Monographs on Statistics and Applied Probability 105

Measurement Error in Nonlinear Models

A Modern Perspective

Second Edition

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (1960)
2 Queues *D.R. Cox and W.L. Smith* (1961)
3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5 Population Genetics *W.J. Ewens* (1969)
6 Probability, Statistics and Time *M.S. Barlett* (1975)
7 Statistical Inference *S.D. Silvey* (1975)
8 The Analysis of Contingency Tables *B.S. Everitt* (1977)
9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10 Stochastic Abundance Models *S. Engen* (1978)
11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12 Point Processes *D.R. Cox and V. Isham* (1980)
13 Identification of Outliers *D.M. Hawkins* (1980)
14 Optimal Design *S.D. Silvey* (1980)
15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16 Classification *A.D. Gordon* (1981)
17 Distribution-Free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19 Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen and E. Pesonen* (1984)
21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25 The Statistical Analysis of Composition Data *J. Aitchison* (1986)
26 Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27 Regression Analysis with Applications *G.B. Wetherill* (1986)
28 Sequential Methods in Statistics, 3rd edition
G.B. Wetherill and K.D. Glazebrook (1986)
29 Tensor Methods in Statistics *P. McCullagh* (1987)
30 Transformation and Weighting in Regression
R.J. Carroll and D. Ruppert (1988)
31 Asymptotic Techniques for Use in Statistics
O.E. Bandorff-Nielsen and D.R. Cox (1989)
32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33 Analysis of Infectious Disease Data *N.G. Becker* (1989)
34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36 Symmetric Multivariate and Related Distributions
K.T. Fang, S. Kotz and K.W. Ng (1990)
37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38 Cyclic and Computer Generated Designs, 2nd edition
J.A. John and E.R. Williams (1995)
39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40 Subset Selection in Regression *A.J. Miller* (1990)
41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44 Inspection Errors for Attributes in Quality Control
N.L. Johnson, S. Kotz and X. Wu (1991)
45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)
46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)
47 Longitudinal Data with Serial Correlation—A State-Space Approach
R.H. Jones (1993)
48 Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49 Markov Models and Optimization *M.H.A. Davis* (1993)
50 Networks and Chaos—Statistical and Probabilistic Aspects
O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall (1993)
51 Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)
52 Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)
53 Practical Risk Theory for Actuaries
C.D. Daykin, T. Pentikäinen and M. Pesonen (1994)
54 Biplots *J.C. Gower and D.J. Hand* (1996)
55 Predictive Inference—An Introduction *S. Geisser* (1993)
56 Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)
57 An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)
58 Nonparametric Regression and Generalized Linear Models
P.J. Green and B.W. Silverman (1994)
59 Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)
60 Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)
61 Statistics for Long Memory Processes *J. Beran* (1995)
62 Nonlinear Models for Repeated Measurement Data
M. Davidian and D.M. Giltinan (1995)
63 Measurement Error in Nonlinear Models
R.J. Carroll, D. Rupert and L.A. Stefanski (1995)
64 Analyzing and Modeling Rank Data *J.J. Marden* (1995)
65 Time Series Models—In Econometrics, Finance and Other Fields
D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (1996)
66 Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)
67 Multivariate Dependencies—Models, Analysis and Interpretation
D.R. Cox and N. Wermuth (1996)
68 Statistical Inference—Based on the Likelihood *A. Azzalini* (1996)
69 Bayes and Empirical Bayes Methods for Data Analysis
B.P. Carlin and T.A Louis (1996)
70 Hidden Markov and Other Models for Discrete-Valued Time Series
I.L. Macdonald and W. Zucchini (1997)
71 Statistical Evidence—A Likelihood Paradigm *R. Royall* (1997)
72 Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)
73 Multivariate Models and Dependence Concepts *H. Joe* (1997)

74 Theory of Sample Surveys *M.E. Thompson* (1997)
75 Retrial Queues *G. Falin and J.G.C. Templeton* (1997)
76 Theory of Dispersion Models *B. Jørgensen* (1997)
77 Mixed Poisson Processes *J. Grandell* (1997)
78 Variance Components Estimation—Mixed Models, Methodologies and Applications *P.S.R.S. Rao* (1997)
79 Bayesian Methods for Finite Population Sampling *G. Meeden and M. Ghosh* (1997)
80 Stochastic Geometry—Likelihood and computation *O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout* (1998)
81 Computer-Assisted Analysis of Mixtures and Applications—Meta-analysis, Disease Mapping and Others *D. Böhning* (1999)
82 Classification, 2nd edition *A.D. Gordon* (1999)
83 Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)
84 Statistical Aspects of BSE and vCJD—Models for Epidemics *C.A. Donnelly and N.M. Ferguson* (1999)
85 Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)
86 The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)
87 Complex Stochastic Systems *O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg* (2001)
88 Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)
89 Algebraic Statistics—Computational Commutative Algebra in Statistics *G. Pistone, E. Riccomagno and H.P. Wynn* (2001)
90 Analysis of Time Series Structure—SSA and Related Techniques *N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky* (2001)
91 Subjective Probability Models for Lifetimes *Fabio Spizzichino* (2001)
92 Empirical Likelihood *Art B. Owen* (2001)
93 Statistics in the 21st Century *Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells* (2001)
94 Accelerated Life Models: Modeling and Statistical Analysis *Vilijandas Bagdonavicius and Mikhail Nikulin* (2001)
95 Subset Selection in Regression, Second Edition *Alan Miller* (2002)
96 Topics in Modelling of Clustered Data *Marc Aerts, Helena Geys, Geert Molenberghs, and Louise M. Ryan* (2002)
97 Components of Variance *D.R. Cox and P.J. Solomon* (2002)
98 Design and Analysis of Cross-Over Trials, 2nd Edition *Byron Jones and Michael G. Kenward* (2003)
99 Extreme Values in Finance, Telecommunications, and the Environment *Bärbel Finkenstädt and Holger Rootzén* (2003)
100 Statistical Inference and Simulation for Spatial Point Processes *Jesper Møller and Rasmus Plenge Waagepetersen* (2004)
101 Hierarchical Modeling and Analysis for Spatial Data *Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand* (2004)
102 Diagnostic Checks in Time Series *Wai Keung Li* (2004)
103 Stereology for Statisticians *Adrian Baddeley and Eva B. Vedel Jensen* (2004)
104 Gaussian Markov Random Fields: Theory and Applications *Håvard Rue and Leonhard Held* (2005)
105 Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition *Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu* (2006)

Monographs on Statistics and Applied Probability 105

Measurement Error in Nonlinear Models

A Modern Perspective

Second Edition

Raymond J. Carroll
David Ruppert
Leonard A. Stefanski
Ciprian M. Crainiceanu

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-633-1 (Hardcover)
International Standard Book Number-13: 978-1-58488-633-4 (Hardcover)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To our families and friends

Preface to the First Edition

This monograph is about analysis strategies for regression problems in which predictors are measured with error. These problems are commonly known as *measurement error modeling* or *errors-in-variables*. There is an enormous literature on this topic in linear regression, as summarized by Fuller (1987). Our interest lies almost exclusively in the analysis of nonlinear regression models, defined generally enough to include generalized linear models, transform-both-sides models, and quasilikelihood and variance function problems.

The effects of measurement error are well known, and we basically assume that the reader understands that measurement error in predictors causes biases in estimated regression coefficients, and hence that the field is largely about correcting for such effects. Chapter 3* summarizes much of what is known about the consequences of measurement error for estimating linear regression parameters, although the material is not exhaustive.

Nonlinear errors-in-variables modeling began in earnest in the early 1980s with the publication of a series of papers on diverse topics: Prentice (1982) on survival analysis; Carroll, Spiegelman, Lan, Bailey, and Abbott (1984) and Stefanski and Carroll (1985) on binary regression; Armstrong (1985) on generalized linear models; Amemiya (1985) on instrumental variables; and Stefanski (1985) on estimating equations. David Byar and Mitchell Gail organized a workshop on the topic in 1987 at the National Institutes of Health, which in 1989 was published as a special issue of *Statistics in Medicine*. Since these early papers, the field has grown dramatically, as evidenced by the bibliography at the end of this book. Unlike the early 1980s, the literature is now so large that it is difficult to understand the main ideas from individual papers. Indeed, a first draft of this book, completed in late 1990, consisted only of the material in four of the first five chapters. Essentially all the rest of the material has been developed since 1990. In a field as rapidly evolving as this one, and with the entrance of many new researchers into the area, we can present but a snapshot of the current state of knowledge.

This book can be divided broadly into four main parts: Chapters 1–2, 3–6, 7–8, and 9–14. In addition, there is Appendix A, a review of relevant fitting methods and statistical models.

The first part is introductory. Chapter 1 gives a number of applications where measurement error is of concern, and defines basic terminology of error structure, data sources and the distinction between functional and structural models. Chapter 2 gives an overview of the important ideas from linear regression, particularly the biases caused by measurement error and some estimation techniques.

The second part gives the basic ideas and techniques of what we call *functional modeling*, where the distribution of the true predictor is *not* modeled parametrically. In addition, in these chapters it is assumed that the true predictor is never observable. The focus is on the additive measurement error model, although periodically we describe modifications for the multiplicative error model. Chapters 3 and 4 discuss two broadly applicable functional methods, regression calibration and simulation-extrapolation (SIMEX), which can be thought of as the default approaches. Chapter 5 discusses a broadly based approach to the use of instrumental variables. All three of these chapters focus on estimators which are easily computed but yield only approximately consistent estimates. Chapter 6 is still based on the assumption that the true predictor is never observable, but here we provide functional techniques which are fully and not just approximately consistent. This material is somewhat more daunting in (algebraic) appearance than the approximate techniques, but even so the methods themselves are often easily programmed. Throughout this part of the book, we use examples of binary regression modeling.

The third part of the book concerns *structural modeling*, meaning that the distribution of the true predictor is parametrically modeled. Chapter 7 describes the likelihood approach to estimation and inference in measurement error models, while Chapter 8 briefly covers Bayesian modeling. Here we become more focused on the distinction between functional and structural modeling, and also describe the measurement error problem as a missing data problem. We also allow for the possibility that the true predictor can be measured in a subset of the study population. The discussion is fully general and applies to categorical data as well as to the additive and multiplicative measurement error models. While at this point the use of structural modeling in measurement error models is not very popular, we believe it will become more so in the very near future.

The fourth part of the book is devoted to more specialized topics. Chapter 9 takes up the study of functional techniques which are applicable when the predictor can be observed in a subset of the study. Chapter 10 discusses functional estimation in models with generalized

* Chapter numbers in this preface refer to the first edition, not the present edition.

linear structure and an unknown link function. Chapter 11 describes the effects that measurement error has on hypothesis testing. Nonparametric regression and density function estimation are addressed in Chapter 12. Errors in the response rather than in predictors are described in Chapter 13. In Chapter 14, a variety of topics are addressed briefly: case-control studies, differential measurement error, functional mixture methods, design of two-stage studies and survival analysis.

We have tried to design the text so that it can be read at two levels. Many readers will be interested only in the background material and in the definition of the specific methods that can be employed. These readers will find that the chapters in the middle two parts of the text (functional and structural modeling) begin with preliminary discussion, move into the definition of the methods, and are then followed by a worked numerical example. The end of the example serves as a flag that the material is about to become more detailed, with justifications of the methods, derivations of estimated standard errors, etc. Those readers who are not interested in such details should skip the material following the examples at first (and perhaps last) reading.

It is our intention that the part of the book on functional models (Chapters 3–6) can be understood at an overview level without an extensive background in theoretical statistics, at least through the numerical examples. The structural modeling approach requires that one knows about likelihood and Bayesian methods, but with this exception the material is not particularly specialized. The fourth part of the book (Chapters 9–14) is more technical, and we suggest that those interested mainly in an overview simply read the first section of each of those chapters.

A full appreciation of the text, especially its details, requires a strong background in likelihood methods, estimating equations and quasilikelihood and variance function models. For inference, we typically provide estimated standard errors, as well as suggest use of "the" bootstrap. These topics are all covered in Appendix A, albeit briefly. For more background on the models used in this monograph, we highly recommend reading Chapter 1 of Fuller (1987) for an introduction to linear measurement error models and the first four chapters of McCullagh and Nelder (1989) for further discussion of generalized linear models, including logistic regression.

This is a book about general ideas and strategies of estimation and inference, not a book about a specific problem. Our interest in the field started with logistic regression, and many of our examples are based upon this problem. However, our philosophy is that measurement error occurs in many fields and in a variety of guises, and what is needed is an outline of strategies for handling progressively more difficult problems. While logistic regression may well be the most important nonlinear measurement error model, the strategies here are applied to a hard-core nonlinear regression bioassay problem (Chapter 3), a changepoint problem (Chapter 7), and a 2 × 2 table with misclassification (Chapter 8). Our hope is that the strategies will be sufficiently clear that they can be applied to new problems as they arise.

We have tried to represent the main themes in the field, and to reference as many research papers as possible. Obviously, as in any monograph, the selection of topics and material to be emphasized reflects our own interests. We apologize in advance to those workers whose work we have neglected to cite, or whose work should have been better advertised.

Carroll's research and the writing of this book were supported by grants from the National Cancer Institute (CA–57030 and CA–61067). After January 1, 1996, Splus and SAS computer programs (on SPARC architecture SunOS versions 4 and 5 and for Windows on PCs), which implement (for major generalized linear models) many of the functional methods described in this book, can be obtained by sending a message to qvf@stat.tamu.edu. The body of the text should contain only a valid return email address. This will generate an automatic reply with instructions on how to get the software.

Much of Stefanski's research on measurement error problems has been supported by grants from the National Science Foundation (DMS–8613681 and DMS–9200915) and by funding from the Environmental Monitoring and Assessment Program, U.S. Environmental Protection Agency.

We want to thank Jim Calvin, Bobby Gutierrez, Stephen Eckert, Joey Lin, C. Y. Wang, and Naisyin Wang for helpful general comments; Donna Spiegelman for a detailed reading of the manuscript; Jeff Buzas, John Cook, Tony Olsen, and Scott Overton for ideas and comments related to our research; and Viswanath Devanarayan for computing assistance and comments. Rob Abbott stimulated our initial interest in the field in 1981 with a question concerning the effects of measurement error in the Framingham Heart Study; this example appears throughout our discussion. Larry Freedman and Mitch Gail have commented on much of our work and have been instrumental in guiding us to interesting problems. Nancy Potischman introduced us to the world of nutritional epidemiology, where measurement error is of fundamental concern. Our friend Leon Gleser has been a source of support and inspiration for many years and has been a great influence on our thinking.

This book uses data supplied by the National Heart, Lung, and Blood Institute, NIH, DHHS from the Framingham Heart Study. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the National Heart, Lung, and Blood Institute or of the Framingham Study.

Preface to the Second Edition

Since the first edition of *Measurement Error in Nonlinear Models* appeared in 1995, the field of measurement error and exposure uncertainty has undergone an explosion in research. Some of these areas are the following:

- Bayesian computation via Markov Chain Monte Carlo techniques are now widely used in practice. The first edition had a short and not particularly satisfactory Chapter 9 on this topic. In this edition, we have greatly expanded the material and also the applications. Even if one is not a card-carrying Bayesian, Bayesian computation is a natural way to handle what we call the structural approach to measurement error modeling.
- A new chapter has been added on longitudinal data and mixed models, areas that have seen tremendous growth since the first edition.
- Semiparametric and nonparametric methods are enjoying increased application. The field of semiparametric and nonparametric regression (Ruppert, Wand, and Carroll, 2003) has become extremely important in the past 11 years, and in measurement error problems techniques are now much better established. We have revamped the old chapter on nonparametric regression and density estimation (Chapter 12) and added a new chapter (Chapter 13) to reflect the changes in the literature.
- Methods for handling covariate measurement error in survival analysis have been developing rapidly. The first edition had a section on survival analysis in the final chapter, "Other Topics." This section has been greatly expanded and made into a separate Chapter 14.
- The area of missing data has also expanded vigorously over the last 11 years, especially due to the work of Robins and his colleagues. This work and its connections with measurement error now needs a book-length treatment of its own. Therefore, with some reluctance, we decided to delete much of the old material on validation data as a missing data problem.
- We have completely rewritten the score function chapter, both to keep up with advances in this area and and to make the exposition more transparent.

The background material in Appendix A has been expanded to make the book somewhat more self-contained. Technical material that appeared as appendices to individual chapters in the first edition has now been collected into a new Appendix B.

Carroll's research has been supported since 1990 by a grant from the National Cancer Institute (CA57030). The work of Raymond Carroll partially occurred during multiple visits to Peter Hall at the Centre of Excellence for Mathematics and Statistics of Complex Systems at the Australian National University, whose support is gratefully acknowledged, along with the opportunity to take thousands of photos of kangaroos (http://www.stat.tamu.edu/∼carroll/compressed_kangaroo.jpg). David Ruppert was supported by the National Science Foundation (DMS 04-538) and the National Institutes of Health (CA57030). Leonard Stefanski also received support from the National Science Foundation and the National Institutes of Health.

In this second edition, we especially acknowledge our colleagues with whom we have discussed measurement error problems and worked since 1995, including Scott Berry, Dennis Boos, John Buonaccorsi, Jeff Buzas, Josef Coresh, Marie Davidian, Eugene Demidenko, Laurence Freedman, Wayne Fuller, Mitchell Gail, Bobby Gutierrez, Peter Hall, Victor Kipnis, Liang Li, Xihong Lin, Jay Lubin, Yanyuna Ma, Doug Midthune, Sastry Pantula, Dan Schafer, John Staudenmayer, Sally Thurston, Tor Toteson, Naisyin Wang, and Alan Welsh. Owen Hoffman introduced us to the problem of radiation dosimetry and the ideas of shared Berkson and classical uncertainties.

We once again acknowledge Robert Abbott for introducing us to the problem in 1980, when he brought to Raymond Carroll a referee report demanding that he explain the impact of measurement error on the (logistic regression) Framingham data. We would love to acknowledge that anonymous referee for starting us along the path of measurement error in nonlinear models.

We also thank Mitchell Gail, one of the world's great biostatisticians, for his advice and friendship over the last 25 years.

We are extremely grateful to Rick Rossi for a detailed reading of the manuscript, a reading that led to many changes in substance and exposition. Rick is the only head of a Department of Mathematics and Statistics who is also a licensed trout-fishing guide.

Finally, and with gratitude, we acknowledge our good friend Leon Gleser, who, to quote the first edition, *has been a source of support and inspiration for many years and has been a great influence on our thinking.*

Our book Web site is
http://www.stat.tamu.edu/∼carroll/eiv.SecondEdition.

Guide to Notation

In this section we give brief explanations and representative examples of the notation used in this monograph. For precise definitions, see the text.

$\widehat{A}_n$, $\widehat{B}_n$	components of the sandwich formula
α_0	intercept in model for $E(\mathbf{X}\vert\mathbf{Z},\mathbf{W})$
α_w	coefficient of $\mathbf{W}$ in model for $E(\mathbf{X}\vert\mathbf{Z},\mathbf{W})$
α_z	coefficient of $\mathbf{Z}$ in model for $E(\mathbf{X}\vert\mathbf{Z},\mathbf{W})$
β_0	intercept in a model for $E(\mathbf{Y}\vert\mathbf{X},\mathbf{Z})$
β_x	coefficient of $\mathbf{X}$ in model for $E(\mathbf{Y}\vert\mathbf{X},\mathbf{Z})$
β_z	coefficient of $\mathbf{Z}$ in model for $E(\mathbf{Y}\vert\mathbf{X},\mathbf{Z})$
$\beta_{\underline{1}ZX}$	coefficient of $\mathbf{1}$ in generalized linear regresssion
Δ	indicator of validation data, for example, where $\mathbf{X}$ is observed
$\dim(\beta)$	dimension of the vector β
f_X	density of $\mathbf{X}$
$f_{Y,W,T\vert Z}$	density of $(\mathbf{Y},\mathbf{W},\mathbf{T})$ given $\mathbf{Z}$
$\mathcal{F}(\cdot)$	unknown link function
$\sigma^2 g(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)$	$\mathrm{var}(\mathbf{Y}\vert\mathbf{Z},\mathbf{X})$ in QVF model
$\mathcal{G}$	extrapolant function in SIMEX
$\mathcal{G}_Q$	quadratic extrapolant function
$\mathcal{G}_{RL}$	rational linear extrapolant function
$\gamma_{0,}\mathrm{cm}$	intercept in a regression calibration model
$\gamma^t_{z,}\mathrm{cm}$	coefficient of $\mathbf{Z}$ in a regression calibration model
$\gamma^t_{w,}\mathrm{cm}$	coefficient of $\mathbf{W}$ in a regression calibration model
$\gamma_{0,}\mathrm{em}$	intercept in an error model
$\gamma^t_{x,}\mathrm{em}$	coefficient of $\mathbf{X}$ in an error model
$\gamma^t_{w,}\mathrm{em}$	coefficient of $\mathbf{W}$ in an error model
$H(v)$	$(1+\exp(-v))^{-1}$, for example, the logistic function
h	bandwidth in nonparametric regression or density estimation
$I_n(\Theta)$	Fisher information
k	With equal replication, number of replicates for all subjects
k_i	Number of replicates of i^{th} subject
$K(\cdot)$	kernel used in nonparametric regression or density estimation
κ_{cm}	$\sigma^2_{\mathrm{cm}}/\sigma^2$
$\Lambda(\cdot)$	likelihood ratio
$\mathcal{L}(\cdot)$	generalized score function
$m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma_{\mathrm{cm}})$	$E(\mathbf{X}\vert\mathbf{Z},\mathbf{W})$
$m_{\mathbf{Y}}(\mathbf{Z},\mathbf{X},\beta)$	$E(\mathbf{Y}\vert\mathbf{Z},\mathbf{X})$ in QVF (quasilikelihood variance function) model
$m_{\mathbf{Y},x}(z,x,\beta)$	$(\partial/\partial x)m_{\mathbf{Y}}(z,x,\beta)$
$m_{\mathbf{Y},xx}(z,x,\beta)$	$(\partial^2/\partial x^2)m_{\mathbf{Y}}(z,x,\beta)$
$\pi(\mathbf{Y},\mathbf{Z},\mathbf{W},\alpha)$	probability of selection into a validation study
Ψ, ψ	estimating functions
$\mathbf{S}$	$\mathbf{Y}$ measured with error ($\mathbf{S}=\mathbf{Y}+\mathbf{V}$)
$s_i(y\vert\Theta)$	score function
σ^2_u	variance of $\mathbf{U}$
$\sigma^2_{X\vert Z}$	conditional variance of $\mathbf{X}$ given $\mathbf{Z}$
σ_{xy}	the covariance between random variables X and Y
ρ_{xy}	the correlation between X and Y, which is defined as $\sigma_{xy}/(\sigma_x\sigma_y)$
Σ_{ZX}	covariance matrix between the random vectors $\mathbf{Z}$ and $\mathbf{X}$
$\mathbf{T}$	observation related to $\mathbf{X}$
$\Theta_b(\lambda)$	simulated estimator used in SIMEX
$\Theta(\lambda)$	average of the $\Theta_b(\lambda)$s
$\mathbf{U}$	observation error in an error model
$\mathbf{U}_{b,k}$	pseudo-error in SIMEX
$\mathbf{V}$	measurement error in the response
$\mathbf{W}$	observation related to $\mathbf{X}$
$\mathbf{X}$	covariates measured with error
$\mathbf{Y}$	response
$\overline{\mathbf{Y}}_{i\cdot}$	average of Y_{ij} over j
$[\tilde{\mathbf{Y}}\vert\tilde{\mathbf{Z}},\tilde{\mathbf{X}},\mathcal{B}]$	density of $\tilde{\mathbf{Y}}$ given $(\tilde{\mathbf{Z}},\tilde{\mathbf{X}},\mathcal{B})$ (Bayesian notation)
$\mathbf{Z}$	covariates measured without error
ζ	parameter controlling amount of simulated extra measurement error in SIMEX

If $m(x)$ is any function, then $m'(x)$ and $m''(x)$ are its first and second derivatives and $m^{(m)}$ is its m^{th} derivative for $m > 2$.

For a vector or matrix A, A^t is its transpose and if A is an invertible matrix, then A^{-1} is its inverse.

If $a = (a_1, \ldots, a_n)$ is a vector, then $\|a\|$ is its Euclidean norm, that is, $\|a\| = \left(\sum_{i=1}^n a_i^2\right)^{1/2}$.

If X and Y are random variables, then $[X]$ is the distribution or X and $[X|Y]$ is the conditional distribution of X given Y. This notation is becoming standard in the Bayesian literature.

Contents

CHAPTER 1

INTRODUCTION

1.1 The Double/Triple Whammy of Measurement Error

Measurement error in covariates has three effects:

- It causes bias in parameter estimation for statistical models.
- It leads to a loss of power, sometimes profound, for detecting interesting relationship among variables.
- It masks the features of the data, making graphical model analysis difficult.

We call the first two the *double whammy* of measurement error. Most of the statistical methods described in this book are aimed at the first problem, namely, to correct for biases of estimation caused by measurement error. Later in this chapter, we will describe an example from radiation dosimetry and the profound loss of power for detecting risks that occurs with uncertainties in individual doses. Here, we briefly describe the third issue, the masking of features.

Consider a regression of a response $\mathbf{Y}$ on a predictor $\mathbf{X}$, uniformly distributed on the interval $[-2, 2]$. Suppose that the mean is $\sin(2\mathbf{X})$ and the variance $\sigma_\epsilon^2 = 0.10$. In the top panel of Figure 1.1, we plot 200 simulated observations from such a model that indicate quite clearly the sinusoidal aspect of the regression function. However, suppose that instead of observing $\mathbf{X}$, we observe $\mathbf{W}$, normally distributed with mean $\mathbf{X}$ but with variance 4/9. As we will later describe in Section 3.2.1, this is an attenuation coefficient of 0.75. Thus, what we observe is not $\mathbf{X}$, but an unbiased estimate of it, $\mathbf{W}$. In the bottom panel of Figure 1.1, we plot the observed data $\mathbf{Y}$ versus $\mathbf{W}$. Note that the sinusoid is no longer evident and the main feature of the data has been *hidden*.

It is also worth noting that the variability about the sinusoid is far smaller when $\mathbf{X}$ is observed than the variability about any curve one could reasonably guess at when only $\mathbf{W}$ is observed. This is one substantial cause of the *loss of power*. Finally, if one only observes $(\mathbf{Y}, \mathbf{W})$ and hence the bottom panel of Figure 1.1, it would be essentially impossible to reconstruct the sinusoid, and something different would certainly be used. This is the *bias* caused by measurement error.

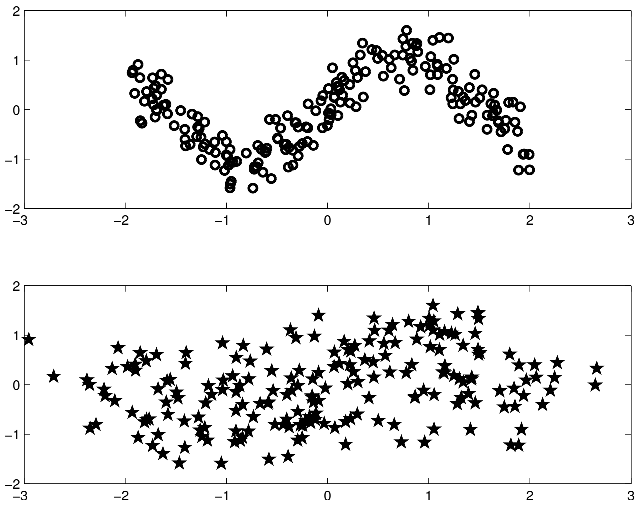


Figure 1.1 *Illustration of the bias, loss of power, and masking of features caused by measurement error in predictors. Top panel regression on the true covariate. Bottom panel regression on the observed covariate.*

1.2 Classical Measurement Error: A Nutrition Example

Much of the measurement error literature is based around what is called *classical measurement error*, in which the truth is measured with additive error, usually with constant variance. We introduce the classical measurement error model via an example from nutrition.

In the National Cancer Institute's OPEN study, see Subar, Thompson, Kipnis, et al. (2001), one interest is in measuring the logarithm of dietary protein intake. True, long-term log-intake is denoted by $\mathbf{X}$, but this cannot be observed in practice. Instead, the investigators measured a biomarker of log-protein intake, namely urinary nitrogen, denoted by $\mathbf{W}$. In this study, 297 subjects had replicated urinary nitrogen measurements. If there were no measurement error, then of course the two biomarker measurements would be equal, but then, since this is a book about measurement error, we would not be wasting space. Indeed, in Figure 1.2 we see that when we plot the second biomarker versus the first, the correlation is relatively high (0.695), but there clearly is some variability in the measurements.

In this context, there is evidence from feeding studies that the protein biomarker captures true protein intake with added variability. Such situations are often called *classical measurement error*. In symbols, let $\mathbf{X}_i$ be the true log-protein intake for individual i, and let $\mathbf{W}_{ij}$ be the j^{th} biomarker log-protein measurement. Then the classical measurement error model states that

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}. \tag{1.1}$$

In this model, $\mathbf{W}_{ij}$ is an unbiased measure of $\mathbf{X}_i$, so that $\mathbf{U}_{ij}$ must have mean zero, that is, in symbols, $E(\mathbf{U}_{ij}|\mathbf{X}_i) = 0$. The error structure of $\mathbf{U}_{ij}$ could be homoscedastic (constant variance) or heteroscedastic. In this particular example, we will show later, in Section 1.7, that the measurement error structure is approximately normal with constant variance, so we can reasonably think that $\mathbf{U}_{ij}|\mathbf{X}_i \sim \text{Normal}(0, \sigma_u^2)$.

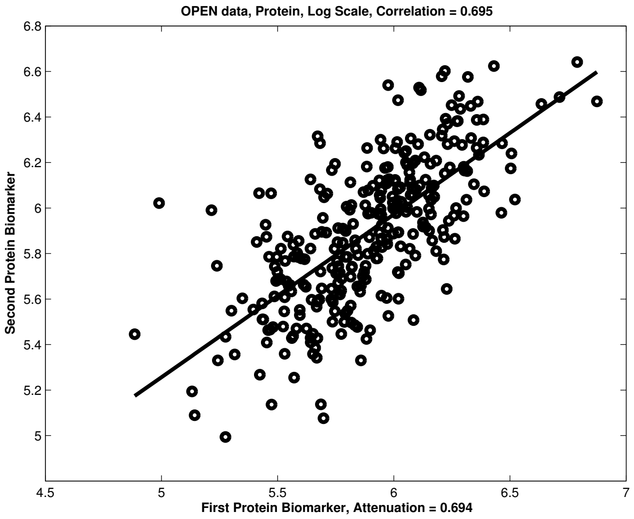


Figure 1.2 *OPEN Study data, scatterplot of the logarithm of the first and second protein biomarker measurements. The fact that there is scatter means that the biomarker has measurement error.*

1.3 Measurement Error Examples

Nonlinear measurement error models commonly begin with an underlying nonlinear model for the response $\mathbf{Y}$ in terms of the predictors. We distinguish between two kinds of predictors: $\mathbf{Z}$ represents those predictors that, for all practical purposes, are measured without error, and $\mathbf{X}$ those that cannot be observed exactly for all study subjects. The distinguishing feature of a measurement error problem is that we can observe a variable $\mathbf{W}$, which is related to an unobservable $\mathbf{X}$. The parameters in the model relating $\mathbf{Y}$ and $(\mathbf{Z}, \mathbf{X})$ cannot, of course, be estimated directly by fitting $\mathbf{Y}$ to $(\mathbf{Z}, \mathbf{X})$, since $\mathbf{X}$ is not observed. The goal of measurement error modeling is to obtain nearly unbiased estimates of these parameters indirectly by fitting a model for $\mathbf{Y}$ in terms of $(\mathbf{Z}, \mathbf{W})$. Attainment of this goal requires careful analysis. Substituting $\mathbf{W}$ for $\mathbf{X}$, but making no adjustments in the usual fitting methods for this substitution, leads to estimates that are biased, sometimes seriously, see Figure 1.1. The problem here is that the parameters of the regression of $\mathbf{Y}$ on $(\mathbf{Z}, \mathbf{W})$ are different from those of $\mathbf{Y}$ on $(\mathbf{Z}, \mathbf{X})$.

In assessing measurement error, careful attention must be given to the type and nature of the error, and the sources of data that allow modeling of this error. The following examples illustrate some of the different types of problems considered in this book.

1.4 Radiation Epidemiology and Berkson Errors

There are many studies relating radiation exposure to disease, including the Nevada Test Site (NTS) Thyroid Disease Study and the Hanford Thyroid Disease Study (HTDS). Full disclosure: One of us (RJC) was involved in litigation concerning HTDS, and his expert report is available at http://www.downwinders.com/files/htds_expert_report.pdf, the plaintiffs' Web site, at least as of May 2005.

Stevens, Till, Thomas, et al. (1992); Kerber, Till, Simon, et al. (1993); and Simon, Till, Lloyd, et al. (1995) described the Nevada test site study, where radiation exposure largely came as the result of above-ground nuclear testing in the 1950s. Similar statistical issues arise in the Hanford Thyroid Disease Study: see Davis, Kopecky, Stram, et al. (2002); Stram and Kopecky (2003); and Kopecky, Davis, Hamilton, et al. (2004), where radiation was released in the 1950s and 1960s. In the Nevada study, over $2,000$ individuals who were exposed to radiation as children were examined for thyroid disease. The primary radiation exposure came from milk and vegetables. The idea of the study was to relate various thyroid disease outcomes to radiation exposure to the thyroid.

Of course, once again, since this is a book about measurement error, the main exposure of interest, radiation to the thyroid, cannot be observed exactly. What is typical in these studies is to build a large dosimetry model that attempts to convert the known data about the above-ground nuclear tests to radiation actually absorbed into the thyroid. Dosimetry calculations in NTS were based on age at exposure, gender, residence history, x-ray history, whether the individual was as a child breast-fed, and a diet questionnaire filled out by the parent, focusing on milk consumption and vegetables. The data were then input into a complex model and, for each individual, the point estimate of thyroid dose and an associated standard error for the measurement error were reported. Roughly similar considerations led to the dose estimates and uncertainties in HTDS.

In both NTS and HTDS, the authors consider analyses taking into account the uncertainties (measurement error) in dose estimates. Indeed, both consider the classical measurement error situation in (1.1). The HTDS study, though, also considered a different type of measurement error, and based most of their power calculations on it. We will go into detail on the power and analysis issues; see Section 1.8.2 of this chapter for power and Section 8.6 for the analysis.

What we see in the classical measurement error model (1.1) is that the observed dose equals the true dose plus (classical) measurement error. This, of course, means that the variability of the observed doses will be greater than the variability of true doses. In HTDS, in contrast, the authors not only consider this classical measurement error, but they also turn the issue around; namely, they assumed that the true dose is equal to the estimated dose plus measurement error. In symbols, this is

$$\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i, \tag{1.2}$$

where $E(\mathbf{U}_i|\mathbf{W}_i) = 0$, so that the true dose has more variability than the estimated dose; contrast with (1.1). Model (1.2) is called a *Berkson* measurement error model, see Berkson (1950).

1.4.1 The Difference Between Berkson and Classical Errors: How to Gain More Power Without Really Trying

Measurement error modeling requires considerable care. In this section, we discuss why it is crucial that one understands the seemingly subtle differences between Berkson and classical errors, and we illustrate some possible pitfalls when choosing between the two error models. As far as we are aware, one cannot be put in jail for using the wrong model, but an incorrect measurement error model often causes erroneous inferences, which to a statistician is worse than going to jail (okay, we have exaggerated). In Section 2.2.2 we provide additional guidance so that the reader can be confident of choosing the correct error model in his/her own work.

The difference between Berkson and classical measurement error is major when one is planning a study *a priori*, especially when one is attempting power calculations. There are some technical similarities between classical and Berkson errors, see Section 3.2.2, but different issues arise in power calculations. What we will indicate here is that *for a given measurement error variance, if you want to convince yourself that you have lots of statistical power despite measurement error, just pretend that the measurement error is Berkson and not classical.*

Suppose that the observed data have a normal distribution with mean zero and variance $\sigma_w^2 = 2.0$. Suppose also that the measurement error has variance $\sigma_u^2 = 1.0$. Then, if one assumes a Berkson model, the true doses have mean zero and variance $\sigma_x^2 = 3.0$. This is so because the variance of $\mathbf{X}$ in (1.2) is the sum of the variance of $\mathbf{W}$ ($\sigma_w^2 = 2.0$) and the variance of the Berkson measurement error $\mathbf{U}$ ($\sigma_u^2 = 1.0$). Now, in major contrast, if one assumes that the measurement error is classical instead of Berkson, then the variance of $\mathbf{X}$ is, from (1.1), the *difference* of the variance of $\mathbf{W}$ (2.0) and the variance of the classical measurement error $\mathbf{U}$ (1.0), that is, 1.0. In other words, if we assume Berkson error, we think that the true dose $\mathbf{X}$ has variance 3.0, while if we assume classical measurement error, we think that the variance of the true dose equals 1.0, a feature reflected in Figure 1.3. Now, for a given set of parameter values of risk, it is generally the case that the power increases when the variance of true exposure $\mathbf{X}$ increases, Hence, assuming Berkson when the error is classical leads to a grossly optimistic overstatement of power.

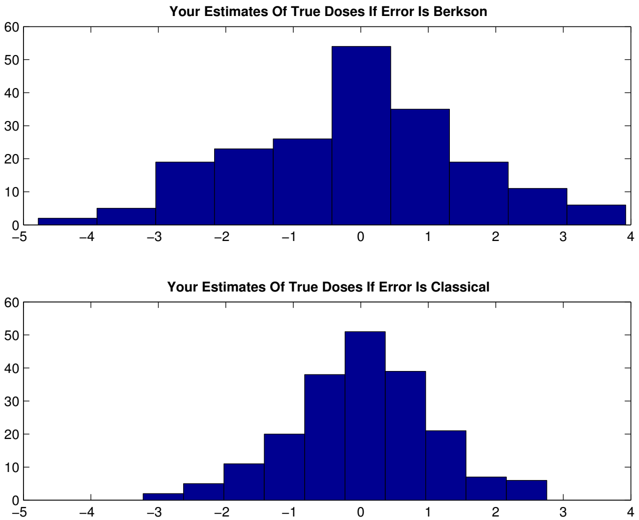


Figure 1.3 *A hypothetical example where the observed doses* $\mathbf{W}$ *have mean zero and variance* 2.0*, while the measurement errors have mean zero and variance* 1.0*. Displayed are the distributions of true dose that you think you have if you think that the errors are Berkson (top) or if you think the errors are classical (bottom). The much smaller variability of true dose under the classical model indicates that the power for detecting effects will be much smaller than if the errors are Berkson.*

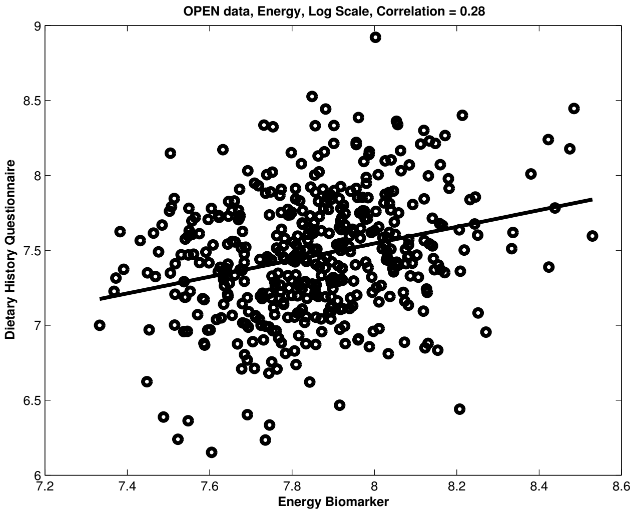


Figure 1.4 *OPEN Study data, scatterplot of the logarithm of energy (calories) using a food frequency questionnaire and a biomarker.*

Further discussion of differences and similarities between power in classical and Berkson error models can be found in Section B.1.

1.5 Classical Measurement Error Model Extensions

It almost goes without saying, but we will say it, that measurement error models can be more complex than the classical additive measurement error model (1.1) or the classical Berkson error model (1.2). Here we illustrate some of the complexities of measurement error modeling via an important nutrition biomarker study.

The study of diet and disease has been a major motivation for nonlinear measurement error modeling. In these studies, it is typical to measure diet via a self–report instrument, for example, a food frequency questionnaire (FFQ), some sort of diary, or a 24-hour recall interview. It has been appreciated for decades that these self-report instruments are only imperfect measures of long-term dietary intakes, and hence that measurement error is a major concern.

To understand the profound nature of measurement error in this context, we consider the National Cancer Institute's OPEN study, which is one of the largest biomarker studies ever done; see Subar, Kipnis, Troiano, et al. (2003) and Kipnis, Midthune, Freedman, et al. (2003). We illustrate this measurement error with energy (caloric) intake mea-

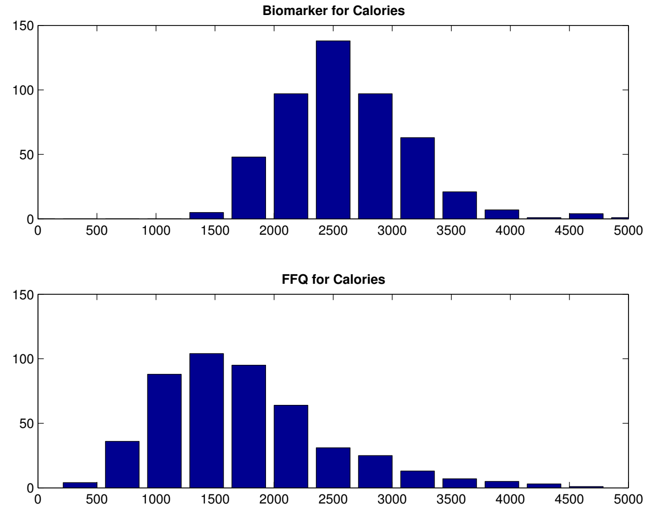


Figure 1.5 *OPEN Study data, histograms of energy (calories) using a biomarker (top panel) and a food frequency questionnaire (bottom panel). Note how individuals report far fewer calories than they actually consume.*

sures. In the OPEN Study, energy intake was measured by the dietary history questionnaire, an FFQ described in Subar, Thompson, Kipnis, et al. (2001). In keeping with our notation, since the FFQ is not the truth, we will denote by $\mathbf{W}$ the log energy intake as measured by the FFQ. In addition, the investigators obtained a near-perfect biomarker measure of energy intake using a technique called doubly-labeled water (DLW), which we call $\mathbf{X}$. DLW is basically what it sounds like: Participants drink water that is enriched with respect to two isotopes, and urine samples allow the measurement of energy expenditure.

That true intake $\mathbf{X}$ and observed intake $\mathbf{W}$ can be very different is seen in Figure 1.4, where we plot the FFQ versus the biomarker along with the associated least squares line. The correlation between truth and observed is only 0.28, indicating that the FFQ is not a very good measure of energy intake. It is also interesting to note the histograms for these two instruments; see Figure 1.5. One can see there that the FFQ is also clearly badly biased downward in general for energy intake, that is, people eat more calories than they are willing to report (no surprise!).

In this example, because of the biases seen in Figures 1.4 and 1.5 the FFQ is not an unbiased measure of true energy intake, and hence the classical measurement error model (1.1) clearly does not hold. A more reasonable model, promoted in a series of papers by Kipnis et al. (1999, 2001, 2003), is to allow for bias as well as variance components

$$\begin{aligned} \mathbf{W}_{ij} &= \gamma_0 + \gamma_1 \mathbf{X}_{ij} + \mathbf{U}_{ij}, \qquad (1.3)\\ \mathbf{U}_{ij} &= r_i + \epsilon_{ij}, \end{aligned}$$

where $r_i \sim \text{Normal}(0, \sigma_r^2)$ and $\epsilon_{ij} \sim \text{Normal}(0, \sigma_\epsilon^2)$. In model (1.3), the linear regression in true intake reflects the biases of the FFQ. The structure of the measurement error random variables $\mathbf{U}_{ij}$ is that they have two components: a shared component r and a random component ϵ. Kipnis et al. (1999, 2001, 2003) call the shared component *person-specific bias*, reflecting the idea that two people who eat exactly the same foods will nonetheless systematically report intakes differently when given multiple FFQs. Fuller (1987) calls the person-specific bias an *equation error*.

Of course, if $\gamma_0 = 0$, $\gamma_1 = 1$, and $r_i \equiv 0$, then we have the standard classical measurement error model (1.1).

1.6 Other Examples of Measurement Error Models

1.6.1 NHANES

The NHANES-I Epidemiologic Study Cohort data set (Jones, Schatzen, Green, et al., 1987) is a cohort study originally consisting of 8,596 women who were interviewed about their nutrition habits and later examined for evidence of cancer. We restrict attention to a subcohort of 3,145 women aged 25–50 who have no missing data on the variables of interest.

The response $\mathbf{Y}$ indicates the presence of breast cancer. The predictor variables $\mathbf{Z}$, assumed to be measured without significant error, include the following: age, poverty index ratio, body mass index, alcohol (Yes, No), family history of breast cancer, age at menarche, and menopausal status. We are primarily interested in the effects of nutrition variables $\mathbf{X}$ that are known to be imprecisely measured, for example, "long-term" saturated fat intake.

If all these underlying variables were observable, then a standard logistic regression analysis would be performed. However, it is both difficult and expensive to measure long-term diet in a large cohort. In the NHANES data, instead of observing $\mathbf{X}$, the measured $\mathbf{W}$ was a 24-hour recall, that is, each participant's diet in the previous 24 hours was recalled and nutrition variables computed. That the measurement error is large in 24-hour recalls has been documented previously (Beaton, Milnor, & Little, 1979; Wu, Whittemore, & Jung, 1986). Indeed, there is evidence to support the conclusion that more than half of the variability in the observed data is due to measurement error.

There are several sources of the measurement error. First, there is the error in the ascertainment of food consumption in the previous 24 hours, especially amounts. Some of this type of error is purely random, while another part is due to systematic bias, for example, some people resist giving an accurate description of their consumption of snacks. The size of potential systematic bias can be determined in some instances (Freedman, Carroll, & Wax, 1991), but in the present study we have available only the 24-hour recall information, and any systematic bias is unidentifiable.

The major source of "error" is the fact that a single day's diet does not serve as an adequate measure of the previous year's diet. There *are* seasonaL differences in diet, as well as day-to-day variations. This points out the fact that measurement error is much more than simple recording or instrument error and encompasses many different sources of variability.

There is insufficient information in the NHANES data to model measurement error directly. Instead, the measurement error structure was modeled using an *external* data set, the CSFII (Continuing Survey of Food Intakes by Individuals) data (Thompson, Sowers, Frongillo, et al., 1992). The CSFII data contain the 24-hour recall measures $\mathbf{W}$, as well as three additional 24-hour recall phone interviews. Using external data, rather than assessing measurement error on an internal subset of the primary study, entails certain risks that we discuss later in this chapter. The basic problem is that parameters in the external study may differ from parameters in the primary study, leading to bias when external estimates are transported to the primary study.

1.6.2 Nurses' Health Study

While the OPEN Study focused on the properties of instruments for measuring nutrient intakes, the real interest is in relating disease and nutrient intakes. A famous and still ongoing study concerning nutrition and breast cancer has been considered by Rosner, Willett, & Spiegelman (1989) and Rosner, Spiegelman, & Willett (1990), namely, the Nurses' Health Study. The study has over 80,000 participants and includes many breast cancer cases. The variables are much the same as in the OPEN study, with the exceptions that (1) alcohol is assessed differently and (2) a food-frequency questionnaire was used instead of 24-hour recall interviews. The size of the measurement error in the nutrition variables is still quite large. Here, $\mathbf{X}$ = (long-term average alcohol intake, long-term average nutrient intake) and $\mathbf{W}$ = (alcohol intake measured by FFQs, nutrient intake measured by FFQs). It is known that $\mathbf{W}$ is both highly variable and biased as an estimator of $\mathbf{X}$.

The Nurses' Health Study was designed so that a direct assessment of measurement error is possible. Specifically, 173 nurses recorded alcohol and nutrient intakes in diary form for four different weeks over the course of a year. The average, $\mathbf{T}$, of these diary entries is taken to be an unbiased estimate of $\mathbf{X}$. We will call $\mathbf{T}$ a *second measure* of $\mathbf{X}$. Thus, in contrast to NHANES, measurement error was assessed on data internal to the primary study. Because $\mathbf{T}$ is unbiased for $\mathbf{X}$, $E(\mathbf{T}|\mathbf{W}) = E(\mathbf{X}|\mathbf{W})$, so we can estimate $E(\mathbf{X}|\mathbf{W})$ by regressing $\mathbf{T}$ on $\mathbf{W}$. Estimating $E(\mathbf{X}|\mathbf{W})$ is the crucial first step in regression calibration, a widely used method of correcting for measurement error; see Chapter 4.

1.6.3 The Atherosclerosis Risk in Communities Study

The Atherosclerosis Risk in Communities (ARIC) study is a multipurpose prospective cohort study described in detail by The ARIC Investigators (1989). From 1987 through 1989, 15,792 male and female volunteers were recruited from four U.S. communities (Forsyth County, NC; suburban Minneapolis, MN; Washington County, MD; and Jackson, MS) for a baseline visit including at-home interviews, clinic examination, and laboratory measurements. Participants returned approximately every three years for second (1990–1992), third (1993–1995), and fourth (1996–98) visits. Time to event data were obtained from annual participant interviews and review of local hospital discharge lists and county death certificates. The "event" was primary coronary kidney disease (CKD).

One purpose of the study was to explain the race effect on the progression of CKD. In particular, African-Americans have maintained approximately four times the age- and sex-adjusted rate of end-stage renal disease (ESRD) compared to whites during the last two decades (USRDS, 2003), while the prevalence of decreased kidney function (CKD Stage 3) in the U.S. is lower among African-Americans than whites. These patterns suggest that that African-Americans progress faster through the different stages of kidney disease.

In Chapter 14 we investigate the race effect on the probability of progression to CKD using a survival data approach. An important confounder is the baseline kidney function, which is typically measured by the estimated glomerular filtration rate (eGFR), which is a noisy version of GFR obtained from a prediction equation. The nature of the adjustment is more complex because of the nonmonotonic relationship between eGFR and progression probability.

1.6.4 Bioassay in a Herbicide Study

Rudemo, Ruppert, & Streibig (1989) consider a bioassay experiment with plants, in which eight herbicides were applied. For each of these eight combinations, six (common) nonzero doses were applied and the dry weight $\mathbf{Y}$ of five plants grown in the same pot was measured. In this instance, the predictor variable **X** of interest is the amount of the herbicide actually absorbed by the plant, a quantity that cannot be measured. Here the response is continuous, and if **X** were observable, then a nonlinear regression model would have been fit, probably by nonlinear least squares. The four-parameter logistic model (not to be confused with logistic regression where the response is binary) is commonly used.

However, **X** is not observable; instead, we know only the nominal concentration **W** of herbicide applied to the plant. The sources of error include not only the error in diluting to the nominal concentration, but also the fact that two plants receiving the same amount of herbicide may absorb different amounts.

In this example, the measurement error was not assessed directly. Instead, the authors assumed that the true amount **X** was linearly related to the nominal amount **W** with nonconstant variance. This error model, combined with the approach discussed in Chapter 4, was used to construct a new model for the observed data.

1.6.5 Lung Function in Children

Tosteson, Stefanski, & Schafer (1989) described an example in which the response was the presence ($\mathbf{Y} = 1$) or absence ($\mathbf{Y} = 0$) of wheeze in children, which is an indicator of lung dysfunction. The predictor variable of interest is $\mathbf{X} =$ personal exposure to NO_2. Since **Y** is a binary variable, if **X** were observable, the authors would have used logistic or probit regression to model the relationship of **Y** and **X**. However, **X** was not available in their study. Instead, the investigators were able to measure a bivariate variable **W**, consisting of observed kitchen and bedroom concentrations of NO_2 in the child's home. School-aged children spend only a portion of their time in their homes, and only a portion of that time in their kitchens and bedrooms. Thus, it is clear that the true NO_2 concentration is not fully explained by what happens in the kitchen and bedroom.

While **X** was not measured in the primary data set, two independent, external studies were available in which both **X** and **W** were observed. We will describe this example in more detail later in this chapter.

1.6.6 Coronary Heart Disease and Blood Pressure

The Framingham study (Kannel, Neaton, Wentworth, et al., 1986) is a large cohort study following individuals for the development **Y** of coronary heart disease. The main predictor of interest in the study is systolic blood pressure, but other variables include age at first exam, body mass, serum cholesterol, and whether or not the person is a smoker. In principle at least, **Z** consists only of age, body mass, and smoking status, while the variables **X** measured with error are serum cholesterol and systolic blood pressure. It should be noted that in a related analysis MacMahon, Peto, Cutler, et al. (1990) consider only the last as a variable measured with error. We will follow this convention in our discussion.

Again, it is impossible to measure long-term systolic blood pressure **X**. Instead, what is available is the blood pressure **W** observed during a clinic visit. The reason that the long-term **X** and the single-visit **W** differ is that blood pressure has major daily, as well as seasonal, variation. Generally, the classical measurement error model (1.1) is used in this context.

In this experiment, we have an extra measurement of blood pressure **T** from a clinic visit taken 4 years before **W** was observed. Hence, unlike any of the other studies we have discussed, in the Framingham study we have information on measurement error for each individual. One can look at **T** as simply a replicate of **W**. However, **T** may be a biased measure of **X** because of temporal changes in the distribution of blood pressure in the population. Each way of looking at the data is useful and leads to different methods of analysis.

1.6.7 A-Bomb Survivors Data

Pierce, Stram, Vaeth, et al. (1992) considered analysis of A-bomb survivor data from the Hiroshima and Nagasaki explosions. They discuss various responses **Y**, including the number of chromosomal aberrations. The true radiation dose **X** cannot be measured; instead, estimates **W** are available. They assume, as an approximation, that $\mathbf{W} = 0$ if and only if $\mathbf{X} = 0$. They adopt a fully parametric approach, specifying that when **X** and **W** are positive, then **W** is lognormal with median **X** and coefficient of variation of 30%. They assume that if **X** is positive, it has a Weibull distribution. In symbols, they propose the multiplicative model

$$\mathbf{W} = \mathbf{X}\,\mathbf{U}, \qquad \log(\mathbf{U}) \sim \text{Normal}(\mu_u, \sigma_u^2),$$

where $\log(\mathbf{U})$ is normally distributed with mean zero and variance 0.0862.

1.6.8 Blood Pressure and Urinary Sodium Chloride

Liu & Liang (1992) described a problem of logistic regression where the response **Y** is the presence of high systolic blood pressure (greater than 140). However, in this particular study blood pressure was measured many times and the average recorded, so that the amount of measurement error in the average systolic blood pressure is reasonably small. The predictors **Z** measured without error are age and body mass index. The predictor **X** subject to measurement error is urinary sodium chloride, which is subject to error because of intra-individual variation over time and also possibly due to measurement error in the chemical analyses. In order to understand the effects of measurement error, 24-hour urinary sodium chloride was measured on 6 consecutive days.

1.6.9 Multiplicative Error for Confidentiality

Hwang (1986) used survey data released by the U. S. Department of Energy on energy consumption by U. S. households. The exact values of certain variables, for example, heating and cooling degree days, were not given since this information might allow the homeowners to be identified. Instead the Department of Energy multiplied these variables by computer-generated random numbers. The Department of Energy released the method for generating the random errors, so this is a rare case where the error distribution is known exactly.

1.6.10 Cervical Cancer and Herpes Simplex Virus

In this example, the question is whether exposure to herpes simplex virus increases the risk of cervical cancer. The data are listed in Carroll, Gail, & Lubin (1993). The response **Y** is the indicator of invasive cervical cancer, **X** is exposure to herpes simplex virus, type 2 (HSV-2) measured by a refined western blot procedure, and **W** is exposure to HSV-2 measured by the western blot procedure. See Hildesheim, Mann, Brinton, et al. (1991) for biological background to this problem. There are 115 complete observations where $(\mathbf{Y}, \mathbf{X}, \mathbf{W})$ is observed and 1,929 incomplete observations where only $(\mathbf{Y}, \mathbf{W})$ is observed. There are 39 cases $(\mathbf{Y} = 1)$ among the complete data and 693 cases among the incomplete data. Among the complete data, there is substantial misclassification, that is, observations where $\mathbf{X} \neq \mathbf{W}$. Also, there is evidence of differential error, meaning that the probability of misclassification depends on the response, that is, $P(\mathbf{X} = \mathbf{W} | \mathbf{X} = \mathbf{x}, \mathbf{Y} = 0) \neq P(\mathbf{X} = \mathbf{W} | \mathbf{X} = \mathbf{x}, \mathbf{Y} = 1)$.

1.7 Checking the Classical Error Model

Suppose that the classical error additive measurement error model (1.1) holds, and that the errors **U** are symmetric and have constant variance in both **X** and any covariates **Z** measured without error, that is, $\text{var}(\mathbf{U}|\mathbf{Z}, \mathbf{X}) = \sigma^2$ (a constant). Then, if the instrument **W** can be replicated, the sample standard deviation of the **W**-values for an individual are uncorrelated with the individual means, and they are also uncorrelated with **Z**. Further, suppose that these errors are normally distributed.

Then differences of the replicates within an individual are normally distributed. This leads to simple graphical devices:

- Plot the sample standard deviation of the **W**-values for an individual against her/his sample mean, call it $\overline{\mathbf{W}}$. If there are no obvious trends, this suggests that the measurement error variance does not depend on **X**.
- Plot the sample standard deviation of the **W**-values for an individual against her/his covariates **Z**. If there are no obvious trends, this suggests that the measurement error variance does not depend on **Z**.
- Form the differences between replications within an individual, and then form a q-q plot of these differences across individuals. If the q-q plot shows no evidence of nonnormality, this suggests that the measurement errors are also roughly normally distributed.

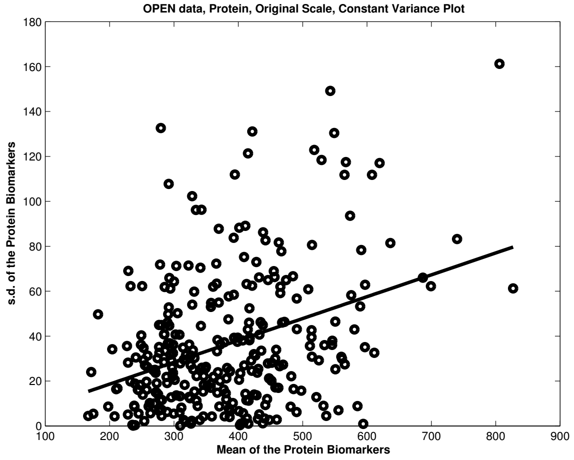


Figure 1.6 *OPEN Study data, plot of the within-individual standard deviation versus mean of the actual untransformed protein biomarkers. The obvious regression slope indicates that the variance of the measurement error depends on true protein intake.*

For example, consider the protein biomarker in the OPEN study; see Section 1.2. In Figure 1.6 we plot the standard deviation of the replicates versus the mean in the original protein scale. The fact that there is an obvious regression slope and the standard deviation of the biomarker varies by a factor of four over the range of the biomarker's mean is strong evidence that, at the very least, the variance of the measurement error depends on true intake.

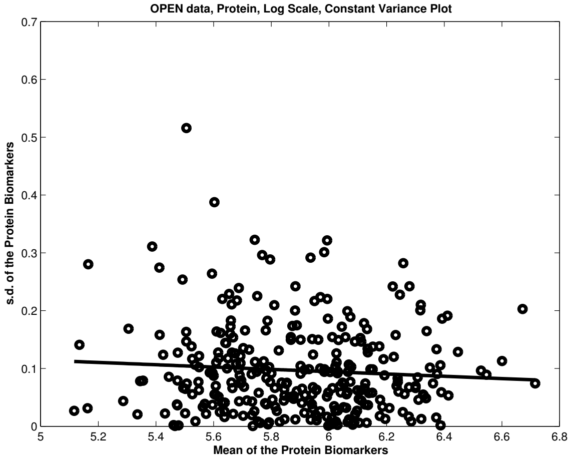


Figure 1.7 *OPEN Study data, plot of the within-individual standard deviation versus mean of the log protein biomarkers. The lack of any major regression slope indicates approximately constant variance measurement error.*

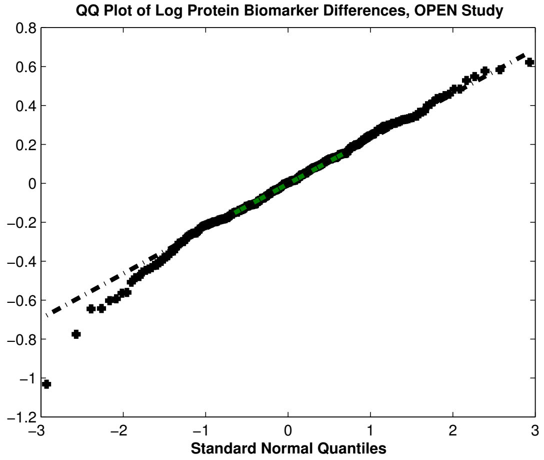


Figure 1.8 *OPEN Study data, q-q plot of the differences of the log protein biomarkers. The nearly straight line of the data indicate nearly normally distributed measurement errors.*

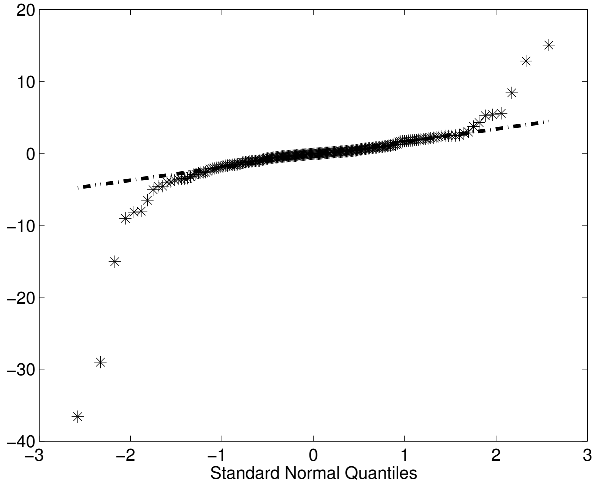


Figure 1.9 *Normal q-q plot of the differences between independent Lognormal(0,1) random variables,* $n = 200$.

A standard way to remove nonconstant variability is via a transformation, and the obvious first attempt is to take logarithms. Figure 1.7 is the standard deviation versus the mean plot in this transformed scale. In contrast to Figure 1.6, here we see no major trend, suggesting that the transformation was successful in removing most of the nonconstant variation. Figure 1.8 gives the q-q plot of the differences: this is not a perfect straight line, but it is reasonably close to straight, suggesting that the transformation has also helped make the data much closer to normally distributed.

Using differences between replicates to assess normality has its pitfalls. The difference between two iid random variables has a symmetric distribution even when the random variable themselves are highly skewed. Thus, nonnormality of measurement errors is somewhat hidden by using differences. For example, Figure 1.9 is a normal q-q plot of the differences between 200 pairs of Lognormal(0,1) random variables; see Section A.2 for the lognormal distribution. Note that the q-q plot shows no sign of asymmetry. Nonnormality is evident only in the presence of heavier-than-Gaussian tails.

1.8 Loss of Power

Classical measurement error causes loss of power, sometimes a profound loss of power. We illustrate this in two situations: linear regression and radiation epidemiology.

1.8.1 Linear Regression Example

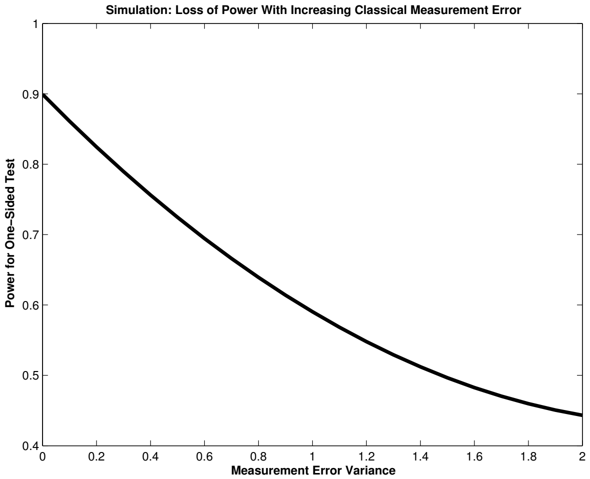


Figure 1.10 *An illustration of the loss of power when there is classical measurement error. When* $\mathbf{X}$ *is observed, the measurement error variance* $= 0.0$, *and the power is* 90%. *When* $\mathbf{X}$ *is not observed and the measurement error variance* $= 1.0$, $1/2$ *of the variability of the observed* $\mathbf{W}$ *is due to noise, and the power is only* 62%. *When* $2/3$ *of the variability of* $\mathbf{W}$ *is due to noise, the power is only* 44%.

Here we consider the simple linear regression model

$$\mathbf{Y}_i = \beta_0 + \beta_x \mathbf{X}_i + \epsilon_i,$$

where $\beta_0 = 0.0$, $\beta_x = 0.69$, $\text{var}(\mathbf{X}) = \text{var}(\epsilon) = 1.0$, and the sample size is $n = 20$. The results here are based on exact calculations using the program nQuery Advisor. The slope was chosen so that, *when* $\mathbf{X}$ *is observed*, there is approximately 90% power for a one-sided test of the null hypothesis $H_0 : \beta_x = 0$.

We added classical measurement error to the true $\mathbf{X}$s using the model (1.1), where we varied the variance of the measurement errors $\mathbf{U}$ from 0.0 to 2.0. When $\text{var}(\mathbf{U}) = 0.0$, we are in the case that there is no classical measurement error, and the power is 90%. When the measurement error variance is $\text{var}(\mathbf{U}) = 1.0$, this means that the observed predictors have variance $\text{var}(\mathbf{W}) = \text{var}(\mathbf{X}) + \text{var}(\mathbf{U}) = 2.0$, and hence $1/2$ of the variability in the observed predictors is due to noise. At the extreme with $\text{var}(\mathbf{U}) = 2.0$, $2/3$ of the variability in the observed predictors is due to noise.

The results are displayed in Figure 1.10. Here we see that while the power would be 90% if $\mathbf{X}$ could be observed, when the measurement error variance equals the variance of $\mathbf{X}$, and hence $1/2$ of the variability in $\mathbf{W}$ is due to noise, the power crashes to 62%. Even worse, when $2/3$ of the variability in the observed $\mathbf{W}$ is noise, the power falls below 50%. This is the first of the double whammy of measurement error; see Section 1.1.

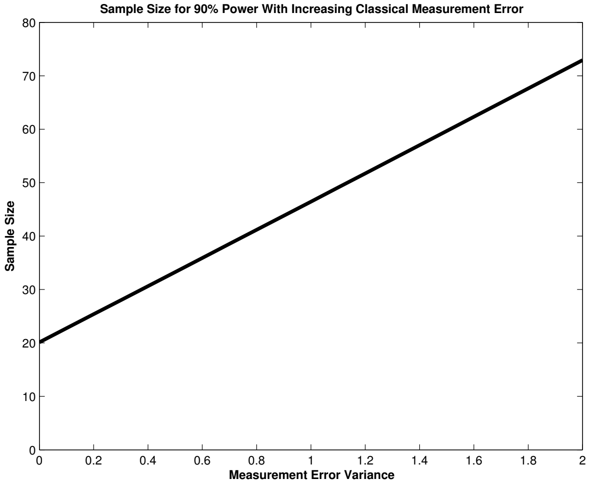


Figure 1.11 *The sample size version of Figure 1.10. When there is no measurement error, the sample size needed for* 90% *power is* $n = 20$. *When* $\mathbf{X}$ *is not observed and the measurement error variance* $= 1.0$, $1/2$ *of the variability of the observed* $\mathbf{W}$ *is due to noise, the necessary sample size for* 90% *power more than doubles to* $n = 45$. *When* $2/3$ *of the variability of* $\mathbf{W}$ *is due to noise, the required sample size is* $n > 70$.

The flip side of a loss of power due to classical measurement error is that sample sizes necessary to gain a given power can increase dramatically. The following power calculations were done assuming all variances are known, and so should be interpreted qualitatively. In Figure 1.11, we show that while only $n = 20$ is required for 90% power when there is no measurement error, when $1/2$ of the variability in the observed predictor $\mathbf{W}$ is due to noise, we require at least $n = 45$ observations, an increase of

200%! Even more dramatic, when 2/3 of the variability in the observed predictor $\mathbf{W}$ is due to noise, the sample size must increase by over 350%!

1.8.2 Radiation Epidemiology Example

In this section, we describe a second simulation showing the effects of classical measurement error. In particular, we show that if one assumes that the measurement error is entirely Berkson but it is partially classical, then one grossly overestimates power.

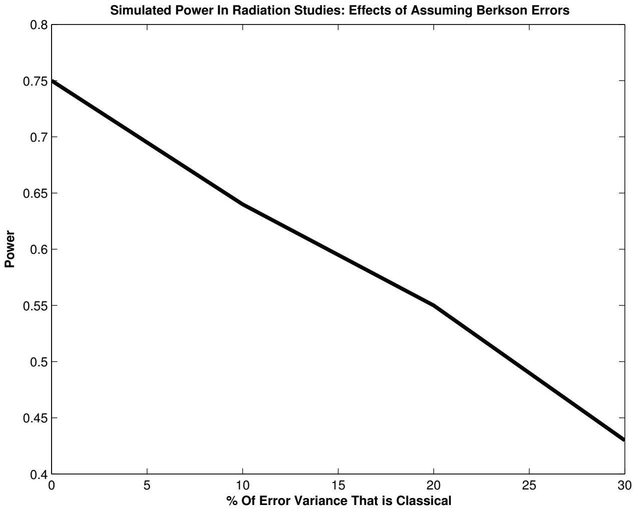


Figure 1.12 *Simulation results for radiation epidemiology, with true excess relative risk* 4.0*. Displayed are the power for detecting an effect from an analysis ignoring measurement error when the percentage of the total error that is classical varies from* 0% *(all Berkson) to* 30% *(majority Berkson). Note the very rapid loss of power that accrues when error become classical in nature. This simulation study shows that if one thinks that all measurement error is Berkson, but* 30% *is classical, then one overestimates the power one really has.*

In the Hanford Thyroid Disease Study (HTDS) litigation, we used what is called an excess relative risk model. Let $\mathbf{Z}$ denote gender, $\mathbf{X}$ denote true but unobservable dose to the thyroid, and $\mathbf{Y}$ be some indicator of thyroid disease. Let $H(x) = \{1 + \exp(-x)\}^{-1}$ be the logistic distribution function, often simply called the logistic function. Then the model fit is

$$\text{pr}(\mathbf{Y} = 1|\mathbf{X}, \mathbf{Z}) = H\left\{\beta_0 + \beta_z \mathbf{Z} + \log(1 + \beta_x \mathbf{X})\right\}. \tag{1.4}$$

The parameter β_x is the excess relative risk parameter.

In our simulations, we used the model of Reeves et al. (2001) and Mallick et al. (2002) to simulate true and calculated doses. This model consists of the variables described above, along with a latent intermediate variable $\mathcal{L}$ between $\mathbf{X}$ and $\mathbf{W}$ that allows for mixtures of Berkson and classical error. This model is

$$\log(\mathbf{X}) = \log(\mathcal{L}) + \mathbf{U}_b, \tag{1.5}$$
$$\log(\mathbf{W}) = \log(\mathcal{L}) + \mathbf{U}_c; \tag{1.6}$$

where $\mathbf{U}_b$ denotes Berkson-type error, and $\mathbf{U}_c$ denotes classical-type error. The standard classical measurement error model (1.1) is obtained by setting $\mathbf{U}_b = 0$. The Berkson model (1.2) is obtained by setting $\mathbf{U}_c = 0$.

The simulation assumed that $\log(\mathcal{L})$, $\mathbf{U}_b$ and $\mathbf{U}_c$ were normally distributed. The details of the simulation were as follows with the values chosen roughly in accord with data in the HTDS:

- The number of study participants was $n = 3,000$, with equal numbers of men and women. With no dose, (β_0, β_z) were chosen so that men had a disease probability of 0.0049, while the disease probability for women was 0.0098. The excess relative risk $\beta_x = 4.0$.
- The mean of $\log(\mathcal{L}) = \log(0.10)$.
- The standard deviation of $\log(\mathbf{W}) = \log(2.7)$.
- The variance of the Berkson errors is σ_b^2, the variance of the classical errors is σ_c^2, and $\sqrt{\sigma_b^2 + \sigma_c^2} = \log(2.3)$.
- The values of σ_b^2 and σ_c^2 were varied to that the total measurement error that is Berkson is 100%, 90%, 80%, and 70%. Stram and Kopecky (2003) mention that there is classical uncertainty in HTDS because of the use of a food frequency questionnaire to measure milk and vegetable consumption.
- A maximum likelihood analysis ignoring measurement errors was employed.

The results of this simulation are displayed in Figures 1.12 and 1.13. Figure 1.12 shows one aspect of the double whammy, namely the profound loss of power when the data are subject to classical measurement error. Note that the power is 75% when all the measurement error is Berkson, but when 30% of the total measurement error variance is classical, the power drops to nearly 40%. In this instance, the practically important part of this simulation has to do with power being announced. If one assumes that all uncertainty is Berkson, then one will announce 75% power to detect an excess relative risk of 4.0, and one would expect to see an effect of this magnitude. However, if in reality 30% of the measurement error is classical, then the actual power is only 40%, and one would not expect to see a statistically significant result. A statistically nonsignificant result could then easily be misinterpreted as a lack of effect, rather than the actuality: a lack of power. If you want to convince yourself that you have lots of statistical power despite measurement error, just pretend that the measurement error is Berkson and not classical.

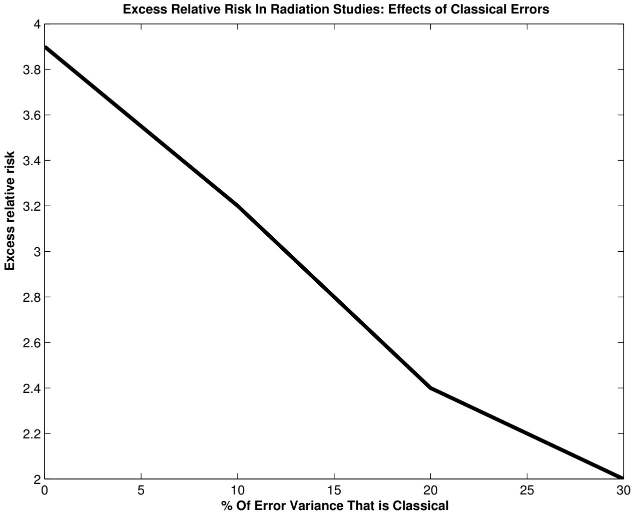


Figure 1.13 *Simulation results for radiation epidemiology, with true excess relative risk* 4.0. *Displayed are the median estimated excess relative risks from an analysis ignoring measurement error when the percentage of the total error that is classical varies from* 0% *(all Berkson) to* 30% *(majority Berkson). Note the very rapid bias that accrues when error become classical in nature.*

Figure 1.13 shows the other aspect of the double whammy, namely the bias caused by classical measurement error. Note that the true value of the excess relative risk is 4.0, and if all uncertainty is Berkson, then the estimated excess relative risk has median very close to the actual value. In other words, there is not much bias in parameter estimation in the Berkson scenario. However, as more of the measurement error becomes classical, a far different story emerges. Thus, if 30% of the measurement error variance is classical, one will tend to observe an excess relative risk of 2.0, half the actual risk. Again, if one assumes all uncertainty is Berkson, one might well conclude that there is little risk of radiation with an excess relative risk of 2.0, but in fact the much larger true relative risk would be masked by the classical measurement error.

1.9 A Brief Tour

As noted in the preface, this monograph is structured into four parts: background material, functional modeling where the marginal distribution of **X** is not modeled, structural modeling where a parametric model for the marginal distribution of **X** is assumed, and specialized topics. Here we provide another brief overview of where we are going.

It is commonly thought that the effect of measurement error is "bias toward the null," and hence that one can ignore measurement error for the purpose of testing whether a predictor is "statistically significant." This lovely and appealing folklore is sometimes true but unfortunately often wrong. The reader may find Chapters 3 and 10 (especially Section 3.6) instructive, for it is in these chapters that we describe in detail the effects of ignoring measurement error.

With continuously measured variables, the classical error model (1.1) is often assumed. The question of how one checks this assumption has not been discussed in the literature. Section 1.7 suggests one such method, namely, plotting the intra-individual standard deviation against the mean, which should show no structure if (1.1) holds. This, and a simple graphical device to check for normality of the errors, are described in Section 8.5. Often, the measured value of **W** is replicated, and the usual assumption is that the replicates are independent.

Having specified an error model, one can use either functional modeling methods (Chapters 4–7) or structural modeling methods (Chapters 8–9). Hypothesis testing is discussed in Chapter 10. Longitudinal data and mixed models are described briefly in Chapter 11. Density estimation and nonparametric regression methods appear in Chapters 12–13. The analysis of survival data and cases where the response is measured with errors occurs in Chapters 14–15.

Bibliographic Notes

The model (1.3) for measurement error of a FFQ actually goes back to Cochran (1968), who cited Pearson (1902) as his inspiration; see Carroll (2003). The paper of Pearson is well worth reading for its innovative use of statistics to understand the biases in self-report instruments, in his case the bisection of lines. Amusingly, Pearson had a colleague who did the actual work of measuring the errors made in bisecting 1,500 lines: "Dr. Lee spent several months in the summer of 1896 in the reduction of the observations," one of the better illustrations of why one does not want to be a postdoc.

Our common notation that **X** stands for the covariate measured with error, **W** is its mismeasured version, and **Z** are the covariates measured without error is not, unfortunately, the only notation in use. Fuller

(1987), for example, used “x” for the covariate measured with error, and “X” for its mismeasured version. Pierce & Kellerer (2004) use “x” for the covariate measured with error and “z” for its mismeasured version. Many authors use “Z” for the mismeasured version of our **X**, and others even interchange the meaning of **Z** and **X**! Luckily, some authors use our convention, for example, Tsiatis & Ma (2004) and Huang & Wang (2001). The annoying lack of a common notation (we, of course, think ours is the best) can make it rather difficult to read papers in the area.

The program Nquery Advisor is available from Statistical Solutions, Stonehill Corporate Center, Suite 104, 999 Broadway, Saugus, MA 01906, http://www.statsol.ie/nquery/nquery.htm. We are not affiliated with that company.

Tosteson, Buzas, Demidenko, & Karagas (2003) studied power and sample size for score tests for generalized regression models with covariate measurement error and provide software which, at the time of this writing, is at http://biostat.hitchcock.org/MeasurementError/Analytics /SampleSizeCalculationsforLogisticRegression.asp.

CHAPTER 2

IMPORTANT CONCEPTS

2.1 Functional and Structural Models

Historically, the taxonomy of measurement error models has been based upon two major defining characteristics. The first is the structure of the error model relating $\mathbf{W}$ to $\mathbf{X}$, and the second is the type and amount of additional data available to assess the important features of this error model, for example, replicate measurements as in the Framingham data or second measurements as in the NHANES study. These two factors, error structure and data structure, are clearly related, since more sophisticated error models can be entertained only if sufficient data are available for estimation. We take up the issue of error models in detail in Section 2.2, although it is a recurrent theme throughout the book.

The second defining characteristic is determined by properties of the unobserved true values $\mathbf{X}_i$, $i = 1, \ldots, n$. Traditionally, a distinction was made between *classical functional* models, in which the $\mathbf{X}$s are regarded as a sequence of unknown fixed constants or parameters, and *classical structural* models, in which the $\mathbf{X}$s are regarded as random variables. The trouble with the classical functional models is that one is then tempted to use the maximum likelihood paradigm to estimate the nuisance parameters, the $\mathbf{X}$s, along with the parameters of interest in a regression model. This approach works in linear regression, but in virtually no other context, see for example Stefanski & Carroll (1985) for one of the many of examples where the method fails.

We believe that it is more fruitful to make a distinction between *functional modeling*, where the $\mathbf{X}$s may be either fixed or random, but in the latter case no, or only minimal, assumptions are made about the distribution of the $\mathbf{X}$s, and *structural modeling*, where models, usually parametric, are placed on the distribution of the random $\mathbf{X}$s. Besides the fact that our approach is cleaner, it also leads to more useful methods of estimation and inference than the old idea of treating the $\mathbf{X}$s as parameters.

Likely the most important concept to keep in mind is the idea of robust model inference. In the functional modeling approach, we make no assumptions about the distribution of the unobserved $\mathbf{X}$s. In contrast, in a typical structural approach, some version of a parametric distribution for the **X**s is assumed, and concern inevitably arises that the resulting estimates and inferences will depend upon the parametric model chosen. Over time, the two approaches have been moving closer to one another. For example, in a number of Bayesian structural approaches, flexible parametric models have been chosen, the flexibility helping in terms of model robustness; see, for example, Carroll, Roeder, & Wasserman (1999) and Mallick, Hoffman, & Carroll (2002). Tsiatis & Ma (2004) described a frequentist method that is functional in the sense that the estimators are consistent no matter what the distribution of **X** is, but the method is also structural in the sense that a pilot distribution for **X** must be specified. The Tsiatis–Ma approach involves solving an integral equation, or at least approximating the solution, and either way of implementation requires major computational effort.

Functional modeling is at the heart of the first part of this book, especially in Chapters 4, 5, and 7. The key point is that even when the **X**s form a random sample from a population, functional modeling is useful because it leads to estimation procedures that are robust to misspecification of the distribution of **X**. As described in Chapter 8, structural modeling has an important role to play in applications (see also Chapter 9), but a concern is the robustness of inference to assumptions made about the distribution of **X**.

Throughout, we will treat $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ as fixed constants, and our analyses will be conditioned on their values. The practice of conditioning on known covariates is standard in regression analysis.

2.2 Models for Measurement Error

2.2.1 General Approaches: Berkson and Classical Models

A fundamental prerequisite for analyzing a measurement error problem is specification of a model for the measurement error process. There are two general types:

- Error models, including classical measurement error models, where the conditional distribution of **W** given $(\mathbf{Z}, \mathbf{X})$ is modeled.
- Regression calibrarion models, including Berkson error models, where the conditional distribution of **X** given $(\mathbf{Z}, \mathbf{W})$ is modeled.

We have already discussed two variants of the *classical error model*, see (1.1) for the simplest additive model, and see (1.3) for an example of biases in the instrument, along with a more complex variance components structure. Somewhat more generally, we can suppose that the relationship between the measured **W** and the unobserved **X** also depends on the observed predictors **Z**, as in for example,

$$\mathbf{W} = \gamma_0 + \gamma_x^t \mathbf{X} + \gamma_z^t \mathbf{Z} + \mathbf{U}, \quad E(\mathbf{U}|\mathbf{X}, \mathbf{Z}) = 0. \tag{2.1}$$

Thus, in the OPEN study described in Section 1.5, it is possible that bias in the FFQ might depend on gender, age, or even body mass index.

In (2.1), the measurement errors **U** have mean zero given the observed and unobserved covariates, but nothing is said otherwise about the structure of **U**. In the OPEN study, the variability might very well be expected to depend on gender. In addition, nothing is said about the distributional structure of **U**, so that, for example, if we have replicated instruments $\mathbf{W}_{ij}$ for the i^{th} person, the $\mathbf{U}_{ij}$ might have a variance components structure, as in (1.3).

By a *regression calibration model* we mean one which focuses on the distribution of **X** given $(\mathbf{Z}, \mathbf{W})$. We have already described the most famous case, the Berkson model; see equation (1.2). More generally, one might be willing to model the distribution of the unobserved covariates directly as a function of the observed versions, as in

$$\mathbf{X} = \gamma_0 + \gamma_1^t \mathbf{W} + \gamma_2^t \mathbf{Z} + \mathbf{U}, \quad E(\mathbf{U}|\mathbf{Z}, \mathbf{W}) = 0. \tag{2.2}$$

The Berkson model says that true **X** is unbiased for nominal **W**, so that $\gamma_0 = \gamma_2 = 0$ and $\gamma_1 = 1$. Mallick & Gelfand (1996) basically started with model (2.2), for example.

2.2.2 Is It Berkson or Classical?

Compared to complex undertakings such as rocket science or automotive repair, determining whether data follow the classical additive measurement error model (1.1) or the standard Berkson error model (1.2) is generally simple in practice. Basically, if the choice is between the two, then the choice is classical if an error-prone covariate is necessarily measured uniquely to an individual, and especially if that measurement can be replicated. Thus, for example, if people fill out a food frequency questionnaire or if they get a blood pressure measurement, then the errors and uncertainties are of classical type. If all individuals in a small group or strata are given the same value of the error-prone covariate, for example, textile workers or miners working in a job classification for a fixed number of years are assigned the same exposure to dust, but the true exposure is particular to an individual, as it almost certainly would be, then the measurement error is Berkson. In the herbicide study, the measured concentration **W** is fixed by design and the true concentration **X** varies due to error, so that the Berkson model is more appropriate.

Other differences between Berkson and classical error models, which might help distinguish between them in practice, are that in the classical model the error **U** is independent of **X**, or at least $E(\mathbf{U}|\mathbf{X}) = 0$, while for Berkson errors **U** is independent of **W** or at least $E(\mathbf{U}|\mathbf{W}) = 0$.

Therefore, $\mathrm{var}(\mathbf{W}) > \mathrm{var}(\mathbf{X})$ for classical errors and $\mathrm{var}(\mathbf{X}) > \mathrm{var}(\mathbf{W})$ for Berkson errors.

In practice, the choice is not always between the classical additive measurement error model (1.1) and the standard Berkson error model (1.2). As we have seen, with a food frequency questionnaire, as well as with other instruments based on self-report, more complex models incorporating biases are required. We still measure quantities unique to the individual, and the measurements can in principle be replicated, but biases must be entertained, and the general classical error model (2.1) is appropriate.

Outside of the Berkson model, the general regression calibration model (2.2) is typically used in ad hoc ways, simply as a modeling device and not based on any fundamental considerations. Briefly, as we will see later in this book, likelihood-type calculations can become fairly simple if one has a regression calibration model in hand and one can estimate it. For example, consider the lung function study of Tosteson et al. (1989). In this study, interest was in the relationship of long-term true NO_2 intake, $\mathbf{X}$, in children on the eventual development of lung disease. The variable $\mathbf{X}$ was not available. The vector $\mathbf{W}$ consists of bedroom and kitchen NO_2 levels as measured by in situ or stationary recording devices. Certainly, $\mathbf{X}$ and $\mathbf{W}$ are related, but children are exposed to other sources of NO_2, for example, in other parts of the house, at school, etc.

The available data consisted of the primary study in which $\mathbf{Y}$ and $\mathbf{W}$ were observed, and two external studies, from different locations, study populations, and investigators, in which $(\mathbf{X}, \mathbf{W})$ were observed. In this problem, the regression calibration model (2.2) seems physically reasonable, because a child's total exposure $\mathbf{X}$ can be thought of as a sum of in-home exposure and other uncontrolled factors. Tosteson, Stefanski, & Schafer (1989) fit (2.2) to each of the external studies, found remarkable similarities in the estimated regression calibration parameters (γ), and concluded that the assumption of a common model for all three studies was a reasonable working assumption.

A general error model (2.1) could also have been fit. However, $\mathbf{W}$ here is bivariate, $\mathbf{X}$ is univariate, and implementation of estimates and inferences is simply less convenient here than it is for a regression calibration model.

2.2.3 Berkson Models from Classical

There is an interesting relationship at a technical level between error models and regression calibration models; see Chapter 4. This relationship is important in regression calibration where a model for $\mathbf{X}$ given $\mathbf{W}$ is needed, but we start with a model for $\mathbf{W}$ given $\mathbf{X}$. If one has a

structural model so that one knows the marginal distribution of $\mathbf{X}$, then an error model can be converted into a regression calibration model by Bayes theorem. Specifically,

$$f_{\mathbf{X}|\mathbf{W}}(x|w) = \frac{f_{\mathbf{W}|\mathbf{X}}(w|x) f_{\mathbf{X}}(x)}{\int f_{\mathbf{W}|\mathbf{X}}(w|x) f_{\mathbf{X}}(x) dx},$$

where $f_{\mathbf{X}}$ is the density of $\mathbf{X}$, $f_{\mathbf{W}|\mathbf{X}}$ is the density of $\mathbf{W}$ given $\mathbf{X}$, and $f_{\mathbf{X}|\mathbf{W}}$ is the density of $\mathbf{X}$ given $\mathbf{W}$. For example, suppose that $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{X}$ and $\mathbf{U}$ are uncorrelated. Then, as discussed in Section A.4, the best linear predictor of $\mathbf{X}$ given $\mathbf{W}$ is $(1 - \lambda)E(\mathbf{X}) + \lambda\mathbf{W}$, and

$$\mathbf{X} = (1 - \lambda)E(\mathbf{X}) + \lambda\mathbf{W} + \mathbf{U}^*, \tag{2.3}$$

where $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is the attenuation, $\mathbf{U}^* = (1-\lambda)\{\mathbf{X} - E(\mathbf{X})\} - \lambda\mathbf{U}$, and a simple calculation shows that $\mathbf{U}^*$ and $\mathbf{W}$ are uncorrelated. If $\mathbf{X}$ and $\mathbf{U}$ are independent and normally distributed, then so are $\mathbf{X}$ and $\mathbf{U}^*$. *Attenuation* means that the magnitude of a regression coefficient is biased towards zero, and the attenuation coefficient measures the size of the attenuation in simple linear regression with classical additive measurement errors; see Section 3.2.1.

Equation (2.3) has the form of a Berkson model, even though the error model is classical. Note, however, that the slope of $\mathbf{X}$ on $\mathbf{W}$ is λ, not 1. Therefore, the variance of $\mathbf{X}$ is smaller than the variance of $\mathbf{W}$ in keeping with the classical rather than Berkson errors.

2.2.4 Transportability of Models

In some studies, the measurement error process is not assessed directly, but instead data from other independent studies, called *external data sets*, are used. In this section, we discuss the appropriateness of using information from independent studies and the manner in which this information should be used.

We say that parameters of a model can be transported from one study to another if the model holds with the same parameter values in both studies. Typically, in applications only a subset of the model parameters need be transportable. Transportability means that not only the model but also the relevant parameter estimates can be transported without bias.

In many instances, approximately the same classical error model holds across different populations. For example, consider systolic blood pressure at two different clinical centers. Assuming similar levels of training for technicians making the measurements and a similar protocol, for example, sitting after a resting period, it is reasonable to expect that the distribution of the error in the recorded measure $\mathbf{W}$ does not depend on the clinical center one enters, or on the technician making the measurement, or on the value of $\mathbf{X}$ being measured, except possibly for heteroscedasticity. Thus, in classical error models it is often reasonable to assume that the error distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ is the same across different populations. However, even here some care is needed because a major component of the measurement error might be sampling error. If the populations differ in temporal variation or sampling frequency, then the error distribution would differ.

Much, much more rarely, the same regression calibration model can sometimes be assumed to hold across different studies. For example, consider the NO_2 study described in Section 1.6.5. If we have two populations of suburban children, then it may be reasonable to assume that the sources of NO_2 exposure other than the bedroom and kitchen will be approximately the same, and the error models are transportable. However, if one study consists of suburban children living in a nonindustrial area and the second study consists of children living in an inner city near an industrialized area, the assumption of transportable error models would be tenuous at best.

2.2.5 Potential Dangers of Transporting Models

The use of independent-study data to assess error model structure carries with it the danger of introducing estimation bias into the primary study analysis.

First, consider the regression calibration model for NO_2 intake. The primary data set of Tosteson Stefanski, & Schafer (1989) (Section 1.6.5) is a sample from Watertown, Massachusetts. Two independent data sets were used to fit the parameters in (2.2): one from the Netherlands and one from Portage, Wisconsin. The parameter estimates for this model in the two external data sets were essentially the same, leading Tosteson et al. (1989) to conclude that the common regression relationship from the Dutch and Portage studies was likely to be appropriate for the Watertown study as well. However, as these authors note in some detail, it is important to remember that this is an *assumption*, plausible in this instance, but still one not to be made lightly. If Watertown were to have a much different pattern of NO_2 exposure than Portage or the Netherlands, then the estimated parameters in model (2.2) from the latter two studies, while similar, might be biased for the Watertown study, and the results for Watertown hence incorrect.

The issue of transporting results for error models is critical in the classical measurement error model as well. Consider the MRFIT study (Kannel et al., 1986), in which $\mathbf{X}$ is long-term systolic blood pressure. The external data set is the Framingham data (MacMahon, Peto, Cutler, et al., 1990). Carroll & Stefanski (1994) discussed these studies in detail, but here we use the studies only to illustrate the potential pitfalls of extrapolating across studies. It is reasonable to assume that classical measurement error model (1.1) holds with the same measurement error variance for both studies, which reduces to stating that the distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ is the same in the two studies. However, the distribution of $\mathbf{X}$ appears to differ substantially in the two studies, with the MRFIT study having smaller variance. Under these circumstances, while the error model is probably transportable, a regression calibration model formed from Framingham would not be transportable to MRFIT. The problem is that, by Bayes's theorem, the distribution of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ depends on both the distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ and the distribution of $\mathbf{X}$ given $\mathbf{Z}$, and the latter is not transportable.

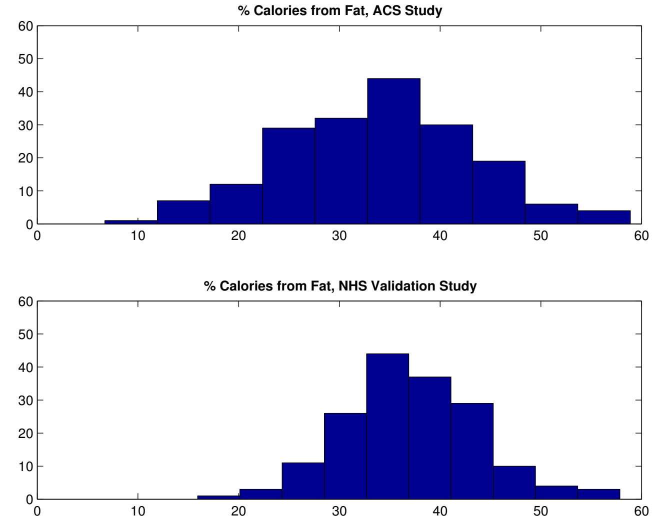


Figure 2.1 *Comparison of % calories from fat using a food frequency questionnaire from two studies. Note how the distributions seem very different, calling into question whether the distribution of true intake can be transported between studies.*

That this is not merely a theoretical exercise is illustrated in Figure 2.1, which shows the distribution of calories from fat (fat density) for two study populations. In this figure, we plot the observed values $\mathbf{W}$ from two studies: the validation arm of the Nurses' Health Study (NHS) and a study done by the American Cancer Society (ACS), both using the same food frequency questionnaire (FFQ). What we see in this figure is that the observed fat density for the ACS study seems to have much more variability than the data in the NHS. Assuming that the error properties of the FFQ were the same in the two studies, it would clearly make no sense to pretend that the distribution of the exact predictor $\mathbf{X}$ is the same in the two studies, that is, the distribution of the exact predictor is not transportable.

2.2.6 Semicontinuous Variables

Some variables—such as nutrient intakes of food groups, such as red meat, or environmental exposures—have a positive probability of being zero and otherwise have a positive continuous distribution. Such variables have been called semicontinuous by Schafer (1997). An example is radiation exposure in the atomic bomb survivors study described in Section 1.6.7. As mentioned in that section, Pierce, Stram, Vaeth, et al. (1992) assume that $\mathbf{W} = 0$ if and only if $\mathbf{X} = 0$. In many studies, this assumption is unlikely to hold and is, at best, a useful approximation.

An alternative model was used by Li, Shao, & Palta (2005). These authors assume that there exists a latent continuous variable $\mathbf{V}$ such that

$$\mathbf{X} = \max(0, \mathbf{V}) \text{ and } \mathbf{W} = \max(0, \mathbf{V} + \mathbf{U})$$

where $\mathbf{U}$ is measurement error. When both $\mathbf{X}$ and $\mathbf{W}$ are positive, then the usual classical measurement error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$ holds. Notice that it is possible for $\mathbf{X}$ to be zero while $\mathbf{W}$ is positive, or vice versa.

2.2.7 Misclassification of a Discrete Covariate

So far in this chapter, it has been assumed that the mismeasured covariate is continuously distributed. For discrete covariates, measurement error means misclassification. A common situation is a binary covariate, where $\mathbf{X}$ and $\mathbf{W}$ are both either 0 or 1, for example, the diagnoses of herpes simplex virus by the refined western blot and western blot tests discussed in Section 1.6.10. In such cases, the misclassification model can be parameterized using the misclassification probabilities $\text{pr}(\mathbf{W} = 1|\mathbf{X} = 0)$ and $\text{pr}(\mathbf{W} = 0|\mathbf{X} = 1)$; see Section 8.4.

2.3 Sources of Data

In order to perform a measurement error analysis, as seen in (2.1)-(2.2), one needs information about either $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Z})$ (classical measurement error) or about $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ (regression calibration). In this section, we will discuss various data sources that allow estimation of

the critical distributions. These data sources can be partitioned into two main categories:

- *Internal* subsets of the primary data.
- *External* or independent studies.

Within each of these broad categories, there are three types of data, all of which we assume to be available in a random subsample of the data set in question:

- *Validation* data in which $\mathbf{X}$ is observable directly. This is the relatively rare circumstance where a measurement error problem is also a missing data problem.
- *Replication* data, in which replicates of $\mathbf{W}$ are available.
- *Instrumental* data, in which another variable $\mathbf{T}$ is observable in addition to $\mathbf{W}$.

An internal validation data set is the ideal, because it can be used with all known techniques, permits direct examination of the error structure, and typically leads to much greater precision of estimation and inference. We cannot express too forcefully that if it is possible to construct an internal validation data set, one should strive to do so. External validation data can be used to assess any of the models (1.1)–(2.2) in the external data, but one is always making an assumption when transporting such models to the primary data.

Usually, one would make replicate measurements if there were good reason to believe that the replicated mean is a better estimate of $\mathbf{X}$ than a single observation, that is, the classical error model is the target. Such data cannot be used to test whether $\mathbf{W}$ is unbiased for $\mathbf{X}$, as in the classical measurement error model (1.1), or biased, as in the general measurement error model (2.1). However, if one is willing to assume (1.1), then replication data can be used to estimate the variance of the measurement error $\mathbf{U}$.

Data sets sometimes contain a second measurement $\mathbf{T}$, which may or may not be unbiased for $\mathbf{X}$, in addition to the primary measurement $\mathbf{W}$. If $\mathbf{T}$ is internal, then it need not be unbiased to be useful. In this case, $\mathbf{T}$ is called an instrumental variable (IV) and can be used in an instrumental variable analysis provided that $\mathbf{T}$ possesses certain other statistical properties (Chapter 6). If $\mathbf{T}$ is external, then it is useful in general only if it is unbiased for $\mathbf{X}$. In this case, $\mathbf{T}$ can be used in a regression calibration analysis (Chapter 4).

2.4 Is There an “Exact” Predictor? What Is Truth?

We have based our discussion on the existence of an exact predictor $\mathbf{X}$ and measurement error models that provide information about this predictor. However, in practice, it is often the case that the term *exact* or *true* needs to be carefully defined prior to discussion of error models.

In almost all cases, one has to take an operational definition for the exact predictor. In the measurement error literature, the term *gold standard* is often used for the operationally defined exact predictor, though sometimes this term is used for an exact predictor that cannot be operationally defined. In the NHANES study, the operational definition is the average saturated food intake over a year-long period *as measured by the average of 24-hour recall instruments*. One can think of this as the best measure of exposure that could possibly be determined in practice, and even here it is extremely difficult to measure this quantity. Having made this operational definition for $\mathbf{X}$, we are in a position to undertake an analysis, for clearly the observed measure $\mathbf{W}$ is unbiased for $\mathbf{X}$ when measured on a randomly selected day. In this case, the measurement error model (1.1) is reasonable. However, in order to ascertain the distributional properties of the measurement error, one requires a replication experiment. The simplest way to take replicates is to perform 24-hour recalls on a few consecutive days (see also Section 1.6.8), but the problem here is that such replicates are probably not conditionally independent given the long-term average, and a variance component model such as (1.3) would likely be required. After all, if one is on an ice cream jag, several consecutive days of ice cream may show up in the 24-hour recall, even though it is rarely eaten.

This type of replication does not measure the true error, which is highly influenced by intraindividual variation in diet. Hence, with replicates on consecutive days, estimating the variance of the measurement error by components-of-variance techniques will underestimate the measurement error.

The same problem may occur in the urinary sodium chloride example (Section 1.6.8), because the replicates were recorded on consecutive days. Liu & Liang (1992) suggested that intraindividual variation is an important component of variability, and the design is not ideal for measuring this variation.

If one wants to estimate the measurement error variance consistently, it is much simpler if replicates can be taken far enough apart in time that the errors can reasonably be considered independent (see Chapter 4 for details). Otherwise, assumptions must be made about the form of the correlation structure; see Wang, Carroll, & Liang (1996). In the CSFII component of the NHANES study, measurements were taken at least two months apart, but there was still some small correlation between errors. In the Nurses' Health Study (Section 1.6.2), the exact predictor is the long-term average intake as measured by food records. Replicated food records were taken at four different points during the year, thus properly accounting for intraindividual variation.

Using an operational definition for an "exact" predictor is often reasonable and justifiable on the grounds that it is the best one could ever possibly hope to accomplish. However, such definitions may be controversial. For example, consider the breast cancer and fat controversy. One way to determine whether changing one's fat intake lowers the risk of developing breast cancer is to do a clinical trial, where the treatment group is actively encouraged to change their dietary behavior. Even this is controversial, because noncompliance can occur in either the treatment or the control arm. If instead one uses prospective data, as in the NHANES study, along with an operational definition of long-term intake, one should be aware that the results of a measurement error analysis could be invalid if true long-term intake and operational long-term intake differ in subtle ways. Suppose that the operational definition of fat and calories could be measured, and call these $(\text{Fat}_O, \text{Calories}_O)$, while the actual long-term intake is $(\text{Fat}_A, \text{Calories}_A)$. If breast cancer risk is associated with age and fat intake through the logistic regression model

$$\begin{aligned}&\Pr(\mathbf{Y} = 1|\text{Fat}_A, \text{Calories}_A, \text{Age})\\&\quad = H\left(\beta_0 + \beta_1\text{Age} + \beta_2\text{Calories}_A + \beta_3\text{Fat}_A\right),\end{aligned}$$

where here and throughout the book, $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function. Then the important parameter is β_3, with $\beta_3 > 0$ corresponding to the conclusion that increased fat intake at a given level of calories leads to increased cancer risk.

However, suppose that the observed fat and calories are actually biased measures of the long-term average:

$$\begin{aligned}\text{Fat}_O &= \gamma_1\text{Fat}_A + \gamma_2\text{Calories}_A;\\ \text{Calories}_O &= \gamma_3\text{Fat}_A + \gamma_4\text{Calories}_A.\end{aligned}$$

Then a little algebra shows that the regression of disease on the operationally defined measures has a slope for operationally defined fat of

$$(\gamma_4\beta_3 - \gamma_3\beta_2) \,/\, (\gamma_1\gamma_4 - \gamma_2\gamma_3).$$

Depending on the parameter configurations, this can take on a sign different from β_3. For example, suppose that $\beta_3 = 0$ and there really is no fat effect. Using the operational definition, a measurement error analysis would lead to a fat effect of $-\gamma_3\beta_2/(\gamma_1\gamma_4 - \gamma_2\gamma_3)$, which may be nonzero. Hence, in this instance, there really is no fat effect, but our operational definition might lead us to find one.

In our experience, researchers in nutrition shy away from terms such as *true intake*, because except for a few recovery biomarkers (protein and energy), the operational definition is not truth. However, the operational definition, for example, the average of many repeated 24-hour recalls, is generally clear to subject-matter experts and not particularly controversial.

2.5 Differential and Nondifferential Error

It is important to make a distinction between *differential* and *nondifferential* measurement error. Nondifferential measurement error occurs when $\mathbf{W}$ contains no information about $\mathbf{Y}$ other than what is available in $\mathbf{X}$ and $\mathbf{Z}$. The technical definition is that measurement error is nondifferential if the distribution of $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ depends only on $(\mathbf{X}, \mathbf{Z})$. In this case, $\mathbf{W}$ is said to be a *surrogate*. In other words, $\mathbf{W}$ is a surrogate if it is *conditionally independent* of the response given the true covariates; measurement error is *differential* otherwise.

For instance, consider the Framingham example of Section 1.6.6. The predictor of major interest is long-term systolic blood pressure ($\mathbf{X}$), but we can only observe blood pressure on a single day ($\mathbf{W}$). It seems plausible that a single day's blood pressure contributes essentially no information over and above that given by true long-term blood pressure, and hence that measurement error is nondifferential. The same remarks apply to the nutrition examples in Sections 1.6.1 and 1.6.2: Dietary intake on a single day should not contribute information about overall health that is not already present in long-term diet intake.

Many problems can plausibly be classified as having nondifferential measurement error, especially when the true and observed covariates occur at a fixed point in time and the response is measured at a later time.

There are two exceptions to keep in mind. First, in case-control or choice-based sampling studies, the response is obtained first and then subsequent follow-up ascertains the covariates. In nutrition studies, this ordering of measurement typically causes differential measurement error. For instance, here the true predictor would be long-term diet before diagnosis, but the nature of case-control studies is that reported diet is obtainable only after diagnosis. A woman who develops breast cancer may well change her diet, so the reported diet as measured after diagnosis is clearly still correlated with cancer outcomes, even after taking into account long-term diet before diagnosis.

A second setting for differential measurement error occurs when $\mathbf{W}$ is not merely a mismeasured version of $\mathbf{X}$, but is a separate variable acting as a type of proxy for $\mathbf{X}$.

For example, in an important paper with major implications for the analysis of retrospective studies in the presence of missing data, Satten & Kupper (1993) described an example of estimating the risk of coronary heart disease where $\mathbf{X}$ is an indicator of elevated LDL (low density lipoprotein cholesterol level), taking the values 1 and 0 according as the LDL does or does not exceed 160. For their value $\mathbf{W}$ they use total cholesterol. In their particular data set, both $\mathbf{X}$ and $\mathbf{W}$ are available, and it transpires that the relationship between $\mathbf{W}$ and $\mathbf{Y}$ is differential, that is, there is still a relationship between the two even after accounting for $\mathbf{X}$. While the example is somewhat forced on our part, one should be aware that problems in which $\mathbf{W}$ is not merely a mismeasured version of $\mathbf{X}$ may well have differential measurement error.

It is also important to realize that the definition of a surrogate depends on the other variables, $\mathbf{Z}$, in the model. For example, consider a model in which $\mathbf{Z}$ has two components, say $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. Then it is possible that $\mathbf{W}$ is a surrogate in the model containing both $\mathbf{Z}_1$ and $\mathbf{Z}_2$ but not in the model containing only $\mathbf{Z}_1$. Buzas et al. (2004) pointed out that this has implications when different models for the response are considered, and they gave a simple example illustrating this phenomenon. We next present a modified version of their algebraic example.

Suppose that $\mathbf{X}, \mathbf{Z}_1, \epsilon_1, \epsilon_2, \mathbf{U}_1$ and $\mathbf{U}_2$ are mutually independent normal random variables with zero means. Define $\mathbf{Z}_2 = \mathbf{X} + \epsilon_1 + \mathbf{U}_1$, $\mathbf{Y} = \beta_1 + \beta_{z_1}\mathbf{Z}_1 + \beta_{z_2}\mathbf{Z}_2 + \beta_x\mathbf{X} + \epsilon_2$, and $\mathbf{W} = \mathbf{X} + \epsilon_1 + \mathbf{U}_2$. Because of joint normality, it is straightforward to show that $E(\mathbf{Y} \mid \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{X}, \mathbf{W}) = E(\mathbf{Y} \mid \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{X})$ and consequently that $\mathbf{W}$ is a surrogate in the model containing both $\mathbf{Z}_1$ and $\mathbf{Z}_2$. However,

$$\begin{aligned} E(\mathbf{Y} \mid \mathbf{Z}_1, \mathbf{X}) &= \beta_0 + \beta_{z_1}\mathbf{Z}_1 + (\beta_{z_2} + \beta_x)\mathbf{X}, \\ E(\mathbf{Y} \mid \mathbf{Z}_1, \mathbf{X}, \mathbf{W}) &= E(\mathbf{Y} \mid \mathbf{Z}_1, \mathbf{X}) + \beta_{z_2} E(\epsilon_1 \mid \mathbf{Z}_1, \mathbf{X}, \mathbf{W}). \quad (2.4) \end{aligned}$$

The last expectation in (2.4) is not equal to zero because $\mathbf{W}$ depends on ϵ_1. Thus $\mathbf{W}$ is not a surrogate in the model that contains only $\mathbf{Z}_1$ unless $\beta_{z_2} = 0$. So the presence or absence of $\mathbf{Z}_2$ in the model determines whether or not $\mathbf{W}$ is a surrogate. The driving feature of this example is that the measurement error, $\mathbf{W} - \mathbf{X}$, is correlated with the covariate $\mathbf{Z}_2$. Problems in which measurement error is correlated with error-free predictors arise in practice and are amenable to the methods of regression calibration in Chapter 4 and instrumental variable estimation in Chapter 6.

The reason why nondifferential measurement error is important is that, as we will show in subsequent chapters, one can typically estimate parameters in models for responses given true covariates, even when the true covariates ($\mathbf{X}$) are not observable. With differential measurement error, this is not the case: Outside of a few special situations, one must observe the true covariate on some study subjects. Most of this book focuses on nondifferential measurement error models.

2.5.0.1 A Technical Derivation

Here is a little technical argument illustrating why nondifferential measurement error is so useful. With nondifferential measurement error, the relationship between $\mathbf{Y}$ and $\mathbf{W}$ is greatly simplified relative to the case of differential measurement error. In simple linear regression, for example, it means that the regression in the observed data is a linear regression of $\mathbf{Y}$ on $E(\mathbf{X}|\mathbf{W})$, because

$$\begin{aligned} E(\mathbf{Y}|\mathbf{W}) &= E\{E(\mathbf{Y}|\mathbf{X},\mathbf{W})|\mathbf{W}\} \\ &= E\{E(\mathbf{Y}|\mathbf{X})|\mathbf{W}\} \\ &= E(\beta_0 + \beta_x\mathbf{X}|\mathbf{W}) \\ &= \beta_0 + \beta_x E(\mathbf{X}|\mathbf{W}). \end{aligned}$$

The assumption of nondifferential measurement error is used to justify the second equality above. This argument is the basis of the regression calibration method; see Chapter 4.

2.6 Prediction

In Chapter 3 we discuss the biases caused by measurement error for estimating regression parameters, and the effects on hypothesis testing are described in Chapter 10. Much of the rest of the book is taken up with methods for removing the biases caused by measurement error, with brief descriptions of inference at each step.

Prediction of a response is, however, another matter. Generally, *there is no need for the modeling of measurement error to play a role in the prediction problem.* If a predictor $\mathbf{X}$ is measured with error and one wants to predict a response *based on the error-prone version* $\mathbf{W}$ *of* $\mathbf{X}$, then except for a special case discussed below, it rarely makes any sense to worry about measurement error. The reason for this is quite simple: $\mathbf{W}$ is error-free as a measurement *of itself*! If one has an original set of data $(\mathbf{Y},\mathbf{Z},\mathbf{W})$, one can fit a convenient model to $\mathbf{Y}$ as a function of $(\mathbf{Z},\mathbf{W})$. Predicting $\mathbf{Y}$ from $(\mathbf{Z},\mathbf{W})$ is merely a matter of using this model for prediction, that is, substituting known values of $\mathbf{W}$ and $\mathbf{Z}$ into the regression model for $\mathbf{Y}$ on $(\mathbf{Z},\mathbf{W})$; the prediction errors from this model will minimize the expected squared prediction errors in the class of all linear unbiased predictors. Predictions with $(\mathbf{Z},\mathbf{W})$ naively substituted for $(\mathbf{Z},\mathbf{X})$ in the regression of $\mathbf{Y}$ on $(\mathbf{Z},\mathbf{X})$ will be biased and can have large prediction errors.

Another potential prediction method is to use the methodology discussed throughout this book to estimate the regression of $\mathbf{Y}$ on $(\mathbf{Z},\mathbf{X})$ and then to substitute into this model $\{\mathbf{Z}, E(\mathbf{X}|\mathbf{W})\}$. Though this seems like a nice idea, it turns out to be equivalent to simply ignoring the measurement error, that is, to substituting $(\mathbf{Z},\mathbf{W})$ into the fitted model for the regression of $\mathbf{Y}$ on $(\mathbf{Z},\mathbf{W})$.

The one situation requiring that we correctly model the measurement error occurs when we develop a prediction model using data from one population but we wish to predict in another population. A naive prediction model that ignores measurement error may not be transportable. In more detail, if $\mathbf{Y} = \beta_0 + \beta_1\mathbf{X} + \epsilon$ in both populations, then $\mathbf{Y} = \beta_0^* + \lambda\beta_x\mathbf{W} + \epsilon^*$, where β_0^* and $\lambda = \sigma_x^2/(\sigma_x^2+\sigma_u^2)$ may differ between populations if either σ_x^2 or σ_u^2 does. Thus, the regression of $\mathbf{Y}$ on $\mathbf{W}$ may be different for the two populations.

Bibliographic Notes

An interesting discussion about the issues of Berkson and classical modeling is given throughout *Uncertainties in Radiation Dosimetry and Their Impact on Dose response Analysis*, E. Ron and F. O. Hoffman, editors, National Cancer Institute Press, 1999. This book, which arose from a conference on radiation epidemiology, has papers or discussions by many leading statisticians. Although we have stated (Section 2.2.2) that “Compared to complex undertakings such as rocket science or automotive repair, determining whether data follow the classical additive measurement error model (1.1) or the standard Berkson error model (1.2) is generally simple in practice,” the conference discussions make it clear that radiation, where the errors are a complex mixture of classical and Berkson errors, is a case where it is difficult to sort through what models to use.

CHAPTER 3

LINEAR REGRESSION AND ATTENUATION

3.1 Introduction

This chapter summarizes some of the known results about the effects of measurement error in linear regression and describes some of the statistical methods used to correct for those effects. Our discussion of the linear model is intended only to set the stage for our main topic, nonlinear measurement error models, and is far from complete. A comprehensive account of linear measurement error models can be found in Fuller (1987).

3.2 Bias Caused by Measurement Error

Many textbooks contain a brief description of measurement error in linear regression, usually focusing on simple linear regression and arriving at the conclusion that the effect of measurement error is to bias the slope estimate in the direction of zero. Bias of this nature is commonly referred to as *attenuation* or *attenuation to the null*.

In fact, though, even this simple conclusion must be qualified, because it depends on the relationship between the measurement, **W**, and the true predictor, **X**, and possibly other variables in the regression model as well. In particular, the effect of measurement error depends on the model under consideration and on the joint distribution of the measurement error and the other variables. In linear regression, the effects of measurement error vary depending on (i) the regression model, be it simple or multiple regression; (ii) whether or not the predictor measured with error is univariate or multivariate; and (iii) the presence of bias in the measurement. The effects can range from the simple attenuation described above to situations where (a) real effects are hidden; (b) observed data exhibit relationships that are not present in the error-free data; and (c) even the signs (±) of estimated coefficients are reversed relative to the case with no measurement error.

The key point is that the measurement error distribution determines the effects of measurement error, and thus appropriate methods for cor-

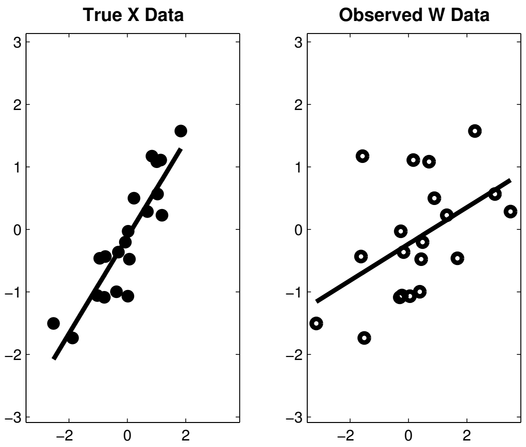


Figure 3.1 *Illustration of additive measurement error model. The left panel displays the true* $(\mathbf{Y}, \mathbf{X})$ *data, while the right panel displays the observed* $(\mathbf{Y}, \mathbf{W})$ *data. Note how the true* $\mathbf{X}$ *data plot has less variability and a more obvious nonzero effect.*

recting for the effects of measurement error depend on the measurement error distribution.

3.2.1 Simple Linear Regression with Additive Error

The basic effects of classical measurement error on simple linear regression can be seen in Figures 3.1 and 3.2. These effects are the double whammy of measurement error described in Section 1.1, namely loss of power when testing and bias in parameter estimation. The third whammy, masking of features, occurs only in nonlinear models, since obviously a straight line has no features to mask.

The left panel of Figure 3.1 displays error-free data $(\mathbf{Y}, \mathbf{X})$ generated from the linear regression model $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, where $\mathbf{X}$ has mean $\mu_x = 0$ and variance $\sigma_x^2 = 1$, the intercept is $\beta_0 = 0$, the slope is $\beta_x = 1$, and the error about the regression line ϵ is independent of $\mathbf{X}$, has mean zero and variance $\sigma_\epsilon^2 = 0.25$. The right panel displays the error-contaminated data $(\mathbf{Y}, \mathbf{W})$ where $\mathbf{W} = \mathbf{X} + \mathbf{U}$, and $\mathbf{U}$ is independent of $\mathbf{X}$, has mean zero, and variance $\sigma_u^2 = 1$. This is the classical additive measurement error model; see Section 1.2. Note how the $(\mathbf{Y}, \mathbf{X})$ data are more tightly grouped around a well delineated line, while the error-prone $(\mathbf{Y}, \mathbf{W})$ data have much more variability about a much less obvious line. This is the loss of power through additional variability.

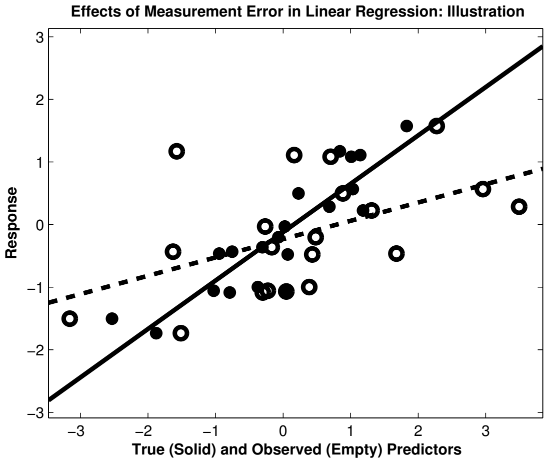


Figure 3.2 *Illustration of additive measurement error model. Here we combine the data in Figure 3.1 and add in least squares fitted lines: The solid line and solid circles are for the true* $\mathbf{X}$ *data, while the dashed line and empty circles are for the observed, error-prone* $\mathbf{W}$ *data. Note how the slope to the true* $\mathbf{X}$ *data is steeper, and the variability about the line is much smaller.*

In Figure 3.2 we combine the data sets: The solid circles and solid line are the $(\mathbf{Y}, \mathbf{X})$ data and least squares fit, while the empty circles and dashed line are the $(\mathbf{Y}, \mathbf{W})$ data and their least squares fit. Here we see the bias in the least squares line due to classical measurement error.

We can understand the phenomena in Figures 3.1–3.2 through some theoretical calculations. For example, it is well known that an ordinary least squares regression of $\mathbf{Y}$ on $\mathbf{W}$ is a consistent estimate not of β_x, but instead of $\beta_{x*} = \lambda \beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1. \tag{3.1}$$

Thus ordinary least squares regression of $\mathbf{Y}$ on $\mathbf{W}$ produces an estimator that is attenuated to zero. The attenuating factor, λ, is called the *reliability ratio* (Fuller, 1987). This attenuation is particularly pronounced in Figures 3.1–3.2.

One would expect that because $\mathbf{W}$ is an error-prone predictor, it has a weaker relationship with the response than does $\mathbf{X}$, as seen in Figure 3.1. This can be seen both by the attenuation and also by the fact that the residual variance of this regression of $\mathbf{Y}$ on $\mathbf{W}$ is

$$\text{var}(\mathbf{Y}|\mathbf{W}) = \sigma_\epsilon^2 + \frac{\beta_x^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \sigma_\epsilon^2 + \lambda \beta_x^2 \sigma_u^2. \tag{3.2}$$

This facet of the problem is often ignored, but it is important. Measurement error causes a double whammy: Not only is the slope attenuated, but the data are more noisy, with an increased error about the line.

It is not surprising that measurement error, as another source of error, increases variability about the line. Indeed, we can substitute $\mathbf{X} = \mathbf{W} - \mathbf{U}$ into the regression model to obtain the model $\mathbf{Y} = \beta_0 + \beta_x \mathbf{W} + (\epsilon - \beta_x \mathbf{U})$, with error $(\epsilon - \beta_x \mathbf{U})$ that has variance $\sigma_\epsilon^2 + \beta_x^2 \sigma_u^2 > \sigma_\epsilon^2$ and covariate $\mathbf{W}$. What may be surprising is that this additional error causes bias. However, the error and the covariate have a common component $\mathbf{U}$, which causes them to be correlated. The correlation between the error and covariate is the source of the bias.

In light of the effects of classical measurement error discussed above, one might expect that the least squares estimate of slope calculated from measured $(\mathbf{Y}, \mathbf{W})$ is more variable than the slope estimator calculated from the true $(\mathbf{Y}, \mathbf{X})$ data. This is not always the case. Buzas, Stefanski, and Tosteson (2004) pointed out that the naive estimate of slope can be *less variable* than the true data estimator. In fact, for the classical error model, the variance of the naive estimator is less than the variance of the true-data estimator asymptotically if and only if $\beta_x^2 \sigma_x^2 / (\sigma_x^2 + \sigma_u^2) < \sigma_\epsilon^2 / \sigma_x^2$, which is possible when σ_ϵ^2 is large, or σ_u^2 is large, or β_x^2 is small. So, relative to the case of no measurement error, classical errors can result in more precise estimates of the wrong, that is, biased, quantity. This phenomenon explains, in part, why naive-analysis confidence intervals often have disastrous coverage probabilities; not only are they centered on the wrong value, but they sometimes have shorter length than would be obtained with the true data. This phenomenon cannot occur with Berkson errors, for which the variance of the naive estimator is never less than the variance of the true-data estimator asymptotically.

3.2.2 Regression Calibration: Classical Error as Berkson Error

There is another way of looking at the bias that will give further insight, namely that by a simple mapping, classical measurement error can be made into a Berkson model. Define $\mathbf{W}_{\text{blp}} = (1 - \lambda)\mu_x + \lambda \mathbf{W}$, the best linear predictor of $\mathbf{X}$ based on $\mathbf{W}$. Then, by (A.8) of Appendix A,

$$\mathbf{X} = \mathbf{W}_{\text{blp}} + \mathbf{U}^*, \tag{3.3}$$

where $\mathbf{U}^*$ is uncorrelated with $\mathbf{W}$, and var$U^* = \lambda \sigma_u^2$. Compare (3.3) with the formal definition of a Berkson error model (1.2) in Section 1.4. Effectively, we have a formal transformation of the classical error model into a Berkson error model, where the observed predictor is now the best linear predictor of $\mathbf{X}$ from $\mathbf{W}$. The calculation leading to (3.3) is at the heart of the regression calibration method of Chapter 4.

Equation (3.3) has important consequences in fitting the linear regression model and correction for the bias due to classical measurement error: Little (generally nothing) can be done to eliminate the loss of power. Substituting (3.3) for $\mathbf{X}$ into the regression model, we have

$$\begin{aligned} \mathbf{Y} &= \beta_0 + \beta_x(1 - \lambda)\mu_x + \beta_x \lambda \mathbf{W} + (\epsilon + \beta_x \mathbf{U}^*) \\ &= \beta_0 + \beta_x \mathbf{W}_{\text{blp}} + \epsilon + \beta_x \mathbf{U}^*. \end{aligned} \tag{3.4}$$

In (3.4) the error $\epsilon + \beta_x \mathbf{U}^*$ is uncorrelated with the regressor $\mathbf{W}_{\text{blp}}$ and has variance $\sigma_\epsilon^2 + \lambda \beta_x^2 \sigma_u^2$ in agreement with (3.2). Moreover, the regression of $\mathbf{Y}$ on $\mathbf{W}$ has intercept $\beta_0 + \beta_x(1 - \lambda)\mu_x$ and slope $\lambda \beta_x$, which explains the attenuation of the slope and the additive bias of the intercept.

However, these considerations show a way to eliminate bias. By (3.4), we have $\mathbf{Y} = \beta_0 + \beta_x \mathbf{W}_{\text{blp}} + \epsilon + \beta_x \mathbf{U}^*$, so if we replace the unknown $\mathbf{X}$ by $\mathbf{W}_{\text{blp}}$, which is known since it depends only on $\mathbf{W}$, then we have a regression model with intercept equal to β_0, slope equal to β_x, and error uncorrelated with the regressor. Therefore, regressing $\mathbf{Y}$ on $\mathbf{W}_{\text{blp}}$ gives unbiased estimates of β_0 and β_x. In fact, regressing $\mathbf{Y}$ on $\mathbf{W}_{\text{blp}}$ is equivalent to the method-of-moments correction for attenuation discussed in Section 3.4.1. Replacing $\mathbf{X}$ with its predictor $\mathbf{W}_{\text{blp}}$ is the key idea behind the technique of regression calibration discussed in Chapter 4. Of course, $\mathbf{W}_{\text{blp}}$ is "known" only if we know λ and μ_x. In practice, these parameters need to be estimated.

3.2.3 Simple Linear Regression with Berkson Error

Suppose that we have linear regression, $\mathbf{Y}_i = \beta_0 + \beta_x \mathbf{X}_i + \epsilon_i$, with unbiased Berkson error, that is, $\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i$. Then $E(\mathbf{X}_i|\mathbf{W}_i) = \mathbf{W}_i$ so that $E(\mathbf{Y}_i|\mathbf{W}_i) = \beta_0 + \beta_x \mathbf{W}_i$. As a consequence, the naive estimator that regresses $\mathbf{Y}_i$ on $\mathbf{W}_i$ is unbiased for β_0 and β_x. This unbiasedness can be seen in Figure 3.3 which illustrates linear regression with Berkson errors. In the figure, $(\mathbf{Y}_i, \mathbf{X}_i)$ and $(\mathbf{Y}_i, \mathbf{W}_i)$ are plotted, as well as fits to both $(\mathbf{Y}_i, \mathbf{X}_i)$ and $(\mathbf{Y}_i, \mathbf{W}_i)$. The $\mathbf{W}_i$ are equally spaced on $[-1, 3]$, $\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i$, $\mathbf{U}_i = \text{Normal}(0, 1)$, $\epsilon_i = \text{Normal}(0, 0.5)$, $n = 50$, $\beta_0 = 1$, and $\beta_x = 1$.

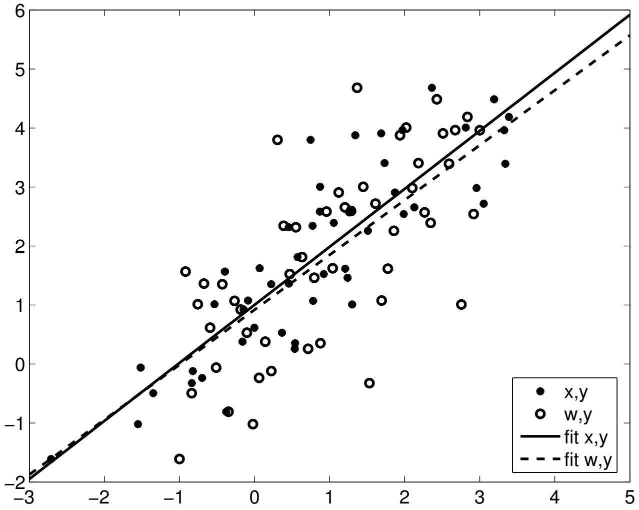


Figure 3.3 *Simple linear regression with unbiased Berkson errors. Theory shows that the fit of* $\mathbf{Y}_i$ *to* $\mathbf{W}_i$ *is unbiased for the regression of* $\mathbf{Y}_i$ *on* $\mathbf{X}_i$*, and the two fits are, in fact, similar.*

3.2.4 Simple Linear Regression, More Complex Error Structure

Despite admonitions of Fuller (1987) and others to the contrary, it is a common perception that the effect of measurement error is always to attenuate the line. In fact, attenuation depends critically on the classical additive measurement error model. In this section, we discuss two deviations from the classical additive error model that do not lead to attenuation.

We continue with the simple linear regression model, but now we make the error structure more complex in two ways. First, we will no longer insist that $\mathbf{W}$ be unbiased for $\mathbf{X}$. The intent of studying this departure from the classical additive error model is to study what happens when one pretends that one has an unbiased surrogate, but in fact the surrogate is biased.

A second departure from the additive model is to allow the errors in the linear regression model to be correlated with the errors in the predictors. This is differential measurement error; see Section 2.5. One example where this problem arises naturally is in dietary calibration studies (Freedman et al., 1991). In a typical dietary calibration study, one is interested in the relationship between a self-administered food frequency questionnaire (FFQ, the value of $\mathbf{Y}$) and usual (or long-term) dietary intake (the value of $\mathbf{X}$) as measures of, for example, the percentage of calories from fat in a person's diet. FFQs are thought to be biased for usual intake, and in a calibration study researchers will obtain a second measure (the value of $\mathbf{W}$), typically either from a food diary or from an interview in which the study subject reports his or her diet in the previous 24 hours. In this context, it is often assumed that the diary or recall is unbiased for usual intake. In principle, then, we have simple linear regression with an additive measurement error model, but in practice a complication can arise. It is often the case that the FFQ and the diary or recall are given very nearly contemporaneously in time, as in the Women's Health Trial Vanguard Study (Henderson et al., 1990). In this case, it makes little sense to pretend that the error in the relationship between the FFQ ($\mathbf{Y}$) and usual intake ($\mathbf{X}$) is uncorrelated with the error in the relationship between a diary or recall ($\mathbf{W}$) and usual intake. This correlation has been demonstrated (Freedman, Carroll, and Wax, 1991), and in this section we will discuss its effects.

To express the possibility of bias in $\mathbf{W}$, we write the model as $\mathbf{W} = \gamma_0 + \gamma_1\mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $\mathbf{X}$ and has mean zero and variance σ_u^2. To express the possibility of correlated errors, we will write the correlation between ϵ and $\mathbf{U}$ as $\rho_{\epsilon u}$. The classical additive measurement error model sets $\gamma_0 = 0$, $\rho_{\epsilon u} = 0$, and $\gamma_1 = 1$, so that $\mathbf{W} = \mathbf{X} + \mathbf{U}$.

If $(\mathbf{X}, \epsilon, \mathbf{U})$ are jointly normally distributed, then the regression of $\mathbf{Y}$ on $\mathbf{W}$ is linear with intercept

$$\beta_{0*} = \beta_0 + \beta_x\mu_x - \beta_{x*}(\gamma_0 + \gamma_1\mu_x),$$

and slope

$$\beta_{x*} = \frac{\beta_x\gamma_1\sigma_x^2 + \rho_{\epsilon u}\sqrt{\sigma_\epsilon^2\sigma_u^2}}{\gamma_1^2\sigma_x^2 + \sigma_u^2}. \tag{3.5}$$

Examination of (3.5), shows that if $\mathbf{W}$ is biased ($\gamma_1 \neq 1$) or if there is significant correlation between the measurement error and the error about the true line ($\rho_{\epsilon u} \neq 0$), it is possible for $|\beta_{x*}| > |\beta_x|$, an effect exactly the opposite of attenuation. Thus, correction for bias induced by measurement error clearly depends on the nature, as well as the extent, of the measurement error.

For purposes of completeness, we note that the residual variance of the linear regression of $\mathbf{Y}$ on $\mathbf{W}$ is

$$\mathrm{var}(\mathbf{Y}|\mathbf{W}) = \sigma_\epsilon^2 + \frac{\beta_x^2\sigma_u^2\sigma_x^2 - \rho_{\epsilon u}^2\sigma_\epsilon^2\sigma_u^2 - 2\beta_x\gamma_1\sigma_x^2\rho_{\epsilon u}\sqrt{\sigma_\epsilon^2\sigma_u^2}}{\gamma_1^2\sigma_x^2 + \sigma_u^2}.$$

3.2.4.1 *Diagnostic for Correlation of Errors in Regression and Measurement Errors*

In some cases, there is a simple graphical diagnostic to check whether the errors in the regression are correlated with the classical measurement errors. The methods are related to the graphical diagnostics used to detect whether the additive error model is reasonable; see Section 1.7.

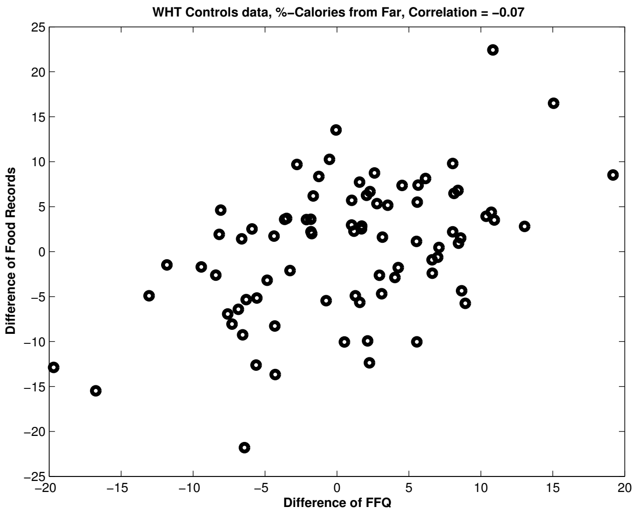


Figure 3.4 *Women's Health Trial Vanguard Study Data. This is a plot for % calories from fat of the differences of food records and the differences of food frequency questionnaires. With replicated* **Y** *and* **W***, this plot is a diagnostic for whether errors in a regression are correlated with the classical measurement errors.*

Specifically, suppose that the error-prone instrument is replicated, so that we observe $\mathbf{W}_{ij} = \gamma_0 + \gamma_1 \mathbf{X}_i + \mathbf{U}_{ij}$. The difference $\mathbf{W}_{i1} - \mathbf{W}_{i2} = \mathbf{U}_{i1} - \mathbf{U}_{i2}$ is "pure" error, unrelated to $\mathbf{X}_i$. Suppose further that the response is replicated, so that we observe $\mathbf{Y}_{ij} = \beta_0 + \beta_x \mathbf{X}_i + r_i + \epsilon_{ij}$, where r_i is person-specific bias or equation error; see Section 1.5. Then differences $\mathbf{Y}_{i1} - \mathbf{Y}_{i2} = \epsilon_{i1} - \epsilon_{i2}$ are the model errors. A plot of the two sets of differences will help reveal whether the regression errors and the measurement errors are correlated. This is illustrated in Figure 3.4, where there appears to be a very strong correlation between the model errors and the measurement errors. A formal test can be performed by regressing one set of differences on the other and testing the null hypothesis that the slope is zero. This plotting method and the test assume that the covariances of errors separated in time are small. This assumption seems reasonable if the time separation is at all large.

3.2.5 *Summary of Simple Linear Regression*

Before continuing with a discussion of the effects of measurement error in multiple linear regression, we summarize the primary effects of measurement error in simple linear regression for various types of error models that we study throughout the book. Table 3.1 displays the important error-model parameters and linear regression model parameters for the case that $(\mathbf{Y}, \mathbf{X}, \mathbf{W})$ are multivariate normal for a hierarchy of error model types. In all cases, the underlying regression model is

$$\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon, \tag{3.6}$$

where $\mathbf{X}$ and ϵ are independent and ϵ has mean zero and variance σ^2.

3.2.5.1 *Differential Error Measurement*

The least restrictive type of error model is one in which **W** is not unbiased and the error is differential. This is also the most troublesome type of error in the sense that correcting for bias requires the most additional information or data. The first row in Table 3.1 shows how the parameters in the regression of **Y** on **W** depend on the true-data regression model parameters, β_0, β_x, σ^2, in this case. Note that to recover β_x from the regression of **Y** on **W** one would have to know or be able to estimate the covariances, σ_{xw} and $\sigma_{\epsilon w}$. Also, with a differential-error measurement it is possible for the residual variance in the regression of **Y** on **W** to be *less than* σ^2.

3.2.5.2 *Surrogate Measurement*

As defined in Section, 2.5, a surrogate measurement is one for which the conditional distribution of **Y** given $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ depends only on $(\mathbf{X}, \mathbf{Z})$. In this case, **W** is also said to be a *surrogate*. The second row of Table 3.1 shows how the parameters in the regression of **Y** on **W** depend on β_0, β_x, σ^2 when **W** is a surrogate, with no additional assumptions about the type of error model. With a surrogate, it is apparent that knowledge of or estimability of σ_{xw} is enough to recover β_x from the regression of **Y** on **W**. The residual variance in the regression of **Y** on **W** is *always greater than* σ^2 when **W** is a surrogate. In this sense, a surrogate is always less informative than **X**.

Error Model	ρ^2_{xw}	Intercept	Slope	Residual Variance
Differential	ρ^2_{xw}	$\beta_0 + \beta_x\mu_x - \frac{\beta_x\sigma_{xw}+\sigma_{\epsilon w}}{\sigma^2_w}\mu_w$	$\beta_x\left(\frac{\sigma_{xw}}{\sigma^2_w}\right) + \frac{\sigma_{\epsilon w}}{\sigma^2_w}$	$\sigma^2_\epsilon + \beta^2_x\sigma^2_x - \frac{(\sigma_{xw}\beta_x+\sigma_{\epsilon w})^2}{\sigma^2_w}$
Surrogate	ρ^2_{xw}	$\beta_0 + \beta_x\mu_x - \frac{\beta_x\sigma_{xw}}{\sigma^2_w}\mu_w$	$\beta_x\left(\frac{\sigma_{xw}}{\sigma^2_w}\right)$	$\sigma^2_\epsilon + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
Classical	$\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_{u_c}}$	$\beta_0 + \beta_x\mu_x\left(1-\rho^2_{xw}\right)$	$\beta_x\left(\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_{u_c}}\right)$	$\sigma^2_\epsilon + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
B/C mixture	$\frac{\sigma^4_L(\sigma^2_L+\sigma^2_{u_b})^{-1}}{(\sigma^2_L+\sigma^2_{u_c})}$	$\beta_0 + \beta_x\mu_x\left(1-\frac{\sigma^2_L}{\sigma^2_L+\sigma^2_{u_c}}\right)$	$\beta_x\left(\frac{\sigma^2_L}{\sigma^2_L+\sigma^2_{u_c}}\right)$	$\sigma^2_\epsilon + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
Berkson	$\frac{\sigma^2_x-\sigma^2_{u_b}}{\sigma^2_x}$	β_0	β_x	$\sigma^2_\epsilon + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
No error	1	β_0	β_x	σ^2_ϵ

Table 3.1: Table entries are error model squared correlations, and intercepts, slopes and residual variances of the linear model relating $\mathbf{Y}$ to $\mathbf{W}$ when $(\mathbf{Y}, \mathbf{X}, \mathbf{W})$ is multivariate normal for the cases $\mathbf{W}$ is: a general differential measurement, a general surrogate, an unbiased classical-error measurement, an unbiased classical/Berkson mixture error measurement, an unbiased Berkson measurement, and the case of no error ($\mathbf{W} = \mathbf{X}$). Classical error variance, $\sigma^2_{u_c}$; Berkson error variance, $\sigma^2_{u_b}$; B/C mixture error model, $\mathbf{X} = \mathcal{L} + \mathbf{U}_b$, $\mathbf{W} = \mathcal{L} + U_c$, $\mathbf{Y} = \beta_0 + \beta_x\mathbf{X} + \epsilon$.

3.2.5.3 Classical Error Model

In the classical error model, $\mathbf{W}$ is a surrogate and $E(\mathbf{W} \mid \mathbf{X}) = \mathbf{X}$, and we can write $\mathbf{W} = \mathbf{X} + \mathbf{U}_c$ where $\mathbf{U}_c$ is a measurement error. Here we use the subscript c to emphasize that the error is classical and to avoid confusion with the two error models discussed below. We have already discussed this model in detail elsewhere, for example, in Sections 1.2 and 2.2. It is apparent from the third row of Table 3.1 that if the reliability ratio, $\lambda = \sigma^2_x/(\sigma^2_x + \sigma^2_{u_c})$ is known or can be estimated, then β_x can be recovered from the regression of $\mathbf{Y}$ on $\mathbf{W}$.

3.2.5.4 Berkson Error Model

In the Berkson error model, $\mathbf{W}$ is a surrogate and $E(\mathbf{X} \mid \mathbf{W}) = \mathbf{W}$, and we can write $\mathbf{X} = \mathbf{W} + \mathbf{U}_b$ where $\mathbf{U}_b$ is a Berkson error. This model has been discussed in detail elsewhere, for example, Sections 1.4 and 2.2. It is apparent from the fifth row of Table 3.1 that the regression parameters are not biased by Berkson measurement error. However, note that the residual variance in the regression of $\mathbf{Y}$ on $\mathbf{W}$ is greater than σ^2, a consequence of the fact that for this model $\mathbf{W}$ is a surrogate. Both the unbiasedness and increased residual variation are well illustrated in Figure 3.3.

3.2.5.5 Berkson/Classical Mixture Error Model

We now consider an error model that was encountered previously (see Section 1.8.2) on the log-scale, and is discussed again at length in Section 8.6. Here we consider the additive version. The defining characteristic is that the error model contains both classical and Berkson components. Specifically, it is assumed that

$$\mathbf{X} = \mathcal{L} + \mathbf{U}_b, \tag{3.7}$$

$$\mathbf{W} = \mathcal{L} + \mathbf{U}_c. \tag{3.8}$$

When $\mathbf{U}_b = 0$, $\mathbf{X} = \mathcal{L}$ and the classical error model is obtained, whereas the Berkson error model results when $\mathbf{U}_c = 0$, since then $\mathbf{W} = \mathcal{L}$. We denote the variances of the error terms by $\sigma^2_{u_c}$ and $\sigma^2_{u_b}$. This error model has features of both the classical and Berkson error models. Note that there is bias in the regression parameters when $\sigma^2_{u_c} > 0$, as in the classical model. The inflation in the residual variance has the same form as the other nondifferential error models in terms of ρ^2_{xw}, but ρ^2_{xw} depends on both error variances for this model.

The error models in Table 3.1 are arranged from most to least problematic in terms of the negative effects of measurement error. Although we discussed the Berkson/classical mixture error model last, in the hierarchy of error models its place is between the classical and Berkson error models.

3.3 Multiple and Orthogonal Regression

3.3.1 Multiple Regression: Single Covariate Measured with Error

In multiple linear regression, the effects of measurement error are more complicated, even for the classical additive error model.

We now consider the case where $\mathbf{X}$ is scalar, but there are additional covariates $\mathbf{Z}$ measured without error. The linear model is now

$$\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \beta_z^t \mathbf{Z} + \epsilon, \tag{3.9}$$

where $\mathbf{Z}$ and β_z are column vectors, and β_z^t is a row vector. In Appendix B.2 it is shown that if $\mathbf{W}$ is unbiased for $\mathbf{X}$, and the measurement error $\mathbf{U}$ is independent of $\mathbf{X}$, $\mathbf{Z}$ and ϵ, then the least squares regression estimator of the coefficient of $\mathbf{W}$ consistently estimates $\lambda_1\beta_x$, where

$$\lambda_1 = \frac{\sigma^2_{x|z}}{\sigma^2_{w|z}} = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u}, \tag{3.10}$$

and $\sigma^2_{w|z}$ and $\sigma^2_{x|z}$ are the residual variances of the regressions of $\mathbf{W}$ on $\mathbf{Z}$ and $\mathbf{X}$ on $\mathbf{Z}$, respectively. Note that λ_1 is equal to the simple linear regression attenuation, $\lambda = \sigma^2_x/(\sigma^2_x + \sigma^2_u)$, only when $\mathbf{X}$ and $\mathbf{Z}$ are uncorrelated. Otherwise, $\sigma^2_{x|z} < \sigma^2_x$ and $\lambda_1 < \lambda$, showing that collinearity increases attenuation.

The problem of measurement error–induced bias is not restricted to the regression coefficient of $\mathbf{X}$. The coefficient of $\mathbf{Z}$ is also biased in general, unless $\mathbf{Z}$ is independent of $\mathbf{X}$ (Carroll, Gallo, and Gleser, 1985; Gleser, Carroll, and Gallo, 1987). In Section B.2 it is shown that for the model (3.9), the naive ordinary least squares estimates not β_z but rather

$$\beta_{z*} = \beta_z + \beta_x(1 - \lambda_1)\Gamma_z, \tag{3.11}$$

where Γ_z^t is the coefficient of $\mathbf{Z}$ in the regression of $\mathbf{X}$ on $\mathbf{Z}$, that is, $E(\mathbf{X} \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z^t \mathbf{Z}$.

This result has important consequences when interest centers on the effects of covariates measured *without* error. Carroll et al. (1985) and Carroll (1989) showed that in the two-group analysis of covariance where $\mathbf{Z}$ is a treatment assignment variable, naive linear regression produces a consistent estimate of the treatment effect only if the design is balanced, that is, $\mathbf{X}$ has the same mean in both groups and is independent of treatment. With considerable imbalance, the naive analysis may lead to the conclusions that (i) there is a treatment effect when none actually exists; and (ii) the effects are negative when they are actually positive,

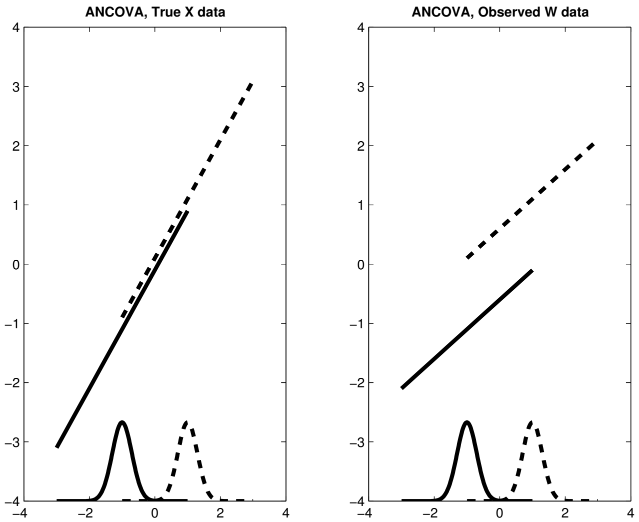


Figure 3.5 *Illustration of the effects of measurement error in an unbalanced analysis of covariance. The left panel shows the actual* $(\mathbf{Y}, \mathbf{X})$ *fitted functions, which are the same, indicating no treatment effect. The density function of* $\mathbf{X}$ *in the two groups are very different, however, as can be seen in the schematic density functions of* $\mathbf{X}$ *at the bottom. The right panel shows what happens when there is measurement error in the continuous covariate: Now the observed data suggest a large treatment effect.*

or vice versa. Figure 3.5 illustrates this process schematically. In the left panel, we show linear regression fits in the analysis of covariance model when there is no effect of treatment, that is, the two lines are the same. At the bottom of this panel, we draw schematic density functions for $\mathbf{X}$ in the two groups: The solid lines are the treatment group with smaller $\mathbf{X}$. The effect of measurement error in this problem is attenuation *around the mean in each group*, leading to the right panel, where the linear regression fits to the observed $\mathbf{W}$ are given. Now note that the lines are not identical, indicating that we would observe a treatment effect, even though it does not exist.

3.3.2 Multiple Covariates Measured with Error

Now suppose that there are covariates $\mathbf{Z}$ measured without error, that $\mathbf{W}$ is unbiased for $\mathbf{X}$, which may consist of multiple predictors, and that the linear regression model is $\mathbf{Y} = \beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z} + \epsilon$. If we write Σ_{ab} to be the covariance matrix between random variables $\mathbf{A}$ and $\mathbf{B}$, then naive ordinary linear regression consistently estimates not (β_x, β_z) but rather

$$\begin{aligned}\begin{pmatrix}\beta_{x*}\\ \beta_{z*}\end{pmatrix} &= \begin{pmatrix}\Sigma_{xx}+\Sigma_{uu} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{pmatrix}^{-1}\\ &\qquad\left\{\begin{pmatrix}\Sigma_{xy}\\ \Sigma_{zy}\end{pmatrix}+\begin{pmatrix}\Sigma_{u\epsilon}\\ 0\end{pmatrix}\right\} \qquad (3.12)\\ &= \begin{pmatrix}\Sigma_{xx}+\Sigma_{uu} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{pmatrix}^{-1}\\ &\qquad\left\{\begin{pmatrix}\Sigma_{xx} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{pmatrix}\begin{pmatrix}\beta_x\\ \beta_z\end{pmatrix}+\begin{pmatrix}\Sigma_{u\epsilon}\\ 0\end{pmatrix}\right\}.\end{aligned}$$

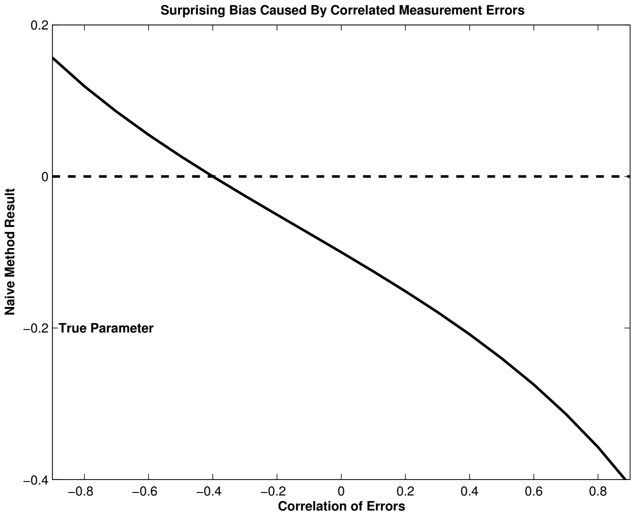


Figure 3.6 *Illustration of the effects of correlated measurement error with two variables measured with error. The true variables are actually uncorrelated, while the errors are correlated, with correlations ranging from* -0.9 *to* 0.9. *Displayed is a plot of what least squares estimates against the correlation of the measurement errors. The true value of the parameter of interest is* -0.2.

Thus, ordinary linear regression is biased. In Section 3.4, we take up the issue of bias correction. However, before doing so, it is worth taking a minute to explore the bias result (3.12). Consider a case of a regression on two error prone covariates, where the coefficient β_x in the regression of $\mathbf{Y}$ on $\mathbf{X}$ is $(1.0, -0.2)^t$, and where the components of $\mathbf{X}$ are independent so that Σ_{xx} is the identity matrix. Let the variance of the measurement errors $\mathbf{U}$ both $= 1.0$, and let their correlation ρ vary from -0.9 to 0.9. In Figure 3.6 we graph what least squares ignoring measurement error is really estimating in the second component (-0.2) of β_x as ρ varies. When the correlation between the measurement error is large but negative, least squares actually suggests that the coefficient is *positive* when it really is negative. Equally surprising, if the correlation between the measurement errors is large and positive, least squares actually suggests a more negative effect than actually exists.

3.4 Correcting for Bias

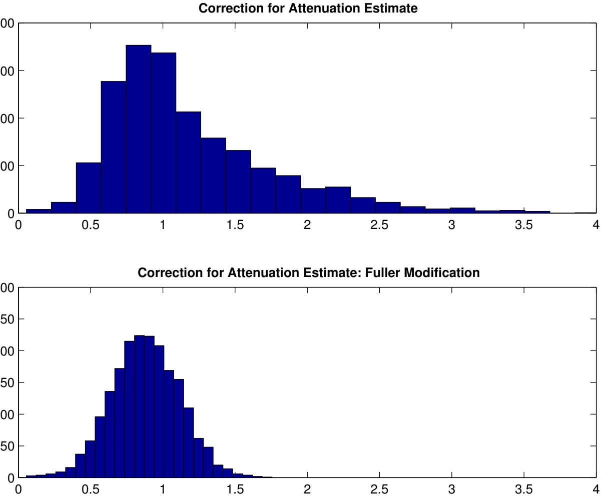
As we have just seen, the ordinary least squares estimator is typically biased under measurement error, and the direction and magnitude of the bias depends on the regression model, the measurement error distribution, and the correlation between the true predictor variables. In this section, we describe two common methods for eliminating bias.

3.4.1 Method of Moments

In simple linear regression with the classical additive error model, we have seen in (3.1) that ordinary least squares is an estimate of $\lambda\beta_x$, where λ is the reliability ratio. If the reliability ratio were known, then one could obtain an unbiased estimate of β_x simply by dividing the ordinary least squares slope $\widehat{\beta}_{x*}$ by the reliability ratio.

Of course, the reliability ratio is rarely known in practice, and one has to estimate it. If $\widehat{\sigma}_u^2$ is an estimate of the measurement error variance (this is discussed in Section 4.4), and if $\widehat{\sigma}_w^2$ is the sample variance of the $\mathbf{W}$s, then a consistent estimate of the reliability ratio is $\widehat{\lambda} = (\widehat{\sigma}_w^2 - \widehat{\sigma}_u^2)/\widehat{\sigma}_w^2$. The resulting estimate is $\widehat{\beta}_{x*}/\widehat{\lambda}$.

In small samples, the sampling distribution of $\widehat{\beta}_{x*}/\widehat{\lambda}$ is highly skewed, and in such cases a modified version of the method-of-moments estimator is recommended (Fuller, 1987; Section 2.5.1). Fuller's modification depends upon a tuning parameter α. Fuller does not give explicit advice about choosing α, but in his simulations $\alpha = 2$ produced more accurate estimates than the unmodified estimator. As an example, in Figure 3.7, in the top panel, we plot the histogram of the corrected estimate when $n = 20$, $\mathbf{X}$ is standard normal, the reliability ratio $= 0.5$, and the error about the line in the regression model is 0.25: The skewness is clear. In the bottom panel, we plot the histogram of Fuller's corrected estimator: It is slightly biased downwards, but very much more symmetric. In this figure, Fuller's method was defined as follows. Let $\widehat{\sigma}_{yw}$ and $\widehat{\sigma}_y^2$ be the sample covariance between $\mathbf{Y}$ and $\mathbf{W}$ and the sample variance of $\mathbf{Y}$, respectively. Define $\widehat{\kappa} = (\widehat{\sigma}_w^2 - \widehat{\sigma}_{yw}^2/\widehat{\sigma}_y^2)/\widehat{\sigma}_u^2$.


Correction for Attenuation Estimate

Correction for Attenuation Estimate: Fuller Modification

Figure 3.7 *Illustration of the small-sample distribution of the method-of-moments estimator of the slope in simple linear regression when* $n = 20$ *and the reliability ratio* $\lambda = 0.5$. *The top panel is the usual method-of-moments estimate, while the bottom panel is Fuller's correction to it.*

Then define $\widehat{\sigma}_x^2 = \widehat{\sigma}_w^2 - \widehat{\sigma}_u^2$ if $\widehat{\kappa} \geq 1 + (n-1)^{-1}$, while otherwise $\widehat{\sigma}_x^2 = \widehat{\sigma}_w^2 - \widehat{\sigma}_u^2\{\widehat{\kappa} - (n-1)^{-1}\}$. Then Fuller's corrected estimate with his $\alpha = 2$ is given as $(\widehat{\beta}_{x*}\widehat{\sigma}_w^2)/\{\widehat{\sigma}_x^2 + 2\widehat{\sigma}_u^2/(n-1)\}$.

The algorithm described above is called the *method-of-moments* estimator. The terminology is apt, because ordinary least squares and the reliability ratio depend only on moments of the observed data.

The method-of-moments estimator can be constructed for the general linear model, not just for simple linear regression. Suppose that $\mathbf{W}$ is unbiased for $\mathbf{X}$, and consider the general linear regression model with $\mathbf{Y} = \beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z} + \epsilon$.

The ordinary least squares estimator is biased even in large samples because it estimates (3.12). When Σ_{uu} and $\Sigma_{u\epsilon}$ are known or can be estimated, (3.12) can be used to construct a simple method-of-moments estimator that is commonly used to correct for the bias. Let S_{ab} be the sample covariance between random variables $\mathbf{A}$ and $\mathbf{B}$. The method-of-moments estimator that corrects for the bias in the case that Σ_{uu} and $\Sigma_{u\epsilon}$ are known is

$$\begin{pmatrix} S_{ww} - \Sigma_{uu} & S_{wz} \\ S_{zw} & S_{zz} \end{pmatrix}^{-1} \begin{pmatrix} S_{wy} - \Sigma_{u\epsilon} \\ S_{zy} \end{pmatrix}, \qquad (3.13)$$

In the case that Σ_{uu} and $\Sigma_{u\epsilon}$ are estimated, the estimates replace the known values in (3.13). It is often reasonable to assume that $\Sigma_{u\epsilon} = 0$, in which case (3.13) simplifies accordingly.

In the event that $\mathbf{W}$ is biased for $\mathbf{X}$, that is, $\mathbf{W} = \gamma_0 + \gamma_x\mathbf{X} + \mathbf{U}$, that is, the error calibration model, the method-of-moments estimator can still be used, provided estimates of (γ_0, γ_x) are available. The strategy is to calculate the estimators above using the error-calibrated variate $\mathbf{W}_* = \widehat{\gamma}_x^{-1}(\mathbf{W} - \widehat{\gamma}_0)$.

3.4.2 Orthogonal Regression

Another well publicized method for linear regression in the presence of measurement error is *orthogonal regression*; see Fuller (1987, Section 1.3.3). This is sometimes known as the linear statistical relationship (Tan and Iglewicz, 1999) or the linear functional relationship. However, for reasons given below, we are skeptical about the general utility of orthogonal regression, in large part because it is so easily misused. Although it is not fundamental to understanding later material on nonlinear models, we take the opportunity to discuss orthogonal regression at length here in order to emphasize the potential pitfalls associated with it. The work appeared as Carroll and Ruppert (1996), but the message is worth repeating. This section can be skipped by those who are interested only in estimation for nonlinear models or who plan never to use orthogonal regression.

Let $\mathbf{Y} = \beta_0 + \beta_x\mathbf{X} + \epsilon$ and $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where ϵ and $\mathbf{U}$ are uncorrelated. Whereas the method-of-moments estimator (Section 3.4) requires knowledge or estimability of the measurement error variance σ_u^2, orthogonal regression requires the same for the ratio $\eta = \sigma_\epsilon^2/\sigma_u^2$.

The orthogonal regression estimator minimizes the orthogonal distance of $(\mathbf{Y}, \mathbf{W})$ to the line $\beta_0 + \beta_x\mathbf{X}$, weighted by η, that is, it minimizes

$$\sum_{i=1}^{n} \left\{ (\mathbf{Y}_i - \beta_0 - \beta_x x_i)^2 + \eta\,(\mathbf{W}_i - x_i)^2 \right\} \qquad (3.14)$$

in the unknown parameters $(\beta_0, \beta_x, x_1, \ldots, x_n)$.

In fact, (3.14) is the sum of squared orthogonal distances between the points $(\mathbf{Y}_i, \mathbf{W}_i)_1^n$ and the line $y = \beta_0 + \beta_x x$, only in the special case that $\eta = 1$. However, the term orthogonal regression is used to describe the method regardless of the value of $\eta < \infty$.

The orthogonal regression estimator is the functional maximum likelihood estimator (Sections 2.1 and 7.1) assuming that $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ are unknown fixed constants, and that the errors $(\epsilon, \mathbf{U})$ are independent and normally distributed.

Orthogonal regression has the appearance of greater applicability than

$\mathbf{W}_i$	$\mathbf{Y}_{i1}$	$\mathbf{Y}_{i2}$
−1.8007	−0.5558	−0.9089
−0.7717	0.2076	0.6499
−0.4287	−1.7365	−1.8542
−0.0857	−0.9018	0.2040
0.2572	−0.2312	−0.3097
0.6002	0.2967	0.5072
0.9432	0.5928	1.5381
1.2862	1.2420	1.2599

Table 3.2 *Orthogonal regression example with replicated response.*

method-of-moments estimation in that only the ratio, η, of the error variances need be known or estimated. However, it is our experience that in the majority of problems η cannot be specified or estimated correctly, and use of orthogonal regression with an improperly specified value of η often results in an unacceptably large *overcorrection* for attenuation due to measurement error.

We illustrate the problem with some data from a consulting problem (Table 3.2). The data include two measurements of a response variable, $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$, and one predictor variable with true value $\mathbf{X}_i$, $i = 1, \ldots, 8$. The data are proprietary and we cannot disclose the nature of the application. Accordingly, all of the variables have been standardized to have sample means and variances 0 and 1, respectively.

We take as the response variable to be used in the regression analysis, $\mathbf{Y}_i = (\mathbf{Y}_{i1} + \mathbf{Y}_{i2})/2$, the average of the two response measurements.

Using an independent experiment, it had been estimated that $\sigma_u^2 \approx 0.0424$, also after standardization. Because the sample standard deviation of $\mathbf{W}$ is 1.0, measurement error induces very little bias here. The estimated reliability ratio is $\widehat{\lambda} = 1 - 0.0424 \approx 0.96$, and so attenuation is only about 4%. The ordinary least squares estimated slope from regressing the average of the responses on $\mathbf{W}$ is 0.65, while the method-of-moments slope estimate is $\widehat{\lambda}^{-1} 0.65 \approx 0.68$.

In a first analysis of these data, our client thought that orthogonal regression was an appropriate method for these data. A components-of-variance analysis resulted in the estimate 0.0683 for the response measurement error variance. If η is estimated by $\widehat{\eta} = 0.0683/0.0424 \approx 1.6118$, then the resulting orthogonal regression slope estimate is 0.88.

The difference in these two estimates, $|0.88 - 0.68|$, is larger than would be expected from random variation alone. Clearly, something is amiss. The method-of-moments correction for attenuation is only $\widehat{\lambda}^{-1} \approx 1.04$, whereas,orthogonal regression in effect, produces a correction for attenuation of approximately $1.35 \approx 0.88/0.65$.

The problem lies in the nature of the regression model error ϵ, which is typically the sum of two components: (i) ϵ_M, the measurement error in determination of the response; and (ii) ϵ_L, the *equation error*, that is, the variation about the regression line of the true response in the absence of measurement error. See Section 1.5 for another example of equation error, which in nutrition is called *person-specific bias*.

If we have replicated measurements, $\mathbf{Y}_{ij}$, of the true response, then $\mathbf{Y}_{ij} = \beta_0 + \beta_x \mathbf{X}_i + \epsilon_{L,i} + \epsilon_{M,ij}$, and of course their average is $\overline{\mathbf{Y}}_{i\cdot} = \beta_0 + \beta_x \mathbf{X}_i + \epsilon_{L,i} + \overline{\epsilon}_{M,i\cdot}$. Here and throughout the book, a subscript "dot" and overbar means averaging. For example, with k replicates,

$$\overline{\mathbf{Y}}_{i\cdot} = k^{-1} \sum_{j=1}^{k} \mathbf{Y}_{ij}; \quad \overline{\epsilon}_{M,i\cdot} = k^{-1} \sum_{j=1}^{k} \epsilon_{M,ij}.$$

The components of variance analysis estimates *only* the variance of the average measurement error $\overline{\epsilon}_{M,i\cdot}$ in the responses, but completely ignores the variability, $\epsilon_{L,i}$, about the line. The net effect is to underestimate η and thus overstate the correction required of the ordinary least squares estimate, because $\mathrm{var}(\overline{\epsilon}_{M,i\cdot})/\sigma_u^2$ is used as the estimate of η instead of the larger, appropriate value $\{\mathrm{var}(\overline{\epsilon}_{M,i\cdot}) + \mathrm{var}(\epsilon_{L,i})\}/\sigma_u^2$.

The naive use of orthogonal regression on the data in Table 3.2 has assumed that there is no additional variability about the line in addition to that due to measurement error in the response, that is, $\epsilon_{L,i} = 0$. To check this, refer to Figure 3.8. Each replicated response is indicated by a solid and filled circle. Remember that there is little measurement error in $\mathbf{W}$. In addition, the replication analysis suggested that the standard deviation of the replicates was less than 10% of the variability of the responses. Thus, in the absence of equation error we would expect to see the replicated pairs falling along a clearly delineated straight line. This is far from the case, suggesting that the equation error $\epsilon_{L,i}$ is a large part of the variability of the responses. Indeed, while the replication analysis suggests that $\mathrm{var}(\overline{\epsilon}_{M,i\cdot}) \approx 0.0683$, a method-of-moments analysis suggests $\mathrm{var}(\epsilon_{L,i}) \approx 0.4860$.

Fuller (1987) was one of the first to emphasize the importance of equation error. In our experience, outside of some special laboratory validation studies, equation error is almost always important in linear regression. In the majority of cases,orthogonal regression is an inappro-

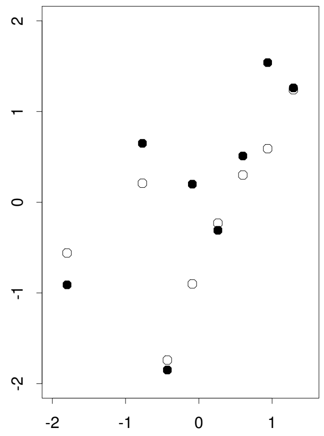


Figure 3.8 *Illustration where the assumptions of orthogonal regression appear violated. The filled and empty circles represent replicated values of the response. Note the evidence of equation error because the replicate responses are very close to each other, indicating little response measurement error, but the circles do not fall on a line, indicating some type of response error.*

priate technique, unless estimation of both the response measurement error and the equation error is possible.

In some cases, **Y** and **W** are measured in the same way, for example, if they are both blood pressure measurements taken at different times. Here, it is often entirely reasonable to *assume* that the variance of ϵ_M equals σ_u^2, and then there is a temptation to ignore equation error and hence set $\eta = 1$. Almost universally, this is a mistake: Equation error generally exists. This temptation is especially acute when replicates are absent, so that σ_u^2 cannot be estimated and the method-of-moments estimator cannot be used.

3.5 Bias Versus Variance

Estimates which do not account for measurement error are typically biased. Correcting for this bias entails what is often referred to as a *bias versus variance* tradeoff. What this means is that, in most problems, the very nature of correcting for bias is that the resulting corrected estimator will be more variable than the biased estimator. Of course, when an estimator is more variable, the confidence intervals associated with it are longer.

Later in this section we will describe theory, but it is instructive to consider an extreme case, using the same simulated data as in Figure 3.7 and Section 3.4.1. In this problem, the sample size is $n = 20$, the true slope is $\beta_x = 1.0$ and the reliability ratio is $\lambda = 0.5$. The top panel of Figure 3.9 gives the histogram of Fuller's modification of the method-of-moments estimator, while the bottom panel gives the histogram of the naive method that ignores measurement error. Note how the naive estimator is badly biased: Indeed, we know it estimates $\lambda\beta_x = 0.5$, and it is tightly bunched around this (wrong) value. The method-of-moments estimator is roughly unbiased, but this correction for bias is at the cost of a much greater variability (2.7 times greater in the simulation).

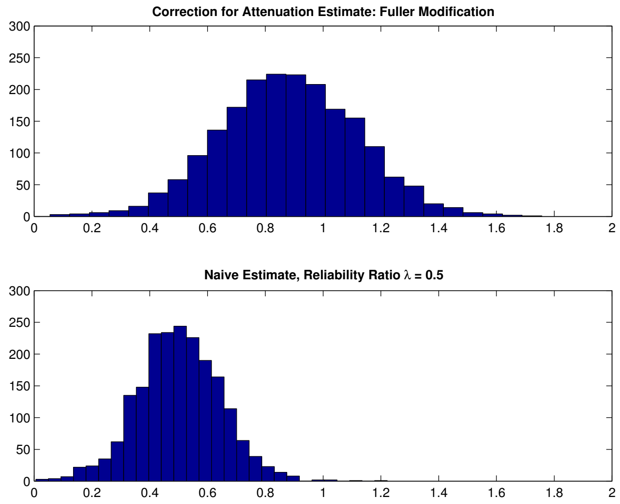


Figure 3.9 *Bias versus variance tradeoff in estimating the slope in simple linear regression. This is an extreme example of simple linear regression, with a sample size of $n = 20$ and a reliability ratio of $\lambda = 0.5$. The true value of the slope is $\beta_x = 1$. The top panel is Fuller's modification of the correction for attenuation estimate; the bottom is the naive estimate that ignores measurement error. The former is much more variable; the latter is very badly biased.*

3.5.1 Theoretical Bias–Variance Tradeoff Calculations

In this section, we will illustrate the bias versus variance tradeoff theoretically in simple linear regression. This material is somewhat technical, and readers may skip it without any loss of understanding of the main points of measurement error models.

Consider the simple linear regression model, $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, with additive independent measurement error, $\mathbf{W} = \mathbf{X} + \mathbf{U}$, under the simplifying assumption of joint normality of $\mathbf{X}$, $\mathbf{U}$, and ϵ. Further, suppose that the reliability ratio λ in (3.1) is known. We make this assumption only to simplify the discussion in this section. Generally, in applications it is seldom the case that this parameter is known, although there are exceptions (Fuller, 1987).

Let $\widehat{\beta}_{x*}$ denote the least squares estimate of slope from the regression of $\mathbf{Y}$ on $\mathbf{W}$. We know that its mean is $E(\widehat{\beta}_{x*}) = \lambda\beta_x$. Denote its variance by σ_*^2.

The method-of-moments estimator of β_x, is $\widehat{\beta}_{x,mm} = \lambda^{-1}\widehat{\beta}_{x*}$ and has mean $E(\widehat{\beta}_{x,mm}) = \beta_x$, and variance $\text{Var}(\widehat{\beta}_{x,mm}) = \lambda^{-2}\sigma_*^2$.

Because $\lambda < 1$, it is clear that while the correction-for-attenuation in $\widehat{\beta}_{x,mm}$ reduces its bias to zero, there is an increase in variability relative to the variance of the biased estimator $\widehat{\beta}_{x*}$. The variability is inflated even further if an estimate $\widehat{\lambda}$ is used in place of λ.

The price for reduced bias is increased variance. This phenomenon is not restricted to the simple model and estimator in this section, but occurs with almost universal generality in the analysis of measurement error models. In cases where the absence of bias is of paramount importance, there is usually no escaping the increase in variance. In cases where some bias can be tolerated, consideration of mean squared error is necessary.

In the following material, we indicate that there are compromise estimators that may outperform both uncorrected and corrected estimators, at least in small samples. Surprisingly, outside of the work detailed in Fuller (1987), such compromise estimators have not been much investigated, especially for nonlinear models.

Remember that mean squared error (MSE) is the sum of the variance plus the square of the bias. This is an interesting criterion to use, because uncorrected estimators have more bias but smaller variance than corrected estimators, and the bias versus variance tradeoff is transparent. Note that

$$\begin{aligned} \text{MSE}(\widehat{\beta}_{x*}) &= \sigma_*^2 + (1-\lambda)^2\beta_x^2; \text{ and} \\ \text{MSE}(\widehat{\beta}_{x,mm}) &= \lambda^{-2}\sigma_*^2. \end{aligned} \tag{3.15}$$

It follows that

$$\text{MSE}\left(\widehat{\beta}_{x,mm}\right) < \text{MSE}\left(\widehat{\beta}_{x*}\right)$$

if and only if

$$\sigma_*^2 < \frac{\lambda^2(1-\lambda)\beta_x^2}{1+\lambda}.$$

Because σ_*^2 decreases with increasing sample size, we can conclude that in sufficiently large samples it is always beneficial, in terms of mean squared error, to correct for attenuation due to measurement error.

Consider now the alternative estimator $\widehat{\beta}_{x,a} = a\beta_{x*}$ for a fixed constant a. The mean squared error of this estimator is $a^2\sigma_*^2 + (a\lambda - 1)^2\beta_x^2$, which is minimized when $a = a_* = \lambda\beta_x^2/(\sigma_*^2 + \lambda^2\beta_x^2)$. Ignoring the fact that a_* depends on unknown parameters, we consider the "estimator" $\widehat{\beta}_{x,*} = a_*\beta_{x*}$, which has smaller mean squared error than either $\widehat{\beta}_{x,mm}$ or $\widehat{\beta}_{x*}$. Note that as $\sigma_*^2 \to 0$, $a_* \to \lambda^{-1}$.

The estimator $\widehat{\beta}_{x,*}$ achieves its mean-squared-error superiority by making a partial correction for attenuation in the sense that $a_* < \lambda^{-1}$. This simple exercise illustrates that estimators that make only partial corrections for attenuation can have good mean-squared-error performance.

Although we have used a simple model and a somewhat artificial estimator to facilitate the discussion of bias and variance, all of the conclusions made above hold, at least to a very good approximation, in general for both linear and nonlinear regression measurement error models.

3.6 Attenuation in General Problems

We have already seen that, even in linear regression with multiple covariates, the effects of measurement error are complex and not easily described. In this section, we provide a brief overview of what happens in nonlinear models.

Consider a scalar covariate $\mathbf{X}$ measured with error, and suppose that there are no other covariates. In the classical error model for simple linear regression, we have seen that the bias caused by measurement error is always in the form of attenuation, so that ordinary least squares preserves the sign of the regression coefficient asymptotically, but is biased towards zero. Attenuation is a consequence then of (i) the simple linear regression model; and (ii) the classical additive error model. Without (i) and (ii), the effects of measurement error are more complex; we have already seen that attenuation may not hold if (ii) is violated.

In logistic regression when $\mathbf{X}$ is measured with additive error, attenuation does not always occur (Stefanski and Carroll, 1985), but it is typical. More generally, in most problems with a scalar $\mathbf{X}$ and no covariates $\mathbf{Z}$, the underlying *trend* between $\mathbf{Y}$ and $\mathbf{X}$ is preserved under nondifferential measurement error, in the sense that the correlation between $\mathbf{Y}$ and $\mathbf{W}$ is positive whenever both $E(\mathbf{Y}|\mathbf{X})$ and $E(\mathbf{W}|\mathbf{X})$ are increasing functions of $\mathbf{X}$ (Weinberg, Umbach, and Greenland, 1993). Technically, this follows because with nondifferential measurement error, $\mathbf{Y}$ and $\mathbf{W}$ are uncorrelated given $\mathbf{X}$, and hence the covariance between $\mathbf{Y}$ and $\mathbf{W}$ is just the covariance between $E(\mathbf{Y}|\mathbf{X})$ and $E(\mathbf{W}|\mathbf{X})$.

Positively, this result says that for the very simplest of problems (scalar **X**, no covariates **Z** measured without error), the general trend in the data is typically unaffected by nondifferential measurement error. However, the result illustrates only part of a complex picture, because it describes only the *correlation* between **Y** and **W** and says nothing about the structure of this relationship.

For example, one might expect that if the regression, $E(\mathbf{Y}|\mathbf{X})$, of **Y** on **X** is nondecreasing in **X**, and if $\mathbf{W} = \mathbf{X} + \mathbf{U}$ where **U** is independent of **X** and **Y**, then the regression of **Y** on **W** would also be nondecreasing. But Hwang and Stefanski (1994) have shown that this need not be the case, although it is true in linear regression with normally distributed measurement error. However, these results show that making inferences about details in the relationship of **Y** and **X**, based on the observed relationship between **Y** and **W**, is a difficult problem in general.

There are other practical reasons why ignoring measurement error is not acceptable. First, estimating the direction of the relationship between **Y** and **X** correctly is nice, but as emphasized by MacMahon et al. (1990) we can be misled if we severely underestimate its magnitude. Second, the result does not apply to multiple covariates, as we have noted in Figure 3.5 for the analysis of covariance and in Figure 3.6 for correlated measurement errors. Indeed, we have already seen that in multiple linear regression under the additive measurement error model, the observed and underlying trends may be entirely different. Finally, it is also the case (Sectiob 10.1) that, especially with multiple covariates, one can use error modeling to improve the power of inferences. In large classes of problems, then, there is simply no alternative to careful consideration of the measurement error structure.

Bibliographic Notes

The linear regression problem has a long history and continues to be the subject of research. Excellent historic background can be found in the papers by Lindley (1953), Lord (1960), Cochran (1968) and Madansky (1959). Furthermore technical analyses are given by Fuller (1980), Carroll and Gallo (1982, 1984), and Carroll et al. (1985). Diagnostics are discussed by Carroll and Spiegelman (1986, 1992) and Cheng and Tsai (1992). Robustness is discussed by Ketellapper and Ronner (1984), Zamar (1988, 1992), Cheng and van Ness (1988), and Carroll et al. (1993). Ganse, Amemiya, and Fuller (1983) discussed an interesting prediction problem. Hwang (1986) and Hasenabeldy et al. (1989) discuss problems with unusual error structure. Boggs et al. (1988) discussed computational aspects of orthogonal regression in nonlinear models.

CHAPTER 4

REGRESSION CALIBRATION

4.1 Overview

In this monograph we will describe two simple, generally applicable approaches to measurement error analysis: regression calibration in this chapter and simulation extrapolation (SIMEX) in Chapter 5.

The basis of regression calibration is the replacement of $\mathbf{X}$ by the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$. After this approximation, one performs a standard analysis. The simplicity of this algorithm disguises its power. As Pierce and Kellerer (2004) state, regression calibration "is widely used, effective (and) reasonably well investigated." Regression calibration shares with multiple imputation the advantage as Pierce and Kellerer note, "A great many analyses of the same cohort data are made for different purposes . . . it is very convenient that (once the replacement is made) essentially the same methods for ongoing analyses can be employed as if $\mathbf{X}$ were observed." Regression calibration is simple and potentially applicable to any regression model, provided the approximation is sufficiently accurate. SIMEX shares these advantages but is more computationally intensive.

Of course, with anything so simple, yet seemingly so general, there have to be some catches. These are:

- Estimating the basic quantity, the regression of $\mathbf{X}$ on $(\mathbf{W}, Z)$, is an art. After all, we do not observe $\mathbf{X}$! There is an active literature on this topic, reviewed in Sections 4.4 and 4.5.
- No simple approximation can always be accurate. Regression calibration tends to be most useful for generalized linear models (GLIM), helpful given the vast array of applications of these models. Indeed, in many GLIM, the approximation is exact or painfully close to being exact. We review this issue in Sections 4.8 and B.3.3.
- On the other hand, the regression calibration approximation can be rather poor for highly nonlinear models, although sometimes fixups are possible; see Section 4.7, where a unique application to a bioassay example is made.

The algorithm is given in Section 4.2. An example using the NHANES data is given in Section 4.3. Basic to the algorithm is a model for

$E(\mathbf{X}|\mathbf{Z},\mathbf{W})$, and methods of fitting such models are discussed in Sections 4.4 and 4.5. Section 4.6 provides brief remarks on calculating standard errors. The expanded regression calibration approximation in Section 4.7 attempts to improve the basic regression calibration approximation; the section includes a second example, the bioassay data. Sections 4.8 and 4.9 are devoted to theoretical justification of regression calibration and expanded regression calibration. Technical details, of which there are many, are relegated to Appendix B.3.

4.2 The Regression Calibration Algorithm

The regression calibration algorithm is as follows:

- Estimate the regression of $\mathbf{X}$ on $(\mathbf{Z},\mathbf{W})$, $m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma)$, depending on parameters γ, which are estimated by $\widehat{\gamma}$. How to do this is described in Sections 4.4 and 4.5.
- Replace the unobserved $\mathbf{X}$ by its estimate $m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\widehat{\gamma})$, and then run a standard analysis to obtain parameter estimates.
- Adjust the resulting standard errors to account for the estimation of γ, using either the bootstrap or sandwich method; consult Appendix A for the discussion of these techniques.

Suppose, for example, that the mean of $\mathbf{Y}$ given $(\mathbf{X},\mathbf{Z})$ can be described by

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = m_{\mathbf{Y}}(\mathbf{Z},\mathbf{X},\mathcal{B}) \tag{4.1}$$

for some unknown parameter $\mathcal{B}$. The replacement of $\mathbf{X}$ in (4.1) by its estimated value in effect proposes a modified model for the observed data, namely

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \approx m_{\mathbf{Y}}\left\{\mathbf{Z}, m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma),\mathcal{B}\right\}. \tag{4.2}$$

It is important to emphasize that the regression calibration model (4.2) is an approximate, working model for the observed data. It is not necessarily the same as the actual mean for the observed data, but in many cases is only modestly different. Even as an approximation, the regression calibration model can be improved; see Section 4.7 for refinements.

4.3 NHANES Example

The purpose of this section is to give an example of the application of regression calibration to logistic regression. In particular, we will illustrate the bias versus variance tradeoff exemplified by Figure 3.9 in Section 3.5.

We consider the analysis of the NHANES–I Epidemiologic Study Cohort data set (Jones, Schatzen, Green, et al., 1987). The predictor variables $\mathbf{Z}$ that are assumed to have been measured without appreciable error are age, poverty index ratio, body mass index, use of alcohol (yes–no), family history of breast cancer, age at menarche (a dummy variable taking on the value 1 if the age is ≤ 12), menopausal status (pre or post), and race. The variable measured with error, $\mathbf{X}$, is long-term average daily intake of saturated fat (in grams). The response is breast cancer incidence. The analysis in this section is restricted to $3,145$ women aged 25–50 with complete data on all the variables listed above; 59 had breast cancer. In general, logistic regression analyses with a small number of disease cases are very sensitive to misclassification, case deletion, etc.

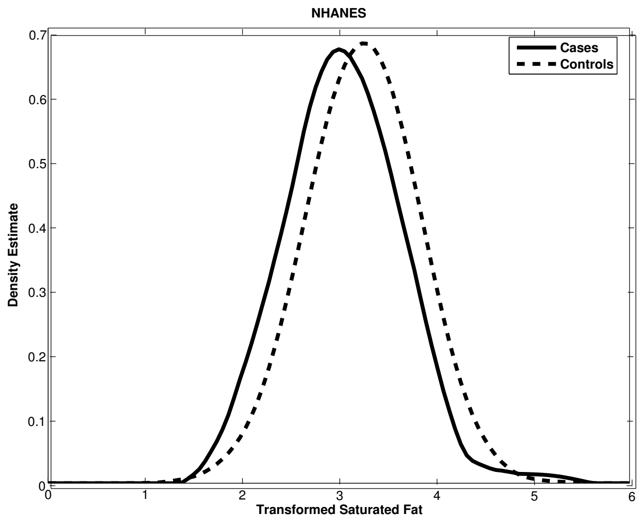


Figure 4.1 *Density estimates of transformed saturated fat for cases and controls: NHANES data.*

Saturated fat was measured via a 24-hour recall, that is, a participant's diet in the previous 24 hours was recalled and nutrition variables computed. It is measured with considerable error (Beaton, Milner, and Little, 1979; Wu, Whittemore, and Jung, 1986), leading to controversy as regards the use of 24-hour recall to assess breast cancer risk (Prentice, Pepe, and Self, 1989; Willett, Meir, Colditz, et al., 1987).

Our analysis concerns the effect of saturated fat on risk of breast cancer, adjusted for the other variables. To give a first indication of the effects, we considered the marginal effect of saturated fat. Specifically, we considered the variable log(5+saturated fat) and computed kernel density estimates (Silverman, 1986) of this variable for the breast cancer cases and for the noncases. The transformation was chosen for illustrative purposes and because it makes the observed values nearly normally distributed. The results are given in Figure 4.1. Note that this figure indicates a small marginal *but protective* effect due to higher levels of saturated fat in the diet, which is in opposition to one popular hypothesis. Thus we should expect the logistic regression coefficient of saturated fat to be negative (hence, the higher the levels of fat, the lower the estimated risk of breast cancer).

Variable	Estimate	Std. Error	p-value
Age /25	2.09	.53	$<$ .001
Poverty index	.13	.08	.10
Body mass index / 100	−1.67	2.55	.51
Alcohol	.42	.29	.14
Family history	.63	.44	.16
Age at menarche	−0.19	.27	.48
Premenopausal	.85	.43	.05
Race	.19	.38	.62
log(5 + saturated fat)	−0.97	.29	$<$ .001

Table 4.1 *Logistic regression in the NHANES data.*

In Table 4.1 we list the result of ignoring measurement error. This analysis suggests that transformed saturated fat is a highly significant predictor of risk with a negative logistic regression coefficient. Results in Chapter 10 show that the p-value is asymptotically valid because there are no other covariates measured with error.

There are at least two problems with these data that suggest that the results should be treated with extreme caution.

The first reason is that few epidemiologists would trust the results of a single 24-hour recall as a measure of long-term daily intake. The second is that if one also adds caloric intake into the model, something often done by epidemiologists, then the statistical significance for saturated fat seen in Table 4.1 disappears, with a p-value of 0.07.

By using data from the Continuing Survey of Food Intake by Individuals (CSFII, see Thompson, Sowers, Frongillo, et al., 1992), we estimate that over 75% of the variance of a single 24-hour recall is made up of measurement error. This analysis is fairly involved and was discussed in too much detail in the first edition: Here, we simply take the estimate as given, namely, that the observed sample variance of $\mathbf{W}$ is 0.233,

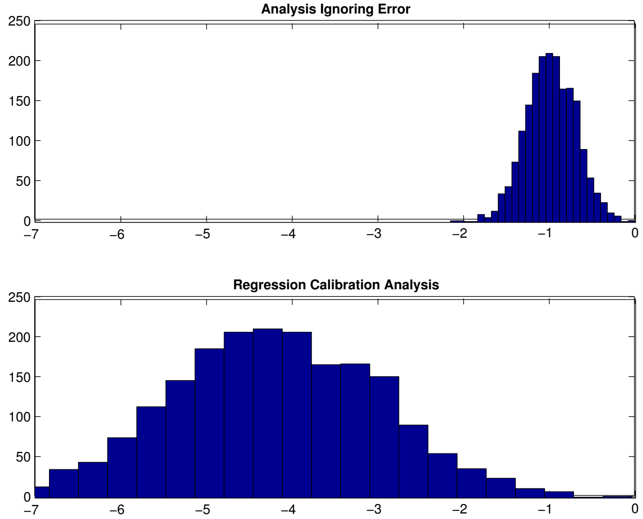


Figure 4.2 *Bootstrap analysis of the estimated coefficient ignoring measurement error (top panel) and accounting for it via regression calibration. Note how the effect of measurement error is to attenuate the coefficient, and the effect of correcting for measurement error is to widen confidence intervals. Compare with Figure 3.9.*

and for the additive measurement error model, the measurement error variance is estimated as $\widehat{\sigma}_u^2 = 0.171$. This error variance estimate is relatively close to the value 0.143 formed using a components-of-variance estimator given by (4.3) below when applied to 24-hour recalls in the American Cancer Society Cancer Prevention Study II (CPS II) Nutrition Survey Validation Study, which has $n = 184$ individuals with four 24-hour recalls per individual.

We applied regression calibration to these data, using the "resampling pairs" bootstrap (Section A.9.2) to get estimated standard errors. The parameter estimate was −4.67 with an estimated variance of 2.26, along with an associated percentile 95% confidence interval from −10.37 to −1.38. What might be most interesting is the results from this bootstrap, given as histograms in Figure 4.2. There we see the bias versus variance tradeoff exemplified by Figure 3.9 in Section 3.5. Specifically, note how the bootstrap, when ignoring measurement error, is tightly bunched around a far too small estimated value, while the bootstrap accounting for measurement error is centered at a very different place and with much more variability.

4.4 Estimating the Calibration Function Parameters

4.4.1 Overview and First Methods

The basic point of using the regression calibration approximation is that one runs a favorite analysis with $\mathbf{X}$ replaced by its regression on $(\mathbf{Z},\mathbf{W})$. In this section, we discuss how to do this regression.

There are two simple cases:

- With *internal validation data*, the simplest approach is to regress $\mathbf{X}$ on the other covariates $(\mathbf{Z},\mathbf{W})$ in the validation data. Of course, this is a missing data problem, and generally one would then use missing data techniques rather than regression calibration. Regression calibration in this instance is simply a poor person's imputation methodology. From a practical matter, for a quick analysis we suggest that one use the $\mathbf{X}$ data where it is available, but add in a dummy variable to distinguish between the cases that $\mathbf{X}$ or its regression calibration versus are used.
- In some problems, for example, in nutritional epidemiology, an *unbiased instrument* $\mathbf{T}$ is available for a subset of the study participants; see Section 2.3. Here, by definition of "unbiased instrument," the regression of $\mathbf{T}$ on $(\mathbf{Z},\mathbf{W})$ is the same as what we want, the regression of $\mathbf{X}$ on $(\mathbf{Z},\mathbf{W})$. This is the method used by Rosner, Spiegelman and Willett (1990) in their analysis of the Nurses' Health Study, see Section 1.6.2. In that study, health outcomes and dietary intakes as measured by a food frequency questionnaire (FFQ) $\mathbf{W}$ were observed on all study participants. On a subset of the study participants, dietary intakes were assessed by food diaries, $\mathbf{T}$. The investigators assumed that the diaries were unbiased for usual dietary intake and applied regression calibration by regressing the intakes from diaries on those from the FFQ.

With validation data or an unbiased instrument, models for $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$ can be checked by ordinary regression diagnostics such as residual plots.

4.4.2 Best Linear Approximations Using Replicate Data

Here we consider the classical additive error model $\mathbf{W}=\mathbf{X}+\mathbf{U}$ where conditional on $(\mathbf{Z},\mathbf{X})$ the errors have mean zero and constant covariance matrix Σ_{uu}. We describe an algorithm yielding a linear approximation to the regression calibration function. The algorithm is applicable when Σ_{uu} is estimated via external data or via internal replicates. The method was derived independently by Carroll and Stefanski (1990) and Gleser (1990), and used by Liu and Liang (1992) and Wang, Carroll, and Liang (1996).

In this subsection, we will discuss using replicates measurements of $\mathbf{X}$, that is, replicated $\mathbf{W}$ measuring the same $\mathbf{X}$. When necessary, the convention made in this book is to adjust the replicates a priori so that they have the same sample means.

Suppose there are k_i replicate measurements, $\mathbf{W}_{i1},\ldots,\mathbf{W}_{ik_i}$, of $\mathbf{X}_i$, and $\overline{\mathbf{W}}_{i\cdot}$ is their mean. Replication enables us to estimate the measurement error covariance matrix Σ_{uu} by the usual components of variance analysis, as follows:

$$\widehat{\Sigma}_{uu}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\mathbf{W}_{ij}-\overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij}-\overline{\mathbf{W}}_{i\cdot}\right)^t}{\sum_{i=1}^{n}(k_i-1)}. \tag{4.3}$$

In (4.3), remember that we are using the "dot and overbar" notation to mean averaging over the "dotted" subscript.

Write Σ_{ab} as the covariance matrix between two random variables, and let μ_a be the mean of a random variable. The best linear approximation to $\mathbf{X}$ given $(\mathbf{Z},\overline{\mathbf{W}})$ is

$$\begin{aligned}E(\mathbf{X}|\mathbf{Z},\overline{\mathbf{W}})&\approx\mu_x &(4.4)\\ &+\begin{pmatrix}\Sigma_{xx}\\ \Sigma_{zx}\end{pmatrix}^t\begin{bmatrix}\Sigma_{xx}+\Sigma_{uu}/k & \Sigma_{xz}\\ \Sigma_{xz}^t & \Sigma_{zz}\end{bmatrix}^{-1}\begin{pmatrix}\overline{\mathbf{W}}-\mu_w\\ \mathbf{Z}-\mu_z\end{pmatrix}.\end{aligned}$$

Here is how one can operationalize (4.4) based on observations $(\mathbf{Z}_i,\overline{\mathbf{W}}_{i\cdot})$, replicate sample sizes k_i and estimated error covariance matrix $\widehat{\Sigma}_{uu}$. We use analysis of variance formulae. Let

$$\begin{aligned}\widehat{\mu}_x&=\widehat{\mu}_w=\sum_{i=1}^{n}k_i\overline{\mathbf{W}}_{i\cdot}/\sum_{i=1}^{n}k_i;\qquad \widehat{\mu}_z=\overline{\mathbf{Z}}_\cdot;\\ \nu&=\sum_{i=1}^{n}k_i-\sum_{i=1}^{n}k_i^2/\sum_{i=1}^{n}k_i;\\ \widehat{\Sigma}_{zz}&=(n-1)^{-1}\sum_{i=1}^{n}\left(\mathbf{Z}_i-\overline{\mathbf{Z}}_\cdot\right)\left(\mathbf{Z}_i-\overline{\mathbf{Z}}_\cdot\right)^t;\\ \widehat{\Sigma}_{xz}&=\sum_{i=1}^{n}k_i\left(\overline{\mathbf{W}}_{i\cdot}-\widehat{\mu}_w\right)\left(\mathbf{Z}_i-\overline{\mathbf{Z}}_\cdot\right)^t/\nu;\\ \widehat{\Sigma}_{xx}&=\left[\left\{\sum_{i=1}^{n}k_i\left(\overline{\mathbf{W}}_{i\cdot}-\widehat{\mu}_w\right)\left(\overline{\mathbf{W}}_{i\cdot}-\widehat{\mu}_w\right)^t\right\}-(n-1)\widehat{\Sigma}_{uu}\right]/\nu.\end{aligned}$$

The resulting estimated calibration function is

$$\begin{aligned}E(\mathbf{X}_i|\mathbf{Z}_i,\overline{\mathbf{W}}_{i\cdot})&\approx\widehat{\mu}_w &(4.5)\\ &+(\widehat{\Sigma}_{xx},\widehat{\Sigma}_{xz})\begin{bmatrix}\widehat{\Sigma}_{xx}+\widehat{\Sigma}_{uu}/k_i & \widehat{\Sigma}_{xz}\\ \widehat{\Sigma}_{xz}^t & \widehat{\Sigma}_{zz}\end{bmatrix}^{-1}\begin{pmatrix}\overline{\mathbf{W}}_{i\cdot}-\widehat{\mu}_w\\ \mathbf{Z}_i-\overline{\mathbf{Z}}_\cdot\end{pmatrix}.\end{aligned}$$

In linear regression, if there are no replicates $(k_i\equiv 1)$ but an external estimate $\widehat{\Sigma}_{uu}$ is available, or if there are exactly two replicates ($k_i \equiv 2$), in which case $\widehat{\Sigma}_{uu}$ is half the sample covariance matrix of the differences $\mathbf{W}_{i1} - \mathbf{W}_{i2}$, regression calibration reproduces the classical method-of-moments estimates, that is, the estimators (3.13) of Section 3.4 with Σ_{uu} estimated from replicates and $\Sigma_{\epsilon u}$ assumed to be 0.

When the number of replicates is not constant, the algorithm can be shown to produce consistent estimates in linear regression and (approximately!) in logistic regression. For loglinear mean models, the intercept is biased, so one should add a dummy variable to the regression indicating whether or not an observation is replicated.

4.4.3 Alternatives When Using Partial Replicates

The linear approximations defined above are only approximations, but they can be checked by using the replicates themselves. As is typical, if only a partial subset of the study has an internal replicate ($k_i = 2$), while most of the data are not replicated ($k_i = 1$), the partial replicates can be used to check the best linear approximations to $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ defined above, by fitting models to the regression of $\mathbf{W}_{i2}$ on $(\mathbf{Z}_i, \mathbf{W}_{i1})$. If necessary, the partial replication data can be used in this way to estimate $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$. A good picture to study is Figure 1.2, where we plot the log protein biomarkers against one another. The scatterplot suggest a linear relationship, and a linear model here seems perfectly reasonable.

4.4.4 James–Stein Calibration

Whittemore (1989) also proposed regression calibration in the case that $\mathbf{X}$ is scalar, there is no $\mathbf{Z}$, and the additive error model applies. If σ_u^2 is unknown and there are k replicates at each observation, then instead of the method-of-moments estimate (4.5) of $E(\mathbf{X}|\mathbf{W})$, she suggested use of the James–Stein estimate, namely

$$\overline{\mathbf{W}}_{\cdot\cdot} + \left\{1 - \frac{n-1}{n-3}\frac{n(k-1)}{n(k-1)+2}\frac{\widehat{\sigma}_u^2/k}{\widehat{\sigma}_{\overline{w}}^2}\right\}(\overline{\mathbf{W}}_{i\cdot} - \overline{\mathbf{W}}_{\cdot\cdot}),$$

where $\widehat{\sigma}_u^2$ is the usual components of variance estimate of σ_u^2 defined in (4.3) and $\widehat{\sigma}_{\overline{w}}^2$ is the sample variance of the terms $(\overline{\mathbf{W}}_{i\cdot})$. Typically, the James–Stein and moments estimates are nearly the same.

4.5 Multiplicative Measurement Error

Until now we have assumed that the measurement errors are additive, but multiplicative errors are common and require special treatment. Multiplicative errors can be converted to additive ones by a log transformation, and first we discuss when a log transformation should be used. Then we introduce alternative strategies for use when a log transformation seems inappropriate.

4.5.1 Should Predictors Be Transformed?

A good example of multiplicative measurement error can be seen in Figures 1.6, 1.7, and 1.8, as described in Section 1.7. This is a case where taking logarithms seems to lead to an additive measurement error model with constant measurement error variance. Other scientists also find multiplicative measurement errors. Lyles and Kupper (1997) state that “there is much evidence for this model” (meaning the multiplicative error model). Pierce, Stram, Vaeth, et al. (1992) study data from Radiation Effects Research Foundation (RERF) in Hiroshima. They state, “It is accepted that radiation dose–estimation errors are more homogeneous on a multiplicative than on an additive scale.” Hwang (1986) studied data on energy consumption and again found multiplicative errors.

While the existence of multiplicative measurement errors is in no doubt, what to do in that situation is a matter of some controversy. Indeed, this has nothing to do with measurement error: taking the logarithm of a predictor is a perfectly traditional way to lessen the effects of leverage in regression. Thus, many authors simply use the transformed data scale. In most of the nutrition examples of which we are aware, investigators use the transformed predictor as $\mathbf{W}$ and carry out analyses: It is trivial then to construct relative risks from the lowest to highest quintiles of a nutrient. Alternatively, they often categorize the observed data into quintiles and run a test for trend against the quintile indicators. It would not be typical to run logistic regression analyses on the original scale data, which are often horribly skew. As we will see, multiplicative measurement error generally means that the largest observed values are very far from the actual values, often an order of magnitude in difference. These considerations dictate against using the original data scale when running an analysis that ignores measurement error.

There are, however, many researchers who prefer to fit a regression model in the original scale, rather than in a transformed scale. We tend to have little sympathy, in the absence of data analysis, for assertions of the type that scientifically one scale is to be preferred. However, it is important to have the flexibility to fit measurement error models on the original data scale. In this section, we describe how to implement regression calibration in the multiplicative context.

4.5.2 Lognormal **X** *and* **U**

In this section, among other things, we will show that in linear regression the effect of multiplicative measurement error is to make the observed untransformed data appear as if they are curved, not linear.

The multiplicative lognormal error model with a scalar $\mathbf{X}$ and an unbiased version of it is

$$\mathbf{W} = \mathbf{X}\,\mathbf{U}, \qquad \log(\mathbf{U}) \sim \text{Normal}\{-(1/2)\sigma_u^2, \sigma_u^2\}. \tag{4.6}$$

For simplicity, we assume that there are no covariates $\mathbf{Z}$ measured without error. If $\mathbf{X}$ is also lognormal and independent of $\mathbf{U}$, then regression calibration takes a simple form. Let $\mu_{w,\log}$ and $\sigma^2_{w,\log}$ be the mean and variance of $\log(\mathbf{W})$, respectively. Let

$$\lambda = \frac{\text{var}\{\log(\mathbf{X})\}}{\text{var}\{\log(\mathbf{W})\}} = \frac{\sigma^2_{w,\log} - \sigma^2_u}{\sigma^2_{w,\log}},$$

and $\alpha = \mu_{w,\log}(1-\lambda) + (1/2)\sigma_u^2$. Then by (A.2) and (A.3)

$$\begin{aligned} E(\mathbf{X}|\mathbf{W}) &= \mathbf{W}^{\lambda} \exp\left(\alpha + \lambda\sigma_u^2/2\right). & (4.7)\\ \text{var}(\mathbf{X}|\mathbf{W}) &= \mathbf{W}^{2\lambda} \exp\left(2\alpha\right)\left\{\exp(2\lambda\sigma_u^2) - \exp(\lambda\sigma_u^2)\right\}. & (4.8) \end{aligned}$$

Replacing $\mu_{w,\log}$ and $\sigma^2_{w,\log}$ by the sample mean and variance of $\log(\mathbf{W})$, and plugging these values into the forms for α and λ allows one to implement regression calibration using (4.7). Of course, one needs an estimate of σ_u^2 as well. This parameter can be estimated using validation or replication data by the methods discussed in Section 4.4, but applied to $\log(\mathbf{W})$, $\log(\mathbf{U})$, and $\log(\mathbf{X})$. Note how the regression calibration function is nonlinear in $\mathbf{W}$.

The way to derive (4.7) and (4.8) is modestly amusing. Take logarithms of both sides of (4.6) to get $\log(\mathbf{W}) = \log(\mathbf{X}) + \log(\mathbf{U})$, and then use equation (A.9) of Appendix A.4 to find that $\log(\mathbf{X}) = \alpha + \lambda\log(\mathbf{W}) + \mathbf{V}$ where $\mathbf{V}$ is Normal$(0, \lambda\sigma_u^2)$, and finally

$$\mathbf{X} = \mathbf{W}^{\lambda}\exp\{\alpha + \mathbf{V}\},$$

from which (4.7) and (4.8) follow from standard moment generating properties; see (A.2) of the appendix.

One of the exciting consequences of (4.7) is that if the regression of $\mathbf{Y}$ on $\mathbf{X}$ is linear in $\mathbf{X}$, say $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_x\mathbf{X}$, then the regression of $\mathbf{Y}$ on the observed $\mathbf{W}$ is nonlinear, that is, by (4.7)

$$E(\mathbf{Y}|\mathbf{W}) = \beta_0 + \left\{\beta_x \exp\left(\alpha + \lambda\sigma_u^2/2\right)\right\}\mathbf{W}^{\lambda}.$$

Therefore, to obtain an asymptotically unbiased slope estimate, one regresses $\mathbf{Y}$ on $\mathbf{W}^{\lambda}\exp\left(\alpha + \lambda\sigma_u^2/2\right)$. Note that the regression of $\mathbf{Y}$ on $\mathbf{W}$ is not linear, even though the regression of $\mathbf{Y}$ on $\mathbf{X}$ is linear. This is because the regression of $\mathbf{X}$ on $\mathbf{W}$ is not linear.

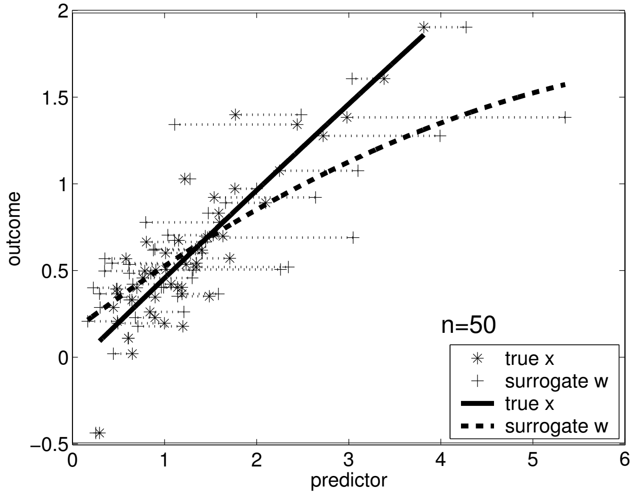


Figure 4.3 *Simulation of multiplicative measurement error:* $\mathbf{W} = \mathbf{X}\mathbf{U}$*,* $\log(\mathbf{X}) = \text{Normal}(0, 1/4)$, $\log(\mathbf{U}) = \text{Normal}(-1/8, 1/4)$, $\mathbf{Y} = (1/2)\mathbf{X} + \text{Normal}(0, 0.04)$, $n = 50$ *observations. The solid line is the fit to the unobserved* $\mathbf{X}$ *data in asterisks, while the dashed line is the spline fit to the observed* $\mathbf{W}$ *data in plus signs. Note how the multiplicative error has induced a curve into what should have been a straight line. Note too the stretching effect of the measurement error.*

Figure 4.3 shows 50 observations of simulated data with multiplicative lognormal measurement errors and $\mathbf{Y}$ linear in $\mathbf{X}$. The data with the true $\mathbf{X}$ values are plotted with asterisks and the data with the surrogates $\mathbf{W}$ are plotted with pluses. A dotted line connects $(\mathbf{Y}_i, \mathbf{X}_i)$ to $(\mathbf{Y}_i, \mathbf{W}_i)$ for each $i = 1, \ldots, n$. Penalized splines (Ruppert, Wand, and Carroll, 2003) were fit to true covariates and the surrogates and plotted as solid and dashed lines, respectively. Notice that, as theory predicts, the spline fit to the true covariates is linear but the spline fit to the surrogates is curved. One can also see attenuation; the derivative (slope) of the curved fit to the surrogates is seen to be everywhere less than the slope of the straight line fit to the true covariates.

Figure 4.4 shows 1,000 observations of simulated data from the same joint distribution as in Figure 4.3. Only $(\mathbf{Y}_i, \mathbf{W}_i)$ data are plotted, but penalized spline fits are shown to both $(\mathbf{Y}_i, \mathbf{W}_i)$ and $(\mathbf{Y}_i, \mathbf{X}_i)$. Figures 4.3 and 4.4 are similar, but, due to the larger sample size, the latter has more extreme $\mathbf{W}$ values and shows the curvature of $E(\mathbf{Y}|\mathbf{W})$ more dramatically.

Figure 4.5 uses the simulated data in Figure 4.4 and shows a plot of $\mathbf{Y}$

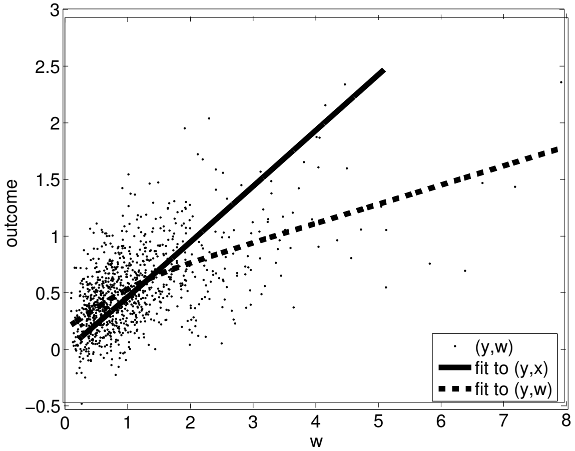


Figure 4.4 *Simulation of multiplicative measurement error:* $n = 1000$*,* $\mathbf{W} = \mathbf{XU}$*,* $\log(\mathbf{X}) = \text{Normal}(0, 1/4)$*,* $\log(\mathbf{U}) = \text{Normal}(-1/8, 1/4)$ $\mathbf{Y} = \mathbf{X}/2 + \text{Normal}(0, 0.04)$*. The solid line is the fit to the unobserved* $\mathbf{X}$ *data in asterisks, while the dashed line if the spline fit to the observed* $\mathbf{W}$ *data in plus signs.*

versus $\mathbf{W}^\lambda$, with $\lambda = 1/2$, and a penalized spline fit to it. As predicted by the theory, the regression of $\mathbf{Y}$ on $\mathbf{W}^\lambda$ appears linear.

In this context, regression calibration means replacing the unknown $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{W})$ given by (4.7). This is nonlinear regression calibration, since $E(\mathbf{X}|\mathbf{W})$ is nonlinear in $\mathbf{W}$. Notice by formula (4.8) for var$(\mathbf{X}|\mathbf{W})$ that, except when $\beta_x = 0$, the regression of $\mathbf{Y}$ on $\mathbf{W}$ is heteroscedastic even if the regression of $\mathbf{Y}$ on $\mathbf{X}$ is homoscedastic. In the presence of heteroscedasticity, ordinary unweighted least-squares is inefficient and, to gain efficiency, statisticians often use quasilikelihood; see Section A.7. Lyles and Kupper (1997) proposed a quasilikelihood estimator that was somewhat superior to nonlinear regression calibration in their simulation study, especially when the covariate measurement error is large.

In this section, we have focused on the case when $\mathbf{X}$ is lognormal, because of the simplicity of the expressions. The methods described above should work reasonably well if the unobserved covariate $\mathbf{X}$ is roughly lognormal, but there are no sensitivity studies done to date to confirm this.

The estimator of $E(\mathbf{X}|\mathbf{W})$ in this section assumes that both $\mathbf{X}$ and $\mathbf{U}$ are lognormally distributed. Pierce and Kellerer (2004) have described a method based on a Laplace approximation that is more nonparametric for the estimation of $E(\mathbf{X}|\mathbf{W})$.

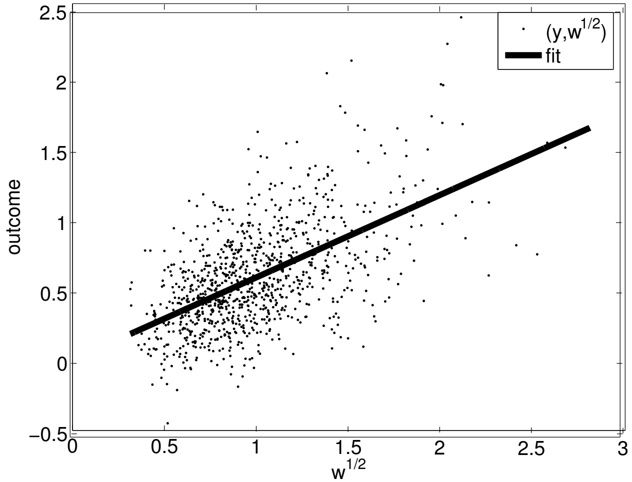


Figure 4.5 *Simulation of multiplicative measurement error:* $n = 1000$*,* $W = XU$*,* $\log(X) = \text{Normal}(0, 1/4)$*,* $\log(U) = \text{Normal}(-1/8, 1/4)$ $Y = X/2 + \text{Normal}(0, 0.04)$*. Note:* $\lambda = 1/2$*. The line is a penalized spline fit to the regression of* $\mathbf{Y}$ *on* $\mathbf{W}^\lambda$*, the regression calibration approximation. Theory predicts that this should be linear.*

4.5.3 Linear Regression

Fuller (1984) and Hwang (1986) independently developed a method-of-moments correction for multiplicative measurement error in linear regression. They make no assumptions that either the measurement errors or the true predictors are lognormal.

Their basic idea is to regress $\mathbf{Y}$ on $\mathbf{W}$ (not $\mathbf{W}^\lambda$) and then make a method-of-moments correction similar to (3.13):

$$\begin{pmatrix} S_{ww}./M_{uu} & S_{wz} \\ S_{zw} & S_{zz} \end{pmatrix}^{-1} \begin{pmatrix} S_{wy} - \Sigma_{u\epsilon} \\ S_{zy} \end{pmatrix}, \tag{4.9}$$

where $A./B$ is coordinate-wise division of equal-size matrices A and B, that is, $(A./B)_{ij} = A_{ij}/B_{ij}$, M_{uu} is the second moment matrix of $\mathbf{U}$, S_{ww} is the sum of cross-products matrix for $\mathbf{W}$, and so forth. In the following $\Sigma_{u\epsilon}$ is assumed to be zero. The Fuller–Hwang method is called the “correction method” by Lyles and Kupper (1997) and is similar to linear regression calibration defined in Section 4.2, because both methods are based on regressing $\mathbf{Y}$ on either $\mathbf{W}$ itself (Fuller–Hwang method) or a linear function of $\mathbf{W}$ (regression calibration). In fact, for scalar $\mathbf{X}$

the Fuller–Hwang estimator is the same as linear regression calibration when the calibration function predicts $\mathbf{X}$ using a linear function of $\mathbf{W}$ without an intercept, that is, a function of form $\lambda\mathbf{W}$, as discussed in Section A.4.2. As shown in that section, $\lambda = 1/E(\mathbf{U}^2)$, and therefore one can show that both the Fuller–Hwang method and regression calibration without an intercept multiply the ordinary least-squares slope estimate by $E(\mathbf{U}^2)$. Thus, the Fuller–Hwang estimator uses a less accurate predictor of X than linear regression calibration when the calibration function is allowed an intercept, which suggests that the Fuller–Hwang method might be inferior to the latter. The two estimators have apparently not been compared, possibly because nonlinear regression calibration seems more appropriate than either of them.

The Fuller–Hwang estimator is consistent but was found to be badly biased in simulations of Lyles and Kupper (1999). The bias is still noticeable for $n = 10,000$ in their simulations. Iturria, Carroll, and Firth (1999) found similar problems when linear regression calibration is used with multiplicative errors.

In addition, Iturria, Carroll, and Firth (1999) studied polynomial regression with multiplicative error. One of their general methods is a special case of the Fuller–Hwang estimator. Their "partial regression" estimator assumes lognormality of $(\mathbf{X}, \mathbf{U})$ and generalizes the nonlinear regression calibration estimator discussed earlier. In a simulation with lognormal $\mathbf{X}$ and $\mathbf{U}$, the partial regression estimator is often much more efficient than the ones that do not assume lognormality. For all these reasons, we favor our approaches over those of the Fuller–Hwang estimator.

The regression calibration methods of this section are not the only ways of handling multiplicative errors. For example, the Bayesian analysis of multiplicative error is discussed in Section 9.5.3.

4.5.4 Additive and Multiplicative Error

A model with both additive and multiplicative error is $\mathbf{W} = \mathbf{X}\mathbf{U}_1 + \mathbf{U}_2$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are independent errors with variances $\sigma^2_{u,1}$ and $\sigma^2_{u,2}$, respectively. This model implies that $\text{var}(\mathbf{W}|\mathbf{X}) = \mathbf{X}^2\sigma^2_{u,1} + \sigma^2_{u,2}$. For sufficiently small values of $\mathbf{X}$, $\text{var}(\mathbf{W}|\mathbf{X}) \approx \sigma^2_{u,2}$, while for sufficiently large values of $\mathbf{X}$ $\text{var}(\mathbf{W}|\mathbf{X}) \approx \mathbf{X}^2\sigma^2_{u,1}$. This model has been studied by Rocke and Durbin (2001) and applied by them to gene expression levels measured by cDNA slides. As far as we are aware, this model has not been applied as a measurement error model in regression, but research on this topic seems well worthwhile. A Berkson model that contains a mixture of additive and multiplicative errors has been proposed by Stram and Kopecky (2003) in their study of the Hanford Thyroid Disease Study.

4.6 Standard Errors

It is possible to provide asymptotic formulae for standard errors (Carroll and Stefanski, 1990), but doing so is extremely tedious because of the multiplicity of special cases. Some explicit formulae are given in the appendix (Section B.3.1) for the case of generalized linear models, and for models in which one specifies only the mean and variance of the response given the predictors.

The bootstrap (Section A.9) requires less programming (and mathematics!) but takes more computer time. In the first edition, we remarked that this can be a real issue because, as Donna Spiegelman has pointed out, many researchers would prefer to have quick standard errors instead of having to use the bootstrap repeatedly while building models for their data. However, faster computers and better software are reducing the time needed to perform bootstrap inference. For example, the rcal function in STATA uses the bootstrap to obtain standard errors in "real time," 1 minute for the ARIC data set.

In its simplest form, the bootstrap can be used to form standard error estimates, and then t-statistics can be constructed using the bootstrap standard errors. The bootstrap percentile method can be used for confidence intervals. Approximate bootstrap pivots can be formed by ignoring the variability in the estimation of the calibration function.

4.7 Expanded Regression Calibration Models

A major purpose of regression calibration is to derive an approximate model for the observed $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ data in terms of the fundamental model parameters. The regression calibration method is one means to this end: Merely replace $\mathbf{X}$ by an estimate of $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$. This method works remarkably well in problems such as generalized linear models, for example, linear regression, logistic regression, Poisson and gamma regression with loglinear links, etc. However, it is often not appropriate for highly nonlinear problems.

It is convenient for our purposes to cast the problems in the form of what are called mean and variance models, often called quasilikelihood and variance function (QVF) models, which are described in more generality and detail in (A.35) and (A.36). Readers unfamiliar with the ideas of quasilikelihood may wish to skip this material at first reading and continue into later chapters.

Mean and variance models specify the mean and variance of a response

$\mathbf{Y}$ as functions of covariates $(\mathbf{X}, \mathbf{Z})$ and unknown parameters. For example, in linear regression, the mean is a linear function of the covariates, and the variance is constant. We write these models in general as

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) &= m_{\mathbf{Y}}(\mathbf{Z},\mathbf{X},\mathcal{B}) && (4.10)\\ \mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) &= \sigma^2\, g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta), && (4.11)\end{aligned}$$

where $g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)$ is some nonnegative function and σ^2 is a scale parameter. The parameter vector θ contains parameters in addition to $\mathcal{B}$ that specify the variance function. In some models, for example, linear, logistic, Poisson and gamma regression, θ is not needed, since there are no additional parameters.

Of course, since $\mathbf{X}$ is not observed, to fit a mean and variance function model, what we need is the mean and variance of $\mathbf{Y}$ given the *observed* data. There are two possible approaches:

- Posit a probability model for the distribution of $\mathbf{X}$ given $(\mathbf{Z},\mathbf{W})$, then compute, *exactly*, $E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = E\{m_{\mathbf{Y}}(\mathbf{Z},\mathbf{X},\mathcal{B})|\mathbf{Z},\mathbf{W}\}$ and
$$\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = \mathrm{var}\{m_{\mathbf{Y}}(\mathbf{Z},\mathbf{X},\mathcal{B})|\mathbf{Z},\mathbf{W}\} + \sigma^2 E\{g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)|\mathbf{Z},\mathbf{W}\}.$$
- Instead of a probability model for the entire distribution, posit a model for the mean and variance of $\mathbf{X}$ given $(\mathbf{Z},\mathbf{W})$ and then do Taylor series expansions to estimate the mean and variance of the response given the observed data. These are the expanded regression calibration approximations.

Regression calibration, in effect, says that $\mathbf{X}$ given $(\mathbf{Z},\mathbf{W})$ is completely specified, with no error, by its mean, so that

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx m_{\mathbf{Y}}\{\mathbf{Z}, m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma),\mathcal{B}\}\,; && (4.12)\\ \mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx \sigma^2 g^2\{\mathbf{Z}, m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma),\mathcal{B},\theta\}\,. && (4.13)\end{aligned}$$

We will show that in some cases, the model can be modified to improve the fit; see Section 4.7.3 for a striking data application.

An example will help explain the possible need for refined approximations. Consider the simple linear homoscedastic regression model $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_x\mathbf{X}$ and $\mathrm{var}(\mathbf{Y}|\mathbf{X}) = \sigma^2$. Suppose the measurement process induces a heteroscedastic Berkson model where $E(\mathbf{X}|\mathbf{W}) = \mathbf{W}$ and $\mathrm{var}(\mathbf{X}|\mathbf{W}) = \sigma^2_{\mathrm{rc}}\mathbf{W}^{2\gamma}$, where *rc* stands for *regression calibration*. The regression calibration approximate model states that the observed data follow a simple linear homoscedastic regression model with $\mathbf{X}$ replaced by $E(\mathbf{X}|\mathbf{W}) = \mathbf{W}$. However, while this gives a correct mean function, the actual variance function for the observed data is heteroscedastic: $\mathrm{var}(\mathbf{Y}|\mathbf{W}) = \sigma^2 + \sigma^2_{\mathrm{rc}}\beta_x^2\mathbf{W}^{2\gamma}$. Hence the regression calibration model gives a consistent estimate of the slope and intercept, but the estimate is inefficient because weighted least squares should have been used. If it is important enough to affect the efficiency of the estimates, the heteroscedasticity should show up in residual plots.

The preceding example shows that a refined approximation can improve efficiency of estimation, while the next describes a simple situation where bias can also be corrected; another example is discussed in the loglinear mean model case in Section 4.8.3. Consider ordinary homoscedastic quadratic regression with $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_{x,1}\mathbf{X} + \beta_{x,2}\mathbf{X}^2$. Use the same heteroscedastic Berkson model as before. Then the regression calibration approximation suggests a homoscedastic model with $\mathbf{X}$ replaced by $\mathbf{W}$, while in fact the observed data have mean $\beta_0 + \beta_{x,1}\mathbf{W} + \beta_{x,2}(\mathbf{W}^2 + \sigma^2_{\mathrm{rc}}\mathbf{W}^{2\gamma})$. If the Berkson error model is heteroscedastic, the regression calibration approximation will lead to a biased estimate of the regression parameters.

It is important to stress that these examples do not invalidate regression calibration as a method, because the heteroscedasticity in the Berkson error model has to be fairly severe before much effect will be noticed. However, there clearly is a need for refined approximations that take over when the regression calibration approximation breaks down.

4.7.1 The Expanded Approximation Defined

We will consider the QVF models (4.10) and (4.11). We will focus entirely on the case that $\mathbf{X}$ is a scalar. Although the general theory (Carroll and Stefanski, 1990) does allow multiple predictors, the algebraic details are unusually complex.

We will to discuss three different sets of approximations:

- A general formula.
- A modification of the general formula that is range preserving, for example, when a function must be positive.
- A simplification of the formula when functions are not too badly curved.

4.7.1.1 The General Development

The expanded approximation starts with both the mean and variance of $\mathbf{X}$ given $(\mathbf{Z},\mathbf{W})$:

$$\begin{aligned} E(\mathbf{X}|\mathbf{Z},\mathbf{W}) &= m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma); && (4.14)\\ \mathrm{var}(\mathbf{X}|\mathbf{Z},\mathbf{W}) &= \sigma^2_{\mathrm{rc}} V^2(\mathbf{Z},\mathbf{W},\gamma). && (4.15)\end{aligned}$$

We wish to construct approximations to the mean and variance function of the observed response given the observed covariates. Carroll and Stefanski (1990) based such approximations on *pretending* that σ^2_{rc} is “small”; if it equals zero, the resulting approximate model is the regression calibration model.

Here is how the approximation works. Let $m_{\mathbf{Y},x}$ and $m_{\mathbf{Y},xx}$ be the first and second derivatives of $m_{\mathbf{Y}}(z, x, \mathcal{B})$ with respect to x, and let $s_x(z, w, \mathcal{B}, \theta, \gamma)$ and $s_{xx}(\cdot)$ be the first and second derivatives of $s(z, x, \mathcal{B}, \theta) = g^2(z, x, \mathcal{B}, \theta)$ with respect to x and evaluated at $x = E(\mathbf{X}|\mathbf{Z} = z, \mathbf{W} = w)$. Defining $m_{\mathbf{X}}(\cdot) = m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \gamma)$ and $V(\cdot) = V(\mathbf{Z}, \mathbf{W}, \gamma)$, simple Taylor series expansions in Section B.3.3 with $\sigma^2_{\mathrm{rc}} \to 0$ yield the following approximate model, which we call the *expanded regression calibration model*:

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx m_{\mathbf{Y}}\left\{\mathbf{Z}, m_{\mathbf{X}}(\cdot), \mathcal{B}\right\} && (4.16)\\ &\quad +(1/2)\sigma^2_{\mathrm{rc}}V^2(\cdot)m_{\mathbf{Y},xx}\left\{\mathbf{Z}, m_{\mathbf{X}}(\cdot), \mathcal{B}\right\};\\ \mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx \sigma^2 g^2\left\{\mathbf{Z}, m_{\mathbf{X}}(\cdot), \mathcal{B}, \theta\right\} && (4.17)\\ &\quad +\sigma^2_{\mathrm{rc}}V^2(\cdot)\left\{m^2_{\mathbf{Y},x}(\cdot) + (1/2)\sigma^2 s_{xx}(\cdot)\right\}. \end{aligned}$$

There are important points to note about the approximate model (4.16)–(4.17):

- By setting $\sigma^2_{\mathrm{rc}} = 0$, it reduces to the regression calibration model, in which we need only estimate $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$.
- It is an approximate model that serves as a guide to final model construction in individual cases. We are not assuming that the measurement error is small, only pretending that it is in order to derive a plausible model for the observed data in terms of the regression parameters of interest. In some instances, terms can be dropped or combined with others to form even simpler useful models for the observed data.
- It is a mean and variance model for the observed data. Hence, the techniques of model fitting and model exploration discussed in Carroll and Ruppert (1988) can be applied to nonlinear measurement error model data.

4.7.1.2 Range-Preserving Modification

One potential problem with the expanded regression calibration model (4.16)–(4.17) is that it might not be range preserving. For example, because of the term $s_{xx}(\cdot)$, the variance function (4.17) need not necessarily be positive. If the original function $m_{\mathbf{Y}}(\cdot)$ is positive, the new approximate mean function (4.16) need not be positive because of the term $f_{xx}(\cdot)$. A range-preserving expanded regression calibration model for the observed data is

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \approx m_{\mathbf{Y}}\left[\mathbf{Z},\ m_{\mathbf{X}}(\cdot) + \frac{1}{2}\sigma^2_{\mathrm{rc}}\frac{V^2(\cdot)m_{\mathbf{Y},xx}(\cdot)}{m_{\mathbf{Y},x}(\cdot)}, \mathcal{B}\right]; \qquad (4.18)$$

$$\begin{aligned} \mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx \sigma^2_{\mathrm{rc}}m^2_{\mathbf{Y},x}\{\mathbf{Z}, m_{\mathbf{X}}(\cdot), \mathcal{B}\}V^2(\cdot) && (4.19)\\ &\quad +\sigma^2 g^2\left[\mathbf{Z}, m_{\mathbf{X}}(\cdot) + \frac{1}{2}\sigma^2_{\mathrm{rc}}\frac{V^2(\cdot)s_{xx}(\cdot)}{s_x(\cdot)}, \mathcal{B}, \theta\right]. \end{aligned}$$

4.7.1.3 Models Without Severe Curvature

When the models for the mean and variance are not severely curved, f_{xx} and s_{xx} are small relative to $m_{\mathbf{Y}}(\cdot)$ and $g^2(\cdot)$, respectively. In this case, setting $\kappa = \sigma^2_{\mathrm{rc}}/\sigma^2$, the mean and variance functions of the observed data greatly simplify to

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx m_{\mathbf{Y}}\left\{\mathbf{Z}, m_{\mathbf{X}}(\cdot), \mathcal{B}\right\}\\ \mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx \sigma^2\left[g^2\left\{\mathbf{Z}, m_{\mathbf{X}}(\cdot), \mathcal{B}, \theta\right\} + \kappa V^2(\cdot)m_{\mathbf{Y}}{}^2_x(\cdot)\right]. \end{aligned}$$

Having estimated the mean function $m_{\mathbf{X}}(\cdot)$, this is just a QVF model in the parameters $(\mathcal{B}, \theta_*)$, where θ_* consists of θ, κ and the other parameters in the function $V^2(\cdot)$. In principle, the QVF fitting methods of Appendix A can be used.

4.7.2 Implementation

The approximations (4.16) and(4.17) require specification of the mean and variance functions (4.14) and (4.15). In the Berkson model, the former is just $\mathbf{W}$ and a flexible model for the latter is $\sigma^2_{\mathrm{rc}}\mathbf{W}^{2\gamma}$, with $\gamma = 0$ indicating homoscedasticity. We will see later in a variety of examples that, for this Berkson class, the model parameters $(\mathcal{B}, \theta)$ are often estimable via QVF techniques using the approximate models, without the need for any validation data. The Berkson framework thus serves as an ideal environment for expanded regression calibration models.

Outside the Berkson class, we have already discussed in Sections 4.4 and 4.5 methods for estimating the conditional mean of $\mathbf{X}$. If possible, one should use available data to estimate the conditional variance function. For example, if there are k unbiased replicates in an additive measurement error model, then the natural counterpart to the best linear estimate of the mean function is the usual formula for the variance in a regression, namely $\mathrm{var}(\mathbf{X}|\mathbf{Z},\mathbf{W}) = \sigma^2_{\mathrm{rc}}$, where if σ^2_x is the variance of $\mathbf{X}$ and σ^2_u is the measurement error variance,

$$\sigma^2_{\mathrm{rc}} = \sigma^2_x - \left(\sigma^2_x, \Sigma_{xz}\right)\begin{bmatrix} \sigma^2_x + \sigma^2_u/k & \Sigma_{xz}\\ \Sigma^t_{xz} & \Sigma_{zz} \end{bmatrix}^{-1}\left(\sigma^2_x, \Sigma_{xz}\right)^t.$$

This can be estimated using the formulae of Section 4.4.2.

H	W	Y	H	W	Y	H	W	Y	H	W	Y
0	0	1.51	0	0	1.43	1	1	0.05	1	2	0.06
1	4	0.15	1	8	0.40	1	16	0.76	1	32	0.95
2	1	0.04	2	2	0.07	2	4	0.13	2	8	0.52
2	16	0.79	2	32	1.17	3	1	0.05	3	2	0.26
3	4	0.28	3	8	0.70	3	16	1.05	3	32	1.30
4	1	0.11	4	2	0.42	4	4	0.59	4	8	0.90
4	16	1.08	4	32	1.24	5	1	0.04	5	2	0.06
5	4	0.19	5	8	0.50	5	16	0.84	5	32	1.17
6	1	0.04	6	2	0.04	6	4	0.24	6	8	0.70
6	16	1.21	6	32	1.01	7	1	0.05	7	2	0.08
7	4	0.14	7	8	0.60	7	16	1.20	7	32	1.30
8	1	0.38	8	2	0.64	8	4	0.88	8	8	1.09
8	16	1.50	8	32	1.30						
0	0	1.01	0	0	1.34	1	1	0.05	1	2	0.07
1	4	0.09	1	8	0.26	1	16	0.55	1	32	1.21
2	1	0.04	2	2	0.06	2	4	0.19	2	8	1.16
2	16	0.96	2	32	1.13	3	1	0.04	3	2	0.17
3	4	0.33	3	8	0.50	3	16	1.11	3	32	1.20
4	1	0.12	4	2	0.30	4	4	0.41	4	8	1.06
4	16	1.29	4	32	1.17	5	1	0.04	5	2	0.07
5	4	0.19	5	8	0.36	5	16	0.88	5	32	1.16
6	1	0.04	6	2	0.05	6	4	0.22	6	8	0.61
6	16	1.15	6	32	1.39	7	1	0.04	7	2	0.18
7	4	0.27	7	8	0.88	7	16	0.97	7	32	1.26
8	1	0.29	8	2	0.98	8	4	1.12	8	8	1.10
8	16	1.13	8	32	1.31						

Table 4.2 *The bioassay data. Here* **Y** *is the response and* **W** *is the nominal dose time 32. The herbicides* **H** *are listed as 1–8, and* **H** = 0 *means a zero dose. The replicates* **R** *are separated by horizontal lines. The herbicide pairs are (1,5), (2,6), (3,7), and (4,8). Continued on next page.*

H	W	Y	H	W	Y	H	W	Y	H	W	Y
0	0	1.21	0	0	1.10	1	1	0.04	1	2	0.09
1	4	0.12	1	8	0.25	1	16	0.56	1	32	1.04
2	1	0.05	2	2	0.06	2	4	0.14	2	8	0.35
2	16	0.90	2	32	1.12	3	1	0.06	3	2	0.21
3	4	0.37	3	8	0.60	3	16	1.01	3	32	0.70
4	1	0.10	4	2	0.20	4	4	0.47	4	8	0.95
4	16	1.07	4	32	0.93	5	1	0.05	5	2	0.07
5	4	0.09	5	8	0.29	5	16	0.78	5	32	1.05
6	1	0.05	6	2	0.07	6	4	0.16	6	8	0.39
6	16	0.78	6	32	0.97	7	1	0.04	7	2	0.11
7	4	0.24	7	8	0.48	7	16	0.94	7	32	1.30
8	1	0.15	8	2	0.26	8	4	0.60	8	8	0.87
8	16	0.61	8	32	0.98						

Table 4.2 continued.

4.7.3 Bioassay Data

Rudemo, Ruppert, and Streibig (1989) described a bioassay problem following a heteroscedastic Berkson error model. In this experiment, four herbicides were applied either as technical grades or as commercial formulations; thus there are eight herbicides: four pairs of two herbicides each. The herbicides were applied at the six different nonzero doses 2^{j-5} for $j = 0, 1, \ldots, 5$. There were also two zero dose observations. The response $\mathbf{Y}$ was the dry weight of five plants grown in the same pot. There were three complete replicates of this experiment done at three different time periods, so that the replicates are a blocking factor. The data are listed in Table 4.2.

Let $\mathbf{Z}_1$ be a vector of size eight with a single nonzero element indicating which herbicide was applied, and let $\mathbf{Z}_2$ be a vector of size four indicating the herbicide pair. Let $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. For zero doses, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ may be defined arbitrarily as any nonzero value. In the absence of measurement error for doses, and if there were no random variation, the relationship between response and dose, $\mathbf{X}$, is expected to be

$$\mathbf{Y} \approx m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = \beta_0 + \frac{\beta_1 - \beta_0}{1 + \left\{ \dfrac{\mathbf{X}}{\beta_3^t \mathbf{Z}_1} \right\}^{\beta_4^t \mathbf{Z}_2}}. \tag{4.20}$$

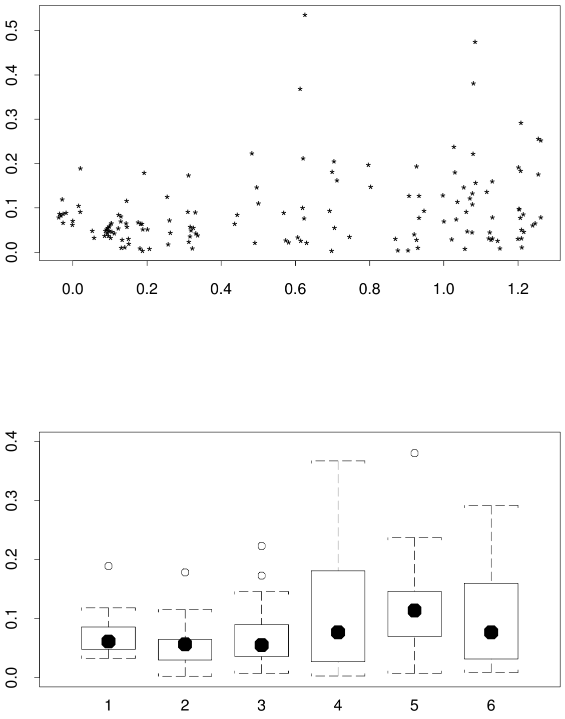


Figure 4.6 *Bioassay data. Absolute residual analysis for an ordinary nonlinear least squares fit. Note the increasing variability for larger predicted values.*

Model (4.20) is typically referred to as the *four-parameter logistic* model. Physically, the parameters β_0 and β_1 should be nonnegative, since they are the approximate dry weight at infinite and zero doses, respectively.

An initial ordinary nonlinear least squares fit to the data with a fixed block effect had a negative estimate of β_0. Figure 4.6 displays a plot of absolute residuals versus predicted means. Also displayed are box plots of the residuals formed by splitting the data into six equal-sized groups ordered on the basis of predicted values. Both figures show that the residuals are clearly heteroscedastic, with the response variance an increasing function of the predicted value.

This problem is exactly of the type amenable to analysis by the *transform-both-sides* (TBS) methodology of Carroll and Ruppert (1988); see also Ruppert, Carroll, and Cressie (1989). Specifically, model (4.20) is a theoretical model for the data in the absence of any randomness, which, when fit, shows a pattern of heteroscedasticity. The TBS methodology suggests controlling for the heteroscedasticity by transforming both sides of the equation:

$$h(\mathbf{Y}, \lambda) \approx h\left\{m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}), \lambda\right\}, \tag{4.21}$$

where the transformation family can be arbitrary but is taken here as the power transformation family:

$$\begin{aligned} h(v, \lambda) &= (v^{\lambda} - 1)/\lambda \text{ if } \lambda \neq 0; \\ &= \log(v) \text{ if } \lambda = 0. \end{aligned}$$

Of course, the actual dose applied $\mathbf{X}$ may be different from the nominal dose applied $\mathbf{W}$. It seems reasonable in this context to consider the Berkson error model with mean $\mathbf{W}$ and variance $\sigma^2_{\mathrm{rc}}\mathbf{W}^{2\gamma}$, the heteroscedasticity indicating the perfectly plausible assumption that the size of the error made depends on the nominal dose applied. With this specification, the regression calibration approximation replaces $\mathbf{X}$ by $\mathbf{W}$. Letting $\mathbf{Y}_{ij}$ be the j^{th} replicate at the i^{th} herbicide–dose combination, the TBS-regression calibration model incorporating randomness is

$$h(\mathbf{Y}_{ij}, \lambda) = h\left\{m_{\mathbf{Y}}(\mathbf{Z}_i, \mathbf{W}_i, \mathcal{B}), \lambda\right\} + \eta_j + \epsilon_{ij}, \tag{4.22}$$

where ϵ_{ij} is the homoscedastic random effect with variance σ^2, and η_j is the fixed block effect. The parameters were fit using maximum likelihood assuming that the errors are normally distributed, as described by Carroll and Ruppert (1988, Chapter 4). This involves maximizing the loglikelihood

$$-\frac{1}{2}\sum_{i,j}\left(\frac{\left[h(\mathbf{Y}_{ij}, \lambda) - h\left\{m_{\mathbf{Y}}(\mathbf{Z}_i, \mathbf{W}_i, \mathcal{B}), \lambda\right\} - \eta_j\right]^2}{\sigma^2}\right.$$
$$\left. + \log(\sigma^2) - 2(\lambda - 1)\log(\mathbf{Y}_{ij})\right).$$

The estimated transformation, $\widehat{\lambda} = 0.117$, is very near the log transformation. The residual plots are given in Figure 4.7, where we still see some unexplained structure to the variability, since the extremes of the predicted means have smaller variability than the centers.

To account for the unexplained variability, we now the consider higher-order approximate models (4.16) and (4.17). Denoting the left-hand

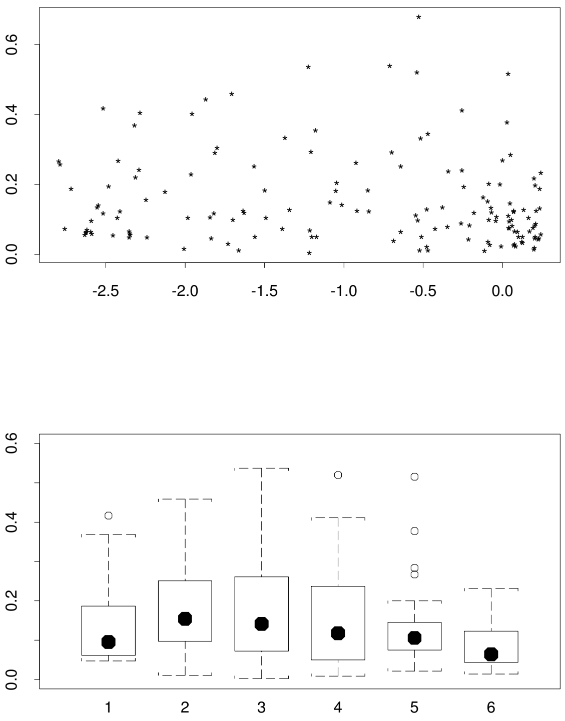


Figure 4.7 *Bioassay data. Absolute residual analysis for an ordinary transform-both-sides fit. Note the unexplained structure of the variability.*

side of (4.21) by $\mathbf{Y}_*$ and the right-hand side by $m_{\mathbf{Y}*}(\cdot)$, and noting that the four-parameter logistic model is one in which $m_{\mathbf{Y},xx}/m_{\mathbf{Y}}$ is typically small, the approximate model (4.17) says that $\mathbf{Y}_*$ has mean $h\left\{m_{\mathbf{Y}}(\mathbf{Z},\mathbf{W},\mathcal{B})\right\}$ and variance $\sigma^2+\sigma^2_{\mathrm{rc}}\mathbf{W}^{2\gamma}\left\{(m_{\mathbf{Y}})^{\lambda-1}(\mathbf{Z},\mathbf{W},\mathcal{B})m_{\mathbf{Y},x}(\mathbf{Z},\mathbf{W},\mathcal{B})\right\}^2$. If we define $\kappa=\sigma^2_{\mathrm{rc}}/\sigma^2$, in contrast to (4.22) an approximate model for the data is

$$\begin{aligned} h(\mathbf{Y}_{ij},\lambda) &= h\left\{m_{\mathbf{Y}}(\cdot),\lambda\right\}+\eta_j \qquad (4.23)\\ &\quad +\epsilon_{ij}\left[1+\kappa\mathbf{W}_i^{2\gamma}\left\{(m_{\mathbf{Y}})^{\lambda-1}(\cdot)m_{\mathbf{Y},x}(\cdot)\right\}^2\right]^{1/2}, \end{aligned}$$

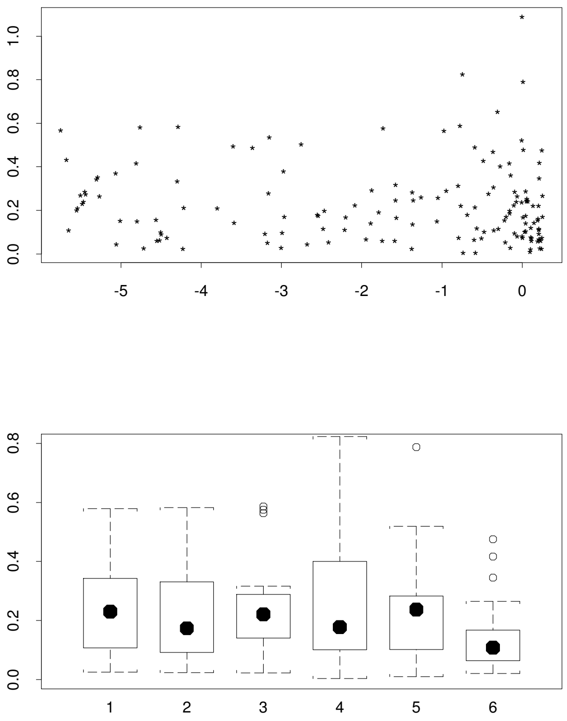


Figure 4.8 *Bioassay data. Absolute residual analysis for a second-order approximate transform-both-sides fit.*

whereas before ϵ_{ij} has variance σ^2. This is a heteroscedastic TBS model, all of whose parameters are identifiable and hence estimable from the observed data. The identifiability of parameters in the Berkson model is a general phenomenon; see Section 4.9. The likelihood of (4.23) is the same as before but with σ^2 replaced by

$$\sigma^2\left[1+\kappa\mathbf{W}_i^{2\gamma}\left\{(m_{\mathbf{Y}})^{\lambda-1}(\cdot)m_{\mathbf{Y},x}(\cdot)\right\}^2\right].$$

This model was fit to the data, and $\widehat{\lambda}\approx -1/3$ with an approximate standard error of 0.12. The corresponding residual plots are given in Figure 4.8. Here we see no real hint of unexplained variability. As a further check, we can contrast the models (4.23) and (4.22) by means of a likelihood ratio test, the two extra parameters being (γ, κ). The likelihood ratio test for the hypothesis that these two parameters equal zero had a chi-squared value of over 30 based on two degrees of freedom, indicating a large improvement in the fit due to allowing for possible heteroscedasticity in the Berkson error model.

4.8 Examples of the Approximations

In this section, we investigate the appropriateness of the regression calibration algorithm in a variety of settings.

4.8.1 Linear Regression

Consider linear regression when the variance of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ is constant, so that the mean and variance of $\mathbf{Y}$ when given $(\mathbf{Z}, \mathbf{X})$ are $\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}$ and σ^2, respectively. As an approximation, the regression calibration model says that the observed data also have constant variance but have regression function given by $E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x^t m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \gamma) + \beta_z^t\mathbf{Z}$. Because we assume nondifferential measurement error (Section 2.5), the regression calibration model accurately reproduces the regression function, but the observed data have a different variance, namely

$$\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \sigma^2 + \beta_x^t \text{var}(\mathbf{X}|\mathbf{Z}, \mathbf{W})\beta_x.$$

Note the difference here: The regression calibration model is a working model for the observed data, which may differ somewhat from the actual or true model for the observed data. In this case, the regression calibration approximation gives the correct mean function, and the variance function is also correct and constant if $\mathbf{X}$ has a constant covariance matrix given $(\mathbf{Z}, \mathbf{W})$.

If, however, $\mathbf{X}$ has nonconstant conditional variance, the regression calibration approximation would suggest the homoscedastic linear model when the variances are heteroscedastic. In this case, while the least squares estimates would be consistent, the usual standard errors are incorrect. There are three options: (i) use least squares and bootstrap by resampling vectors (Section A.9.2); (ii) use least-squares and the sandwich method for constructing standard errors (Section A.6); and (iii) expand the model using the methods of Section 4.7.

4.8.2 Logistic Regression

Regression calibration is also well established in logistic regression, at least as long as the effects of the variable $\mathbf{X}$ measured with error are not "too large" (Rosner, Willett, and Spiegelman, 1989; Rosner, Spiegelman, and Willett, 1990; Whittemore, 1989). Let the binary response $\mathbf{Y}$ follow the logistic model $\Pr(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{X}) = H(\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z})$, where $H(v) = \{1 + \exp(-v)\}^{-1}$ is the logistic distribution function. The key problem is computing the probability of a response $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$. For example, suppose that $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is normally distributed with mean $m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \gamma)$ and (co)variance function $V(\mathbf{Z}, \mathbf{W}, \gamma)$. Let p be the number of components of $\mathbf{X}$. As described in more detail in Chapter 8, the probability that $\mathbf{Y} = 1$ for values of $(\mathbf{Z}, \mathbf{W})$ is

$$\frac{\int H(\cdot)\exp\left[-(1/2)\left\{x - m_{\mathbf{X}}(\cdot)\right\}^t V^{-1}(\cdot)\left\{x - m_{\mathbf{X}}(\cdot)\right\}\right]dx}{(2\pi)^{p/2}|V(\cdot)|^{1/2}}, \tag{4.24}$$

where $H(\cdot) = H(\beta_0 + \beta_x^t x + \beta_z^t\mathbf{Z})$. Formula (4.24) does not have a closed-form solution; Crouch and Spiegelman (1990) developed a fast algorithm that they have implemented in FORTRAN: unfortunately, as far as we know, this algorithm is not in widespread use. Monahan and Stefanski (1991) described a different method easily applicable to all standard computer packages. However, a simple technique often works just as well, namely, to approximate the logistic by the probit. It is well known that $H(v) \approx \Phi(v/1.7)$, where $\Phi(\cdot)$ is the standard normal distribution function (Johnson and Kotz, 1970; Liang and Liu, 1991; Monahan and Stefanski, 1991).

In Figure 4.9 we plot the density and distribution functions of the logistic and normal distributions, and the reader will note that the logistic and normal are very similar. With some standard algebra (Carroll, Bailey, Spiegelman, et al., 1984), one can approximate (4.24) by

$$\Pr(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{W}) \approx H\left[\frac{\beta_0 + \beta_x^t m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \gamma) + \beta_z^t\mathbf{Z}}{\left\{1 + \beta_x^t V(\mathbf{Z}, \mathbf{W}, \gamma)\beta_x/1.7^2\right\}^{1/2}}\right]. \tag{4.25}$$

In most cases, the denominator in (4.25) is very nearly 1, and regression calibration is a good approximation; the exception is for "large" $\beta_x^t V(\cdot)\beta_x$. In general, the denominator in (4.25) means that regression calibration will lead to estimates of the main risk parameters that are slightly attenuated.

The approximation (4.25) is often remarkably good, even when the true predictor $\mathbf{X}$ is rather far from normally distributed. To test this, we dropped $\mathbf{Z}$ and computed the approximations and exact forms of $\text{pr}(\mathbf{Y} = 1|\mathbf{W})$ under the following scenario. For the distribution of $\mathbf{X}$, we chose either a standard normal distribution or the chi-squared distribution with one degree of freedom. The logistic intercept β_0 and slope β_x were chosen so that there was a 10% positive response rate $(\mathbf{Y} = 1)$ on average, and so that $\exp\{\beta_x(q_{90} - q_{10})\} = 3$, where q_a is the a^{th} per-

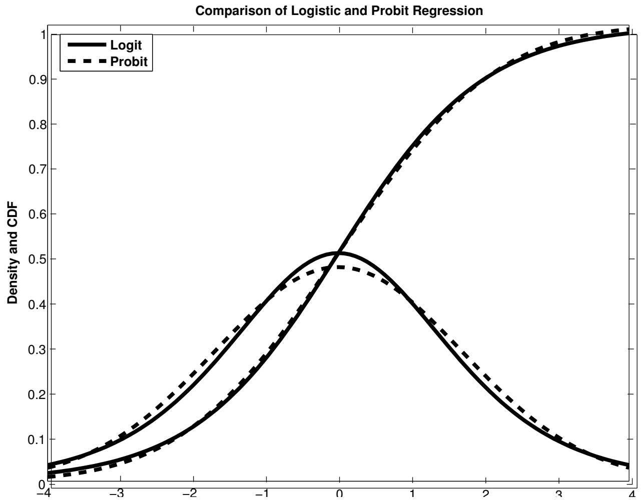


Figure 4.9 *The standard logistic distribution and density functions compared to the normal distribution and density functions with standard deviation 1.70. The point of the graph is to show how close the two are.*

centile of the distribution of $\mathbf{X}$. In the terminology of epidemiology, this means that the "relative risk" is 3.0 in moving from the 10th to the 90th percentile of the distribution of $\mathbf{X}$, a representative situation.

In Figure 4.10 we plot values of $\mathrm{pr}(\mathbf{Y}=1|\mathbf{W})$ against $\mathbf{W}$ in the solid line, for the range from the 5th to the 95th percentile of the distribution of $\mathbf{W}$. The regression calibration approximation is the dotted line. The measurement error is additive on the first row and multiplicative on the second row. The top left plot has $\mathbf{W}=\mathbf{X}+\mathbf{U}$ where $(\mathbf{X},\mathbf{U})$ follow a bivariate standard normal distribution, while the top right plot differs in that both follow a chi-squared distribution with one degree of freedom. The bottom row has $\mathbf{W}=\mathbf{X}\mathbf{U}$, where $\mathbf{U}$ follows a chi-squared distribution with one degree of freedom; on the left, $\mathbf{X}$ is standard normal, while on the right, $\mathbf{X}$ is chi-squared. Note that the solid and dashed lines very nearly overlap. In all of these cases, the measurement error is *very* large, so in some sense we are displaying a worst case scenario. For these four very different situations, the regression calibration approximation works very well indeed.

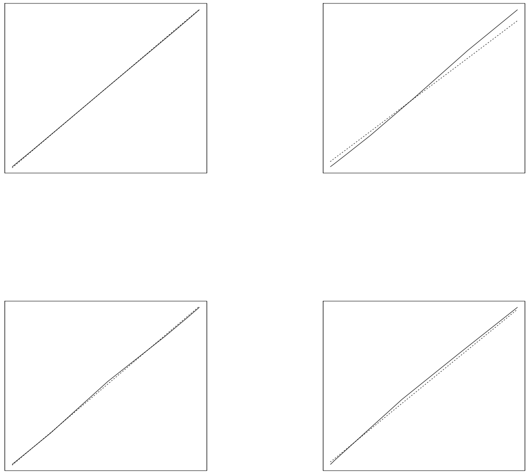


Figure 4.10 *Values of* $\mathrm{pr}(\mathbf{Y}=1|\mathbf{W})$ *are plotted against* $\mathbf{W}$ *in the solid line, while the regression calibration approximation is the dotted line. The measurement error is additive on the first row and multiplicative on the second row. The fact that the lines are nearly indistinguishable is the whole point. See text for more details.*

4.8.3 Loglinear Mean Models

As might occur for gamma or lognormal data, suppose $E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = \exp(\beta_0+\beta_x^t\mathbf{X}+\beta_z^t\mathbf{Z})$ and $\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = \sigma^2\left\{E(\mathbf{Y}|\mathbf{Z},\mathbf{X})\right\}^2$. Suppose that the calibration of $\mathbf{X}$ on $(\mathbf{Z},\mathbf{W})$ has mean $m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma)$, and denote the moment generating function of the calibration distribution by

$$E\left\{\exp(a^t\mathbf{X})|\mathbf{Z},\mathbf{W}\right\} = \exp\left\{a^t m_{\mathbf{X}}(\mathbf{Z},\mathbf{W},\gamma)+v(a,\mathbf{Z},\mathbf{W},\gamma)\right\},$$

where $v(\cdot)$ is a general function which differs from distribution to distribution. If $(\cdot)=(\mathbf{Z},\mathbf{W},\gamma)$, the observed data then follow the model

$$\begin{aligned}
E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &= \exp\left\{\beta_0+\beta_x^t m_{\mathbf{X}}(\cdot)+\beta_z^t\mathbf{Z}+v(\beta_x,\cdot)\right\};\\
\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &= \exp\left\{2\beta_0+2\beta_x^t m_{\mathbf{X}}(\cdot)+2\beta_z^t\mathbf{Z}+v(2\beta_x,\cdot)\right\}\\
&\quad\times\left[\sigma^2+1-\exp\left\{2v(\beta_x,\cdot)-v(2\beta_x,\cdot)\right\}\right].
\end{aligned}$$

If the calibration distribution for $\mathbf{X}$ is normal with constant covariance matrix Σ_{xx}, then $v(a,\cdot) = (1/2)a^t\Sigma_{xx}a$. Remarkably, for $\beta_{0*} = \beta_0 + (1/2)\beta_x^t\Sigma_{xx|z,w}\beta_x$, the observed data *also* follow the loglinear mean model with intercept β_{0*} and a new variance parameter σ_*^2. Thus, the regression calibration approximation is exactly correct for the slope parameters (β_x, β_z)! The conclusion holds more generally, requiring only that $\mathbf{X}-\ m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \gamma)$ have distribution independent of $(\mathbf{Z}, \mathbf{W})$.

4.9 Theoretical Examples

4.9.1 Homoscedastic Regression

The simple homoscedastic linear regression model is $m_{\mathbf{Y}}(z, x, \mathcal{B}) = \beta_0 + \beta_x x + \beta_z z$ with $g^2(\cdot) = V^2(\cdot) = 1$. If the variance function (4.15) is homoscedastic, then the approximate model (4.16)–(4.17) is exact in this case with $E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x m_{\mathbf{X}}(\cdot) + \beta_z\mathbf{Z}$ and $\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \sigma^2 + \sigma_{\text{rc}}^2\beta_x^2$, that is, a homoscedastic regression model. One sees clearly that the effect of measurement error is to inflate the error about the observed line.

In simple linear regression satisfying a Berkson error model with possibly heteroscedastic calibration variances $\sigma_{\text{rc}}^2\mathbf{W}^{2\gamma}$, the approximations are again exact: $E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x\mathbf{W} + \beta_z\mathbf{Z}$ and $\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \sigma^2\left\{1 + \beta_x^2(\sigma_{\text{rc}}^2/\sigma^2)\mathbf{W}^{2\gamma}\right\}$. The reader will recognize this as a QVF model, where the parameter $\theta = (\gamma, \kappa = \sigma_{\text{rc}}^2/\sigma^2)$. As long as $\gamma \neq 0$, all the parameters are estimable by standard QVF techniques, without recourse to validation or replication data.

This problem is an example of a remarkable fact, namely that in Berkson error problems, the approximations (4.16) and (4.17) often lead to an identifiable model, so that the parameters can all be estimated without recourse to validation data. Of course, if one does indeed have validation data, then they can be used to improve upon the approximate QVF estimators.

4.9.2 Quadratic Regression with Homoscedastic Regression Calibration

Ordinary quadratic regression has mean function $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_{x,1}\mathbf{X} + \beta_{x,2}\mathbf{X}^2$. With homoscedastic regression calibration, the observed data have mean function

$$\begin{aligned} E(\mathbf{Y}|\mathbf{W}) &= (\beta_0 + \beta_{x,2}\sigma^2) + \beta_{x,1}m_{\mathbf{X}}(\mathbf{W}) + \beta_{x,2}m_{\mathbf{X}}^2(\mathbf{W}) \\ &= \beta_0^* + \beta_{x,1}m_{\mathbf{X}}(\mathbf{W}) + \beta_{x,2}m_{\mathbf{X}}^2(\mathbf{W}). \end{aligned}$$

As illustrated in Section 4.8.3, the regression calibration model accurately reflects the observed data in terms of the slope parameters, but

it is off by a constant, since its intercept β_0^* differs from β_0. Here, however, the approximate expanded mean model (4.16) is exact, and β_0 can be estimated as long as one has available an estimate of the calibration variance σ^2; see the previous section.

If the error of $\mathbf{X}$ about its conditional mean is homoscedastic and symmetrically distributed, for example, normally distributed, then the expanded regression calibration model accurately reflects the form of the variance function for the observed data. Details are given in Appendix B.3.2. If the error is asymmetric, then the expanded model (4.17) misses a term involving the third error moment.

4.9.3 Loglinear Mean Model

The loglinear mean model of Section 4.8.3 has $E(\mathbf{Y}|\mathbf{X}) = \exp(\beta_0 + \beta_x\mathbf{X})$, and variance proportional to the square of the mean with constant of proportionality σ^2. If calibration is homoscedastic and normally distributed, the actual mean function for the observed data is $E(\mathbf{Y}|\mathbf{W}) = \exp\left\{\beta_0 + (1/2)\beta_x^2\sigma^2 + \beta_x m_{\mathbf{X}}(\mathbf{W})\right\}$. The mean model of regression calibration is $\exp\left\{\beta_0 + \beta_x m_{\mathbf{X}}(\mathbf{W})\right\}$. Regression calibration yields a consistent estimate of the slope β_x but not of the intercept.

In this problem, the range-preserving expanded regression calibration model (4.18) correctly captures the mean of the observed data. Interestingly, it also captures the essential feature of the variance function, since both the actual and approximate variance functions (4.19) are a constant times $\exp\left\{2\beta_0 + 2\beta_x m_{\mathbf{X}}(\mathbf{W})\right\}$.

Bibliographic Notes and Software

This *regression calibration* algorithm was suggested as a general approach by Carroll and Stefanski (1990) and Gleser (1990). Prentice (1982) pioneered the idea for the proportional hazard model, where it is still the default option, and a modification of it has been suggested for this topic by Clayton (1991); see Chapter 14. Armstrong (1985) suggests regression calibration for generalized linear models, and Fuller (1987, pp. 261–262) briefly mentioned the idea. Rosner, Willett and Spiegelman (1989) and Rosner, Spiegelman, and Willett (1990) developed the idea for logistic regression into a methodology particularly popular in epidemiology.

There is a long history of approximately consistent estimates in nonlinear problems, of which regression calibration and the SIMEX method (Chapter 5) and are the most recent such methods. Readers should also consult Stefanski and Carroll (1985), Stefanski (1985), Amemiya and Fuller (1988), Amemiya (1985, 1990a, 1990b), and Whittemore and Keller (1988) for other approaches.

Stata (http://www.stata.com/merror) has code for regression calibration and SIMEX (see next chapter) for generalized linear models. The programs allow for known measurement error variance, measurement error variance estimated by replications, bootstrapping, etc. A detailed example using the Framingham data along with the data are at the book Web site:
http://www.stat.tamu.edu/∼carroll/eiv.SecondEdition/statacode.php.

CHAPTER 5

SIMULATION EXTRAPOLATION

5.1 Overview

In this chapter we describe a measurement error bias-correction method that shares the simplicity, generality, and approximate-inference characteristics of regression calibration. As the previous chapter indicated, regression calibration is ideally suited for problems in which the calibration function $E(\mathbf{X} \mid \mathbf{W})$ can be estimated reasonably well and to problems such as generalized linear models. Simulation extrapolation (SIMEX) is ideally suited to problems with additive measurement error, and more generally to any problem in which the measurement error generating process can be imitated on a computer via Monte Carlo methods.

SIMEX is a simulation-based method of estimating and reducing bias due to measurement error. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error–induced bias versus the variance of the added measurement error, and extrapolating this trend back to the case of no measurement error. The technique was proposed by Cook and Stefanski (1994) and further developed by Stefanski and Cook (1995), Carroll, Küchenhoff, Lombard, and Stefanski (1996), Devanarayan (1996), Carroll and Stefanski (1997), and Devanarayan and Stefanski (2002). SIMEX is closely related to the Monte Carlo corrected score (MCCS) method described in Chapter 7, and the interested reader may want to read the present chapter and the MCCS material in Chapter 7 in combination.

The fact that measurement error in a predictor variable induces bias in parameter estimates is counterintuitive to many people. An integral component of SIMEX is a self-contained simulation study resulting in graphical displays that illustrate the effect of measurement error on parameter estimates and the need for bias correction. The graphical displays are useful when it is necessary to motivate or explain a measurement error model analysis.

SIMEX is very general in the sense that the bias due to measurement error in almost any estimator of almost any parameter is readily estimated and corrected, at least approximately. In the absence of measurement error, it is often the case that competing estimators are available that are consistent for the same parameter, and only differ asymptotically with respect to sampling variability. However, these same estimators can be differentially affected by measurement error. In Section 5.3.1 we present such an illustrative example and show how SIMEX clearly reveals the differences in biases.

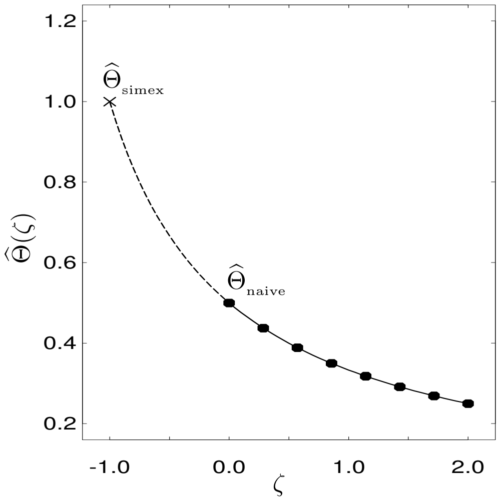
The key features of the SIMEX algorithm are described in the context of linear regression in the following section. Detailed descriptions of the method for different measurement error models are then given, along with an illustrative application to regression through the origin using weighted least squares estimation. Next, the SIMEX method is illustrated for different measurement error models using data from the Framingham Heart Study. The first four sections are sufficient for the reader wanting a working knowledge of the SIMEX method. Following the Framingham example, theoretical aspects of SIMEX estimation are also described.

5.2 Simulation Extrapolation Heuristics

5.2.1 SIMEX in Simple Linear Regression

We describe the basic idea of SIMEX in the context of simple linear regression with additive measurement error. In Section 5.3 we show how to extend SIMEX to nonadditive models and provide additional examples. Suppose that $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, with additive measurement error $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $(\mathbf{Y}, \mathbf{X})$ and has mean zero and variance σ_u^2. The ordinary least squares estimate of β_x, denoted $\widehat{\beta}_{x,\text{naive}}$, consistently estimates $\beta_x \sigma_x^2 / (\sigma_x^2 + \sigma_u^2)$ (Chapter 3) and thus is biased for β_x when $\sigma_u^2 > 0$. For this simple model, the effect of measurement error on the least squares estimator is easily determined mathematically, and simple method-of-moments bias corrections are known (Chapter 3; Fuller, 1987). Thus, in practice, simple linear regression typically would not be a candidate for SIMEX analysis. However, we use it here to show that SIMEX provides essentially the same bias corrections as the method-of-moments.

The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation. In a study of the effect of radiation exposure on tumor development in rats, one is naturally led to an experiment in which radiation dose is varied. Similarly, in a study of the biasing effects of measurement error on an estimator, one is naturally led to an experiment in which the level of measurement error is varied. So if we regard measurement error as a factor whose influence on an estimator is to be determined, we consider simulation experiments in which the level of the measurement error, as measured by its variance, is intentionally varied.



PSfrag replacements

Figure 5.1 *A generic SIMEX plot showing the effect on a statistic of adding measurement error with variance $\zeta\sigma_u^2$ to the data when estimating a parameter Θ. The abscissa (x-axis) is ζ, and the ordinate (y-axis) is the estimated coefficient. The SIMEX estimate is an extrapolation to $\zeta = -1$. The naive estimate occurs at $\zeta = 0$.*

To this end, suppose that in addition to the original data used to calculate $\widehat{\beta}_{x,\text{naive}}$, there are $M - 1$ additional data sets available, each with successively larger measurement error variances, say $(1 + \zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < \cdots < \zeta_M$ are known. The least squares estimate of slope from the m^{th} data set, $\widehat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2 / \{\sigma_x^2 + (1 + \zeta_m)\sigma_u^2\}$ (Chapter 3; Fuller, 1987).

We can formulate this setup as a nonlinear regression model, with data $\{(\zeta_m, \widehat{\beta}_{x,m}),\ m = 1, \ldots, M\}$, dependent variable $\widehat{\beta}_{x,m}$, and independent variable ζ_m. Asymptotically, the mean function of this regression has the form

$$E(\widehat{\beta}_{x,m} \mid \zeta) = \mathcal{G}(\zeta) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1 + \zeta)\sigma_u^2}, \qquad \zeta \geq 0.$$

Note that $\mathcal{G}(-1) = \beta_x$. That is, the parameter of interest is obtained from $\mathcal{G}(\zeta)$ by extrapolation to $\zeta = -1$. The significance of $\zeta = -1$ will become apparent later in this chapter and again in Chapter 7. Heuristically, it suffices to see that the measurement error variances in our data sets are equal to $(1+\zeta_m)\sigma_u^2$. Ideally, we would like error-free data sets, and in terms of ζ_m this corresponds to having $(1+\zeta_m)\sigma_u^2 = 0$, and thus $\zeta_m = -1$.

SIMEX imitates the procedure described above, as illustrated schematically in Figure 5.1.

- In the *simulation step*, additional independent measurement errors with variance $\zeta_m\sigma_u^2$ are generated and added to the original **W** data, thereby creating data sets with successively larger measurement error variances. For the m^{th} data set, the total measurement error variance is $\sigma_u^2 + \zeta_m\sigma_u^2 = (1+\zeta_m)\sigma_u^2$.
- Next, estimates are obtained from each of the generated contaminated data sets.
- The simulation and estimation steps are repeated a large number of times, and the average value of the estimate for each level of contamination is calculated. These averages are plotted against the ζ values and a regression technique, for example, nonlinear least squares, is used to fit an extrapolant function to the averaged, error-contaminated estimates. See Section 5.3.2 for a discussion of extrapolation.
- Extrapolation to the ideal case of no measurement error ($\zeta = -1$) yields the SIMEX estimate.

5.3 The SIMEX Algorithm

5.3.1 Simulation and Extrapolation Steps

We now explain the SIMEX algorithm in detail for four combinations of error models and measured data. In the first, a single measurement for each case is available and the measurement errors are homoscedastic with a known, or independently estimated, variance. In the second, a single measurement for each case is available and the measurement errors are heteroscedastic with known variances. In the third case, replicate measurements are assumed but no additional assumptions are made about the error variances, that is, it is not assumed that they are known and they could be homoscedastic or heteroscedastic. In the fourth case, we show how the method generalizes to certain multiplicative error models and give some illustrative examples.

We do not discuss the extrapolation step in detail in any of the four cases that follow. Once a functional form is selected, fitting the extrapolant function and extrapolating are routine applications of linear or nonlinear regression, using ζ as the independent variable and $\widehat{\Theta}(\zeta)$ given below in equation (5.3) as the dependent variable. However, the choice of functional form is important, and we discuss that in Section 5.3.2.

5.3.1.1 Homoscedastic Errors with Known Error Variance

While SIMEX is a general methodology, it is easiest to understand when there is only a single, scalar predictor **X** subject to additive error, though there could be multiple covariates **Z** measured without error, and $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{U}_i$ is a normal random variable with variance σ_u^2, and is independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{Y}_i$. Typically, minor violations of the assumption of normality of the measurement errors is not critical in practice. We assume that the measurement error variance, σ_u^2, is known or sufficiently well estimated to regard as known.

SIMEX, like regression calibration, is applicable to general estimation methods, for example, least-squares, maximum likelihood, quasilikelihood, etc. In this section, we will not distinguish among the methods, but instead will refer to "the estimator" to mean the chosen estimation method computed as if there were no measurement error. We let Θ denote the parameter of interest.

The first part of the algorithm is the simulation step. As described above, this involves using simulation to create additional data sets of increasingly larger measurement error $(1+\zeta)\sigma_u^2$. For any $\zeta \geq 0$, define

$$\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \sqrt{\zeta}\,\mathbf{U}_{b,i}, \qquad i = 1,\ldots,n, \qquad b = 1,\ldots,B, \tag{5.1}$$

where the computer-generated *pseudo errors*, $\{\mathbf{U}_{b,i}\}_{i=1}^n$, are mutually independent, independent of all the observed data, and identically distributed, normal random variables with mean 0 and variance σ_u^2. We call $\mathbf{W}_{b,i}(\zeta)$ a *remeasurement* of $\mathbf{W}_i$, because it is a measurement of $\mathbf{W}_i$ in the same statistical sense that $\mathbf{W}_i$ is a measurement of $\mathbf{X}_i$.

Note that $\text{var}(\mathbf{W}_i|\mathbf{X}_i) = \sigma_u^2$, whereas

$$\text{var}\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = (1+\zeta)\sigma_u^2 = (1+\zeta)\text{var}(\mathbf{W}_i|\mathbf{X}_i). \tag{5.2}$$

The error variance in the remeasured data has been inflated by a multiplicative factor, $(1+\zeta)$ in this case, that equals zero when $\zeta = -1$. Because $E\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = \mathbf{X}_i$, (5.2) implies that the mean squared error of $\mathbf{W}_{b,i}$ as a measurement of $\mathbf{X}_i$ defined as $\text{MSE}\{\mathbf{W}_{b,i}(\zeta)\} = E[\{\mathbf{W}_{b,i}(\zeta) - \mathbf{X}_i\}^2|\mathbf{X}_i]$ converges to zero as $\zeta \to -1$. This is the key property of the simulated pseudo data, or remeasured data.

Having generated the remeasured predictors, we compute the corresponding naive estimates. Define $\widehat{\Theta}_b(\zeta)$ to be the estimator when the $\{\mathbf{W}_{b,i}(\zeta)\}_1^n$ are used, and define the average of these estimators as

$$\widehat{\Theta}(\zeta) = B^{-1}\textstyle\sum_{b=1}^{B}\widehat{\Theta}_b(\zeta). \tag{5.3}$$

By design, $\widehat{\Theta}(\zeta)$ is the sample mean of $\{\widehat{\Theta}_b(\zeta)\}_1^B$, and hence is the average of the estimates obtained from a large number of experiments with the same amount of measurement error. The reason for averaging over many simulations is that we are interested in estimating the extra *bias* due to added measurement error, not in inducing more variability, and averaging reduces the Monte Carlo simulation variation. It is the points $\{\zeta_m, \widehat{\Theta}(\zeta_m)\}_1^M$ that are plotted as filled circles in Figure 5.1. This is the simulation component of SIMEX.

Note that the components of $\widehat{\Theta}(\zeta)$ are all functions of the same scalar ζ, and there is a separate extrapolation step for each component of $\widehat{\Theta}(\zeta)$. The extrapolation step entails modeling each of the components of $\widehat{\Theta}(\zeta)$ as functions of ζ for $\zeta \geq 0$ and extrapolating the fitted models back to $\zeta = -1$. The vector of extrapolated values yields the simulation extrapolation estimator denoted $\widehat{\Theta}_{\text{simex}}$. In Figure 5.1, the extrapolation is indicated by the dashed line and the SIMEX estimate is plotted as a cross. Heuristically, the significance of $\zeta = -1$ follows from the fact that $\widehat{\Theta}(\zeta)$ is calculated from measurements having variance $\text{var}\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = (1+\zeta)\sigma_u^2$, and we want to extrapolate to the case in which the error variance in the measurements is zero, that is, $(1+\zeta)\sigma_u^2 = 0$, or equivalently $\zeta = -1$. Note that although we cannot add measurement error with negative variance, $\zeta\sigma_u^2 = -\sigma_u^2$, we can add error with positive variance, determine the form of the bias as a function of ζ, and extrapolate to the hypothetical case of adding negative variance ($\zeta = -1$).

5.3.1.2 *Heteroscedastic Errors with Known Error Variances*

Suppose now that $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{U}_i$ is a normal random variable with variance $\sigma_{u,i}^2$, is independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{Y}_i$, and $\sigma_{u,i}^2$ is known. This is not a common error model, but it provides a useful stepping stone to other, more common heteroscedastic error models. In addition, it is appropriate when $\mathbf{W}_i$ is the mean of $k_i \geq 1$ replicate measurements, each having known variance σ_u^2, in which case $\sigma_{u,i}^2 = \sigma_u^2/k_i$.

In this case, the only change in the algorithm is that the remeasurement procedure in (5.1) is replaced by

$$\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \sqrt{\zeta}\,\mathbf{U}_{b,i}, \qquad i = 1, \ldots, n, \qquad b = 1, \ldots, B, \tag{5.4}$$

where the pseudo errors, $\{\mathbf{U}_{b,i}\}_{i=1}^n$, are again mutually independent, independent of all the observed data, and identically distributed, normal random variables with mean 0 and variance $\sigma_{u,i}^2$. Note that

$$\text{var}\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = (1+\zeta)\sigma_{u,i}^2 = (1+\zeta)\text{var}(\mathbf{W}_i|\mathbf{X}_i), \tag{5.5}$$

and $E\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = \mathbf{X}_i$. So just as in the preceding case, we see that the two variances, $\text{var}(\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i)$ and $\text{var}(\mathbf{W}_i|\mathbf{X}_i)$, differ by a multiplicative factor that vanishes when $\zeta = -1$, and consequently that $\text{MSE}\{\mathbf{W}_{b,i}(\zeta)\} = E[\{\mathbf{W}_{b,i}(\zeta) - \mathbf{X}_i\}^2|\mathbf{X}_i] \to 0$ as $\zeta \to -1$, the key property of the remeasured data.

The averaged naive estimates, $\widehat{\Theta}(\zeta)$, are calculated in exactly the same way as for the case of the homoscedastic error model. The SIMEX estimator, $\widehat{\Theta}_{\text{simex}}$, is again obtained by modeling and extrapolation to $\zeta = -1$, as this is the value of ζ for which $(1+\zeta)\sigma_{u,i}^2 = 0$ for all i.

5.3.1.3 *Heteroscedastic Errors with Unknown Variances and Replicate Measurements*

We now consider an error model that allows for arbitrary unknown heteroscedastic error variances. SIMEX estimation for this model was developed and studied by Devanarayan (1996) and Devanarayan and Stefanski (2002). For this model $k_i \geq 2$ replicate measurements are necessary for each subject in order to identify the error variances $\sigma_{u,i}^2$. The assumed error model is $\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j}$, where $\mathbf{U}_{i,j}$, $j = 1, \ldots, k_i$, are $\text{Normal}(0, \sigma_{u,i}^2)$, independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{Y}_i$ with all $\sigma_{u,i}^2$ *unknown*. With replicate measurements, the best measurement of $\mathbf{X}_i$ is the mean $\overline{\mathbf{W}}_{i,\cdot}$, and we define the so-called naive estimation procedure as doing the usual, nonmeasurement error analysis, of the data $(\mathbf{Y}_i, \mathbf{Z}_i, \overline{\mathbf{W}}_{i,\cdot})_1^n$.

Because the variances, $\sigma_{u,i}^2$, are unknown, we cannot generate remeasured data as in (5.4). However, recall that the key property of the remeasured data is that the variance of the best measurement of $\mathbf{X}_i$ is inflated by the factor $1+\zeta$. With replicate measurements, we can obtain such variance-inflated measurements by taking suboptimal linear combinations of the replicate measurements. This is done using random linear contrasts.

Suppose that $\mathbf{c}_{b,i} = (c_{b,i,1}, \ldots, c_{b,i,k_i})^t$ is a normalized contrast vector, $\sum_j c_{b,i,j} = 0$ and $\sum_j c_{b,i,j}^2 = 1$. Define

$$\mathbf{W}_{b,i}(\zeta) = \overline{\mathbf{W}}_{i,\cdot} + (\zeta/k_i)^{1/2}\textstyle\sum_{j=1}^{k_i} c_{b,i,j}\mathbf{W}_{i,j}, \tag{5.6}$$

for $i = 1, \ldots, n, \quad b = 1, \ldots, B$. With this definition, a little calculation indicates that $E\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = \mathbf{X}_i$ and

$$\text{var}\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = (1+\zeta)\sigma_{u,i}^2/k_i = (1+\zeta)\text{var}(\overline{\mathbf{W}}_{i,\cdot}|\mathbf{X}_i). \tag{5.7}$$

Thus the remeasurements $\mathbf{W}_{b,i}(\zeta)$ from (5.6) have the same key properties as the remeasurements in (5.1) and (5.4), that is, the variances of the error in the remeasurements are inflated by a multiplicative factor that vanishes when $\zeta = -1$, and $\text{MSE}\{\mathbf{W}_{b,i}(\zeta)\} \to 0$ as $\zeta \to -1$.

Because we want to average over B remeasured data sets, we need a way to generate random, replicate versions of (5.6). We do this by making the contrasts random. We get statistical replicates of $\mathbf{W}_{b,i}(\zeta)$ by sampling $\mathbf{c}_{b,i}$ uniformly from the set of all normalized contrast vectors of dimension k_i. This is easily accomplished using pseudorandom

Normal$(0,1)$ random variables. If $Z_{b,i,1},\ldots,Z_{b,i,k_i}$ are Normal$(0,1)$, then

$$c_{b,i,j} = \frac{Z_{b,i,j} - \overline{Z}_{b,i,\cdot}}{\sqrt{\sum_{j=1}^{k_i}\left(Z_{b,i,j} - \overline{Z}_{b,i,\cdot}\right)^2}}, \tag{5.8}$$

are such that $\sum_j c_{b,i,j} = 0$ and $\sum_j c_{b,i,j}^2 = 1$. Furthermore, the random contrast vector $\mathbf{c}_{b,i} = (c_{b,i,1},\ldots,c_{b,i,k_i})^t$ is uniformly distributed on the set of all normalized contrast vectors of dimension k_i (Devanarayan and Stefanski, 2002).

The averaged naive estimates, $\widehat{\Theta}(\zeta)$, are calculated in exactly the same way as for the previous two cases. Also, the SIMEX estimator, $\widehat{\Theta}_{\text{simex}}$, is again obtained by modeling the relationship between $\widehat{\Theta}(\zeta)$ and ζ, and extrapolating to $\zeta = -1$. Because this version of SIMEX generates pseudo errors from the observed data (via the random contrasts), we call it *empirical* SIMEX to distinguish it from versions of SIMEX that generate pseudo errors from a parametric normal model, for example, the Normal$(0,\sigma_u^2)$ model.

5.3.1.4 Nonadditive Measurement Error

Thus far, we have described the SIMEX algorithm for additive measurement error models. However, SIMEX applies more generally and is often easily extended to other error models (Eckert, Carroll, and Wang, 1997).

For example, consider multiplicative error. Taking logarithms transforms the multiplicative model to the additive model, but as discussed in Section 4.5, some investigators feel that the most appropriate predictor of $\mathbf{Y}$ is $\mathbf{X}$ on the original, not log, scale. In regression calibration, multiplicative error is handled in special ways; see Section 4.5. SIMEX works more naturally, in that one performs the simulation step (5.1) on the logarithm of $\mathbf{W}$, and not on $\mathbf{W}$ itself. To see this, suppose that the observed data error model is

$$\log(\mathbf{W}_i) = \log(\mathbf{X}_i) + \mathbf{U}_i,$$

where $\mathbf{U}_i$ are Normal$(0,\sigma_u^2)$. The remeasured data are obtained as

$$\log\{\mathbf{W}_{b,i}(\zeta)\} = \log(\mathbf{W}_i) + \sqrt{\zeta}\mathbf{U}_{b,i},$$

where $\mathbf{U}_{b,i}$ are Normal$(0,\sigma_u^2)$ pseudorandom variables. Note that upon transformation

$$\mathbf{W}_{b,i}(\zeta) = \exp\{\log(\mathbf{W}_i) + \sqrt{\zeta}\mathbf{U}_{b,i}\}. \tag{5.9}$$

In the previous three examples, the key property of the remeasured data was the fact that variance was increased by the multiplicative factor $1+\zeta$ — see equations (5.2), (5.5) and (5.7) — and that this multiplier vanishes when $\zeta = -1$. The multiplicative model has a similar property.

However, because the error model is not unbiased on the natural scale ($E(\mathbf{W}_i|\mathbf{X}_i) = \mathbf{X}_i e^{\sigma_u^2/2} \neq \mathbf{X}_i$), the relevant measure is mean squared error, not variance. Tedious but routine calculations show that, for the multiplicative model,

$$\text{MSE}\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = c(\zeta,\sigma_u^2)\ \text{MSE}\{\mathbf{W}_i|\mathbf{X}_i\}, \tag{5.10}$$

where

$$c(\zeta,\sigma_u^2) = \frac{\{e^{\sigma_u^2(1+\zeta)} - 1\}^2 + e^{\sigma_u^2(1+\zeta)}\{e^{\sigma_u^2(1+\zeta)} - 1\}}{\{e^{\sigma_u^2} - 1\}^2 + e^{\sigma_u^2}\{e^{\sigma_u^2} - 1\}},$$

is such that $c(0,\sigma_u^2) = 1$, $c(\zeta,\sigma_u^2)$ is increasing in $\zeta > 0$ for all σ_u^2, and $\lim_{\zeta\to-1} c(\zeta,\sigma_u^2) = 0$. Thus (5.10) is the biased error model counterpart of (5.2), (5.5), and (5.7).

In the multiplicative model used above, $\mathbf{W}_i$ is biased for $\mathbf{X}_i$ because $\mathbf{U}_i$ is assumed to have a mean of 0. An alternative assumption is that $\mathbf{W}_i$ is unbiased for $\mathbf{X}_i$, which requires that $E(\mathbf{U}_i) = -\sigma_u^2/2$. This assumption was used in Section 4.5.2. Either assumption is plausible but, unfortunately, neither can be checked without validation data. If one is certain that $E(\mathbf{U}_i) = 0$, then one might divide $\mathbf{W}_i$ by $\exp(\sigma_u^2/2)$ to get a surrogate that is unbiased. However, we did not do this here because, by definition, the *naive* analysis is to leave $\mathbf{W}_i$ unchanged.

For more general, nonmultiplicative error models, suppose that we can transform $\mathbf{W}$ to an additive model by a transformation $\mathcal{H}$, so that $\mathcal{H}(\mathbf{W}) = \mathcal{H}(\mathbf{X}) + \mathbf{U}$. This is an example of the transform-both-sides model; see (4.21). If $\mathcal{H}$ has an inverse function $\mathcal{G}$, then the simulation step generates

$$\mathbf{W}_{b,i}(\zeta) = \mathcal{G}\left\{\mathcal{H}(\mathbf{W}_i) + \sqrt{\zeta}\mathbf{U}_{b,i}\right\}.$$

In the multiplicative model, $\mathcal{H} = \log$ and $\mathcal{G} = \exp$. A standard class of transformation models is the power family discussed in Section 4.7.3. If replicate measurements are available, one can also investigate the appropriateness of different transformations; see Section 1.7 for a detailed discussion. As mentioned there, after transformation the standard deviation of the intraindividual replicates should be uncorrelated with their mean, and one can find the power transformation which makes the two uncorrelated.

We now present a simple, yet instructive example with multiplicative measurement error. In addition to illustrating the SIMEX method in a nonadditive error model, the example also shows that estimators of the same parameter can be differentially affected by measurement error, and that SIMEX provides insight into the differential sensitivity of estimators to measurement error.

The true-data model is regression through the origin,

$$\mathbf{Y}_i = \beta \mathbf{X}_i + \epsilon_i$$

where the equation errors have mean zero and finite variances, and the error model for the observed data is additive on the log scale

$$\log(\mathbf{W}_i) = \log(\mathbf{X}_i) + \mathbf{U}_i, \tag{5.11}$$

where $\mathbf{U}_i$ are Normal$(0, \sigma_u^2)$ with the error variance assumed known.

We consider five weighted least squares estimators with weights proportional to powers of the predictor. The true-data estimators considered are

$$\widehat{\beta}_{(p)} = \frac{\sum_1^n \mathbf{Y}_i \mathbf{X}_i^{1-p}}{\sum_1^n \mathbf{X}_i^{2-p}}, \tag{5.12}$$

for $p = 0$ (ordinary least squares), $p = 1/2$, $p = 1$ (ratio estimation), $p = 3/2$, and $p = 2$ (mean of ratios). The corresponding naive estimators, $\widehat{\beta}_{(p),\text{naive}}$, are obtained by replacing $\mathbf{X}_i$ with $\mathbf{W}_i$ in (5.12). In the absence of measurement error, all five estimators are unbiased, and the choice among them would be made on the basis of efficiency, as dictated by the assumed or modeled heteroscedasticity in ϵ_i.

We generated a data set of size $n = 100$ from the regression-through-the-origin model with the ϵ_i independent and identically distributed Normal$(0, \sigma_\epsilon^2)$, and the predictors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ distributed as a shifted and scaled chi-squared, $\mathbf{X}_i = (\chi_5^2 + 1)/\sqrt{46}$ where $E(\mathbf{X}_i^2) = 1$, $\sigma_\epsilon^2 = 0.125$, $\sigma_u^2 = 0.25$, and $\beta = 1.0$. Then we applied the multiplicative error model SIMEX procedure (5.9) for each of the five estimators (5.12).

The error-free data pairs $(\mathbf{X}_i, \mathbf{Y}_i)$ are plotted in the top-left panel of Figure 5.2, and observed data pairs $(\mathbf{W}_i, \mathbf{Y}_i)$ in the top-right panel. The lower-left panel displays an overlay of the points generated in the SIMEX simulation step $(B = 500)$ for each of the five estimators. A corresponding overlay of the SIMEX extrapolations appears in the lower-right panel. Quadratic extrapolant functions were used. The five recognizable point plots in the lower left panel and the five curves in the lower right panel, in order lower to upper, correspond to the estimators with $p = 0,\ 1/2,\ 1,\ 3/2$ and 2.

Note that although the five estimators are differentially affected by measurement error, the SIMEX estimator for each is corrected appropriately, as evidenced by the clustering of the extrapolations to $\zeta = -1$ around the true parameter values $\beta = 1$ in the lower-right panel. In this example, the simple quadratic extrapolant adequately adjusts for bias.

Figure 5.2 indicates that measurement error attenuates the weighted least squares estimators (5.12) with $p = 0,\ 1/2$, and 1 (decreasing curves); *expands* the estimator (bias away from zero) with $p = 2$ (in-

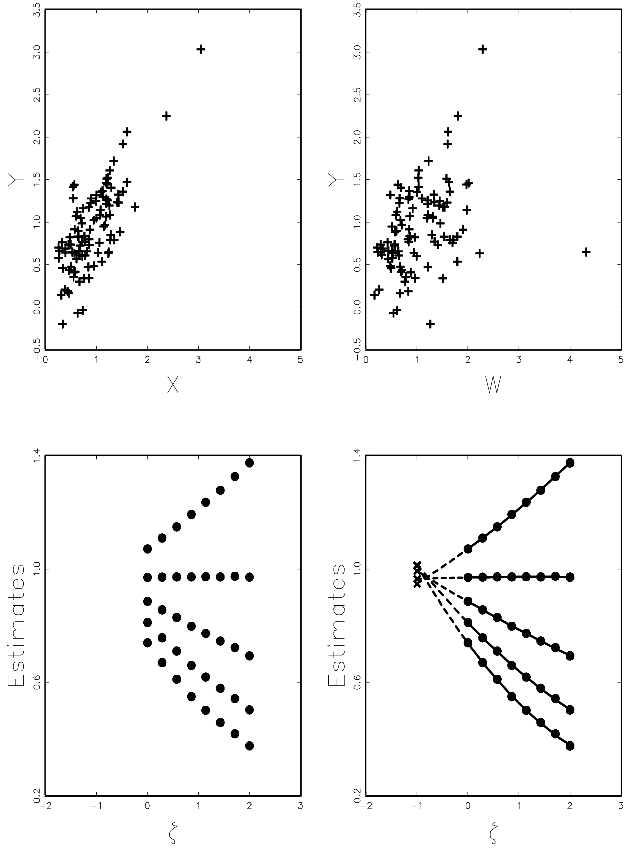


Figure 5.2 *Regression through the origin: weighted least squares estimation with multiplicative measurement error. Top left, true data; top right, observed data; bottom left, $\widehat{\beta}_{(p)}(\zeta)$ estimates calculated in the simulation step $(B = 500)$; bottom right, extrapolation with quadratic extrapolant; bottom two plots, $p = 0,\ 1/2,\ 1,\ 3/2,\ 2$, lower to upper.*

creasing curve); and has no biasing effects on the estimator with $p = 1.5$ (horizontal curve). Remember that *all* of the estimators are consistent with error-free data. Thus, this example shows that bias can depend on the method of estimation, in addition to showing that expansion is possible.

For this simple model, we can do the mathematics to explain the apparent trends in Figure 5.2. Using properties of the normal distribution moment generating function, one can show that as $n \to \infty$, $\widehat{\beta}_{(p),\text{Naive}} \to \beta_{(p)}$ where

$$\beta_{(p)} = \beta\exp\{(2p-3)\sigma_u^2/2\}. \tag{5.13}$$

The exponent in (5.13) is negative when $p < 1.5$ (attenuation), positive when $p > 1.5$ (expansion), and equal to zero when $p = 1.5$ (no bias). Thus, among the class of weighted least squares estimators (5.12), there is one estimator that is robust to measurement error: $\widehat{\beta}_{(p)}$ with $p = 1.5$.

A point we want to emphasize is that the estimators from the SIMEX extrapolation step revealed the robustness of $\widehat{\beta}_{(1.5)}$ quite convincingly, and it can do so for more complicated estimators for which mathematical analysis is intractable. Huang, Stefanski, and Davidian (2006) presented methods for testing the robustness of estimators to measurement error using estimates from the SIMEX simulation step. An overview of their method is given in Section 5.6.3.

Finally, we note that the measurement error robustness of the weighted least squares estimator with $p = 1.5$ depends critically on the assumed error model (5.11). Had we started with an error model for which $\mathbf{W}$ is unbiased for $\mathbf{X}$ on the untransformed scale, that is, $E(\mathbf{W}|\mathbf{X}) = \mathbf{X}$, then it is readily seen the usual ratio estimator ($p = 1$) is consistent for β.

5.3.2 Extrapolant Function Considerations

It follows from the results in Stefanski and Cook (1995) that, under fairly general conditions, asymptotically there is a function of ζ, that, when extrapolated to $\zeta = -1$, the true parameter is obtained. However, this function is seldom known, so it is usually estimated by one of a few simple functional forms. This is what makes SIMEX an approximate method in practice.

As mentioned previously, SIMEX is closely related to the Monte Carlo corrected score (MCCS) method described in Chapter 7. In fact, MCCS is an asymptotically exact-extrapolant version of SIMEX for models satisfying certain smoothness conditions. In other words, MCCS is a version of SIMEX that avoids extrapolation functions. However, MCCS is both mathematically and computationally more involved, whereas SIMEX only requires repeated application of the naive estimation method. Because there are certain regression models for which the asymptotic functional forms are known, and these provide good approximate extrapolant functions for use in other models, SIMEX remains an attractive alternative to MCCS.

In Section 5.3.4.1, we will define what we mean by *non-iid pseudo errors*. In multiple linear regression with these non-iid pseudo errors, the extrapolant function,

$$\mathcal{G}_{\text{RL}}(\zeta,\ \Gamma) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \zeta} = \frac{\gamma_1\gamma_3 + \gamma_2 + \gamma_1\zeta}{\gamma_3 + \zeta}, \tag{5.14}$$

where $\Gamma = (\gamma_1, \gamma_2, \gamma_3)^t$, reproduces the usual method-of-moments estimators; see Section 5.5.1. Because $\mathcal{G}_{\text{RL}}(\zeta,\ \Gamma)$ is a ratio of two linear functions we call it the *rational linear extrapolant*.

SIMEX can be automated in the sense that $\mathcal{G}_{\text{RL}}(\zeta, \Gamma)$ can be employed to the exclusion of other functional forms. However, this is not recommended, especially in new situations where the effects of measurement error are not reasonably well understood. For one thing, as described below and seen in Küchenhoff and Carroll (1995), sometimes the rational linear extrapolant has wild behavior. SIMEX is a technique for studying the effects of measurement error in statistical models and approximating the bias due to measurement error. The extrapolation step should be approached as any other modeling problem, with attention paid to adequacy of the extrapolant based on theoretical considerations, residual analysis, and possibly the use of linearizing transformations. Of course, extrapolation is risky in general even when model diagnostics fail to indicate problems, and this should be kept in mind.

In many problems of interest the magnitude of the measurement error variance, σ_u^2, is such that the curvature in the best or "true" extrapolant function is slight and is adequately modeled by either $\mathcal{G}_{\text{RL}}(\zeta, \Gamma)$ or the simple quadratic extrapolant,

$$\mathcal{G}_{\text{Q}}(\zeta, \Gamma) = \gamma_1 + \gamma_2\zeta + \gamma_3\zeta^2. \tag{5.15}$$

An advantage of the quadratic extrapolant is that it is often numerically more stable than $\mathcal{G}_{\text{RL}}(\zeta, \Gamma)$. Instability of the rational linear extrapolant can occur, for example, when the effects of measurement error on a parameter are negligible and a constant, or nearly constant, extrapolant function is required. Such situations arise, for example, with the coefficient of an error-free covariate $\mathbf{Z}$ that is uncorrelated with $\mathbf{W}$. In this case, in (5.14) $\gamma_2 \approx 0$ and γ_3 is nearly unidentifiable. In cases where $\mathcal{G}_{\text{RL}}(\zeta, \Gamma)$ is used to model a nearly horizontal line, $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are well determined, but $\widehat{\gamma}_3$ is not. Problems arise when $0 < \widehat{\gamma}_3 < 1$, for then the fitted model has a singularity in the range of extrapolation $[-1, 0)$. The problem is easily solved by fitting $\mathcal{G}_{\text{Q}}(\zeta, \Gamma)$ in these cases. The quadratic extrapolant typically results in conservative corrections for attenuation; however, the increase in bias is often offset by a reduction in variability. Of course, problems with the rational linear extrapolant need not be confined to situations as just described.

Simulation evidence and our experience with applications thus far suggest that the extrapolant be fit for ζ in the range $[0, \zeta_{\max}]$, where $1 \leq \zeta_{\max} \leq 2$. We denote the grid of ζ values employed by Λ, that is, $\Lambda = (\zeta_1, \zeta_2, \ldots, \zeta_M)$, where typically $\zeta_1 = 0$ and $\zeta_M = \zeta_{\max}$.

The quadratic extrapolant is a linear model and thus is easily fit. The rational linear extrapolant generally requires a nonlinear least squares program to fit the model. However, it is possible to obtain exact analytic fits to three points, and this provides a means of obtaining good starting values.

Let $\zeta_0^* < \zeta_1^* < \zeta_2^*$ and define $d_{ij} = a_i - a_j$, $0 \leq i < j \leq 2$. Then fitting $\mathcal{G}_{\mathrm{RL}}(\zeta, \Gamma)$ to the points $\{a_j,\ \zeta_j^*\}_0^2$ results in parameter estimates

$$\begin{aligned}
\widehat{\gamma}_3 &= \frac{d_{12}\zeta_2^*(\zeta_1^* - \zeta_0^*) - \zeta_0^* d_{01}(\zeta_2^* - \zeta_1^*)}{d_{01}(\zeta_2^* - \zeta_1^*) - d_{12}(\zeta_1^* - \zeta_0^*)} \\
\widehat{\gamma}_2 &= \frac{d_{12}(\widehat{\gamma}_3 + \zeta_1^*)(\widehat{\gamma}_3 + \zeta_2^*)}{\zeta_2^* - \zeta_1^*} \\
\widehat{\gamma}_1 &= a_0 - \frac{\widehat{\gamma}_2}{\widehat{\gamma}_3 + \zeta_0^*}.
\end{aligned}$$

An algorithm we employ successfully to obtain starting values for fitting $\mathcal{G}_{\mathrm{RL}}(\zeta, \Gamma)$ starts by fitting a quadratic model to $\{\zeta_m, \widehat{\theta}(\zeta_m)\}_1^M$, where the ζ_m are equally spaced over $[0,\ \zeta_{\max}]$. Initial parameter estimates for fitting $\mathcal{G}_{\mathrm{RL}}(\zeta, \Gamma)$ are obtained from a three-point fit to $(\widehat{a}_j, \zeta_j^*)_0^2$, where $\zeta_0^* = 0$, $\zeta_1^* = \zeta_{\max}/2$, $\zeta_2^* = \zeta_{\max}$ and $\widehat{a}_j$ is the predicted value corresponding to ζ_j^* from the fitted quadratic model. In our experience, initial values obtained in this fashion are generally very good and frequently differ insignificantly from the fully iterated, nonlinear least squares parameter estimates.

5.3.3 SIMEX Standard Errors

Inference for SIMEX estimators can be performed either via the bootstrap or the theory of M-estimators (Section A.6), in particular by means of the sandwich estimator. Because of the computational burden of the SIMEX estimator, the bootstrap requires considerably more computing time than do other methods. Without efficient implementation of the estimation scheme at each step, even with current computing resources the SIMEX bootstrap may take an inconveniently long (clock) time to compute. In STATA's implementation of generalized linear models with measurement error (see http://www.stata.com/merror), the implementation is extremely efficient, and bootstrap standard errors for SIMEX take place in fast (clock) time. Even with our own implementation, most bootstrap applications take place in a reasonable (clock) time.

Asymptotic covariance estimation methods based on the sandwich estimator are described in Section B.4.2. These are easy to implement in specific applications but require additional programming. However, when σ_u^2 is known or nearly so, the SIMEX calculations themselves admit a simple standard error estimator. Here, we consider only the case of homoscedastic measurement error. For the case of heteroscedastic error and empirical SIMEX, see Devanarayan (1996).

Let $\widehat{\tau}_b^2(\zeta)$ be any variance estimator attached to $\widehat{\Theta}_b(\zeta)$, for example, the sandwich estimator or the inverse of the information matrix, and let $\widehat{\tau}^2(\zeta)$ be their average for $b = 1, \ldots, B$. Let $s_\Delta^2(\zeta)$ be the sample covariance matrix of the terms $\widehat{\Theta}_b(\zeta)$ for $b = 1, \ldots, B$. Then, as shown in Section B.4.1, variance estimates for the SIMEX estimator can be obtained by extrapolating the components of the differences, $\widehat{\tau}^2(\zeta) - s_\Delta^2(\zeta)$, to $\zeta = -1$. When $\widehat{\tau}^2(\zeta)$ is estimated by the Fisher information matrix or sandwich formula, then the extrapolant is called the *SIMEX Information* or *SIMEX Sandwich* variance estimator, respectively.

5.3.4 Extensions and Refinements

5.3.4.1 Modifications of the Simulation Step

There is a simple modification to the simulation step that is sometimes useful. As described above, the pseudo errors are generated independently of $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n$ as Normal$(0, \sigma_u^2)$ random variables. The Monte Carlo variance in $\widehat{\Theta}(\zeta)$ can be reduced by the use of pseudo errors constrained so that for each fixed b, the sequence $(\mathbf{U}_{b,i})_{i=1}^n$ has mean zero, population variance σ_u^2, that is, $\sum_{i=1}^n \mathbf{U}_{b,i}^2 = n\sigma_u^2$, and its sample correlations with $(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_i)_1^n$ are all zero. We call pseudo errors constrained in this manner *non-iid pseudo errors*. In some simple models, such as linear regression, the Monte Carlo variance is reduced to zero by the use of non-iid pseudo errors.

The non-iid pseudo errors are generated by first generating independent standard normal pseudo errors $(\mathbf{U}_{b,i}^*)_1^n$. Next, fit a linear regression model of the pseudo errors on $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n$, including an intercept. The non-iid pseudo errors are obtained by multiplying the residuals from this regression by the constant

$$c = \left[n\sigma_u^2/\{(n-p-1)\ \mathrm{MSE}\ \}\right]^{1/2},$$

where MSE is the usual linear regression mean squared error, and p is the dimension of $(\mathbf{Y}, \mathbf{Z}^t, \mathbf{W}^t)^t$.

5.3.4.2 Estimating the Measurement Error Variance

When the measurement error variance σ_u^2 is unknown, it must be estimated with auxiliary data, as described in Chapter 4; see especially (4.3). The estimate is then substituted for σ_u^2 in the SIMEX algorithm, and standard errors are calculated as described in Section B.4.2.

5.3.5 Multiple Covariates with Measurement Error

So far, it has been assumed that $\mathbf{X}$ is scalar. For the case of a multivariate $\mathbf{X}$, only a minor change is needed. Suppose that $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$ and $\mathbf{U}_i$ is Normal$(0, \Sigma_u)$, that is, $\mathbf{U}_i$ is multivariate normal with mean zero and covariance matrix Σ_u. Then, to generate the pseudo errors we again use (5.4) and ζ remains a scalar, the only change being that now $\mathbf{U}_{b,i}$ is generated as Normal$(0, \Sigma_u)$. Note that we again have $E\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = \mathbf{X}_i$ and

$$\text{var}\{\mathbf{W}_{b,i}(\zeta)|\mathbf{X}_i\} = (1+\zeta)\Sigma_u = (1+\zeta)\text{var}(\mathbf{W}_i|\mathbf{X}_i), \qquad (5.16)$$

which is the multivariate counterpart of (5.2). Extrapolation is, in principle, the same as in the scalar $\mathbf{X}$ case because ζ is a scalar even for multivariate $\mathbf{X}$. However, the number of remeasured data sets, B, required to achieve acceptable Monte Carlo estimation precision will generally need to be larger when there are multiple covariates measured with error. This is because the Monte Carlo averaging in the simulation step, see (5.3), is effectively a means of numerical integration. As with any numerical integration method, higher-dimensional integration requires greater computational effort for comparable levels of precision. Also, less is known about the general utility of the simple extrapolant functions, for example, the quadratic, in the multivariate $\mathbf{X}$ case, especially for data with large measurement error variances and either strong multicollinearity among the $\mathbf{X}$ variables or high correlation among the measurement errors.

5.4 Applications

5.4.1 Framingham Heart Study

We illustrate the methods using data from the Framingham Heart Study, correcting for bias due to measurement error in systolic blood pressure and serum cholesterol measurements. The Framingham study consists of a series of exams taken two years apart. We use Exam #3 as the baseline. There are 1,615 men aged 31–65 in this data set, with the outcome, $\mathbf{Y}$, indicating the occurrence of coronary heart disease (CHD) within an eight-year period following Exam #3; there were 128 cases of CHD. Predictors employed in this example are the patient's age at Exam #2, smoking status at Exam #1, serum cholesterol at Exam #3, and systolic blood pressure (SBP) at Exam #3, the last is the average of two measurements taken by different examiners during the same visit.

In order to illustrate the various SIMEX methods we do multiple analyses. In the first set of analyses, we treat serum cholesterol as error free, so that the only predictor measured with error is SBP. In these analyses, the error-free covariates $\mathbf{Z}$, are age, smoking status, and serum cholesterol. For $\mathbf{W}$, we employ a modified version of a transformation originally due to Cornfield and discussed by Carroll, Spiegelman, Lan, et al. (1984), setting $\mathbf{W} = \log(\text{SBP} - 50)$. Implicitly, we are defining $\mathbf{X}$ as the long-term average of $\mathbf{W}$. In the final analysis, we illustrate SIMEX when there are two predictors measured with error: SBP and serum cholesterol.

5.4.2 Single Covariate Measured with Error

In addition to the variables discussed above, we also have SBP measured at Exam #2. The mean transformed SBP at Exams #2 and #3 are 4.37 and 4.35, respectively. Their difference has mean 0.02, and standard error 0.0040, so that the large-sample test of equality of means has p-value < 0.0001. Thus in fact, the measurement at Exam #2 is not *exactly* a replicate, but the difference in means from Exam #2 to Exam #3 is close to negligible for all practical purposes.

We present two sets of analyses. Both use the full complement of replicate measurements from Exams #2 and #3. We calculate estimates and standard errors for the naive method, regression calibration, and two versions of SIMEX: SIMEX assuming homoscedastic measurement errors, and empirical SIMEX allowing for possibly heteroscedastic errors. The regression calibration and homoscedastic SIMEX analyses use a pooled estimate of σ_u^2 from the full complement of replicates. In this case, the large degrees of freedom for estimating σ_u^2 means that there is very little penalty in terms of added variability for estimating the measurement error variance.

5.4.2.1 SIMEX and Homoscedastic Measurement Error

This analysis uses the replicate SBP measurements from Exams #2 and #3 for all study participants. The transformed data are $\mathbf{W}_{i,j}$, where i denotes the individual and $j = 1, 2$ refers to the transformed SBP at Exams #2 and #3, respectively. The overall surrogate is $\overline{\mathbf{W}}_{i,\cdot}$, the sample mean for each individual. The model is

$$\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j},$$

	Age	Smoke	Chol	LSBP
Naive	.055	.59	.0078	1.70
Sand.	.010	.24	.0019	.39
Info.	.011	.24	.0021	.41
Reg. Cal.	.053	.60	.0077	2.00
Sand.[1]	.010	.24	.0019	.46
Info.[1]	.011	.25	.0021	.49
Sand.[2]	.010	.24	.0019	.46
Bootstrap	.010	.25	.0019	.46
SIMEX	.053	.60	.0078	1.93
Simex, Sand.[3]	.010	.24	.0019	.43
Simex, Info.[3]	.011	.25	.0021	.47
M-est. [4]	.010	.24	.0019	.44
Empirical SIMEX[5]	.054	.60	.0078	1.94
Simex, Sand.	.011	.24	.0020	.44
Simex, Info.	.012	.25	.0021	.47
M-est.	.011	.24	.0020	.44

Table 5.1 *Estimates and standard errors from the Framingham data logistic regression analysis. This analysis assumes that all observations have replicated SBP. "Naive" = the regression on average of replicated SBP. "Sand." = sandwich standard errors. "Info." = information standard errors. Also,* [1] *= calibration function known;* [2] *= calibration function estimated;* [3] *=* σ_u^2 *known;* [4] *=* σ_u^2 *estimated; and* [5] *= Empirical SIMEX with no assumptions on measurement error variances (standard errors computed as for regular SIMEX). Here "Smoke" is smoking status, "Chol" is cholesterol, and "LSBP" is log(SBP−50).*

where the $\mathbf{U}_{i,j}$ have mean zero and variance σ_u^2. The components of variance estimator (4.3) is $\widehat{\sigma}_u^2 = 0.01259$.

We employ SIMEX using $\mathbf{W}_i^* = \overline{\mathbf{W}}_{i,\cdot}$ and $\mathbf{U}_i^* = \overline{\mathbf{U}}_{i,\cdot}$. The sample variance of $(\mathbf{W}_i^*)_1^n$ is $\widehat{\sigma}_{w,*}^2 = 0.04543$, and the estimated measurement error variance is $\widehat{\sigma}_{u,*}^2 = \widehat{\sigma}_u^2/2 = 0.00630$. Thus, the linear model correction for attenuation, that is, the inverse of the reliability ratio, for these data is 1.16, so that there is only a small amount of measurement error. There are 1,614 degrees of freedom for estimating $\widehat{\sigma}_{u,*}^2$ and thus, for practical purposes, the measurement error variance is known.

In Table 5.1, we list the results of the naive analysis that ignores measurement error, the regression calibration analysis, and the SIMEX analysis. For the naive analysis, "Sand." and "Info." refer to the sandwich and information standard errors discussed in Appendix A; the latter is the output from standard statistical packages.

For the regression calibration analysis, the first set of sandwich and information standard errors are those obtained from a standard logistic regression analysis having substituted the calibration equation for $\mathbf{W}$, and ignoring the fact that the equation is estimated. The second set of sandwich standard errors are as described in Section B.3, while the bootstrap analysis uses the methods of Appendix A.

For the SIMEX estimator, M-estimator refers to estimates derived from the theory of Section B.4.2 for the case where σ_u^2 is estimated from the replicate measurements. Sandwich and Information refer to estimates defined in Section B.4.1, with $\widehat{\tau}^2(\zeta)$ derived from the naive sandwich and naive information estimates, respectively. The M-estimation sandwich and SIMEX sandwich standard errors yield nearly identical standard errors because σ_u^2 is so well estimated.

Figure 5.3 contains plots of the logistic regression coefficients $\widehat{\Theta}(\zeta)$ for eight equally spaced values of ζ spanning $[0, 2]$ (solid circles). The points plotted at $\zeta = 0$ are the naive estimates $\widehat{\Theta}_{\text{naive}}$. For this example, $B = 2000$. Because of double averaging over n and B, taking B this large is not necessary in general (see the related discussion of corrected scores for linear regression in Section 7.2.1). However, there is no harm in taking B large, unless computing time is an issue.

The nonlinear least-squares fits of $\mathcal{G}_{\text{RL}}(\zeta, \Gamma)$ to the components of $\{\zeta_m, \widehat{\Theta}(\zeta_m)\}_1^8$ (solid curves) are extrapolated to $\zeta = -1$ (dashed curves) resulting in the SIMEX estimators (crosses). The open circles are the SIMEX estimators that result from fitting quadratic extrapolants, which are essentially the same as the rational linear extrapolants — not surprising given the small amount of measurement error in this example.

We have stated previously that the SIMEX plot displays the effect of measurement error on parameter estimates. This is especially noticeable in Figure 5.3. In each of the four graphs in Figure 5.3, the range of the ordinate corresponds to a one-standard error confidence interval for the naive estimate constructed using the information standard errors. Thus, Figure 5.3 illustrates the effect of measurement error relative to the variability in the naive estimate. It is apparent that the effect of measurement error is of practical importance only on the coefficient of log(SBP − 50).

The SIMEX sandwich and the M-estimator (with σ_u^2 estimated) meth-

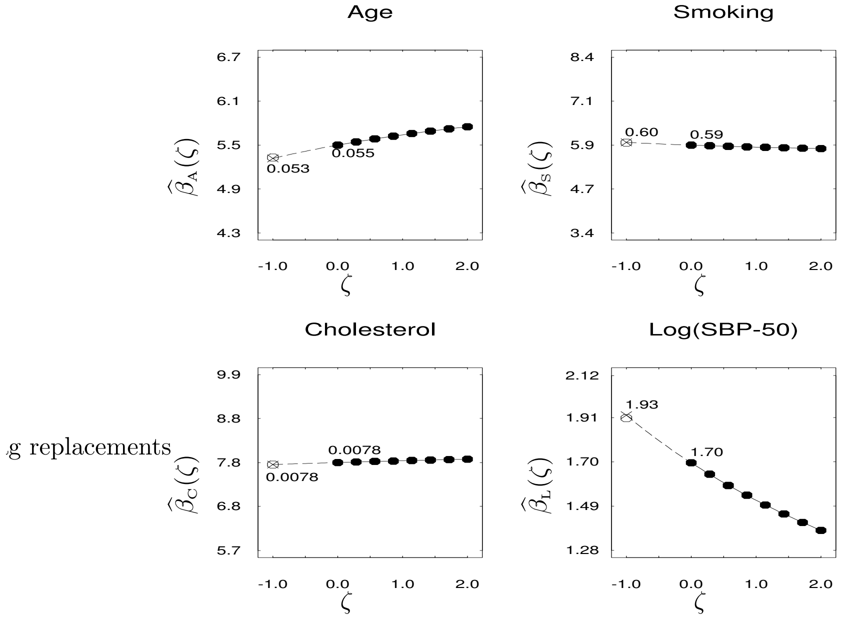


Figure 5.3 *Coefficient extrapolation functions for the Framingham logistic regression modeling. The simulated estimates* $\{\widehat{\beta}_{(\cdot)}(\zeta_m),\ \zeta_m\}_1^8$ *are plotted (solid circles) and the fitted rational linear extrapolant (solid line) is extrapolated to* $\zeta = -1$ *(dashed line), resulting in the SIMEX estimate (cross). Open circles indicate SIMEX estimates obtained with the quadratic extrapolant. The coefficient axis labels for Age are multiplied by* 10^2*, for Smoking by* 10^1*, for Cholesterol by* 10^3*, and for Log(SBP−50) by* 10^0*. Naive and SIMEX estimate values in the graphs are in original units of measurement.*

ods of variance estimation yield similar results in this example. The difference between the SIMEX sandwich and information methods is due to differences in the naive sandwich and information methods for these data.

Figure 5.4 displays the variance extrapolant functions fit to the components of $\widehat{\tau}^2(\zeta)-s_\Delta^2(\zeta)$ used to obtain the SIMEX information variances and standard errors. The figure is constructed using the same conventions used in the construction of Figure 5.3. For these plots, the ranges of the ordinates are $(1/2)\widehat{\text{var}}(\text{naive})$ to $(4/3)\widehat{\text{var}}(\text{naive})$, where $\widehat{\text{var}}(\text{naive})$ is the information variance estimate of the naive estimator.

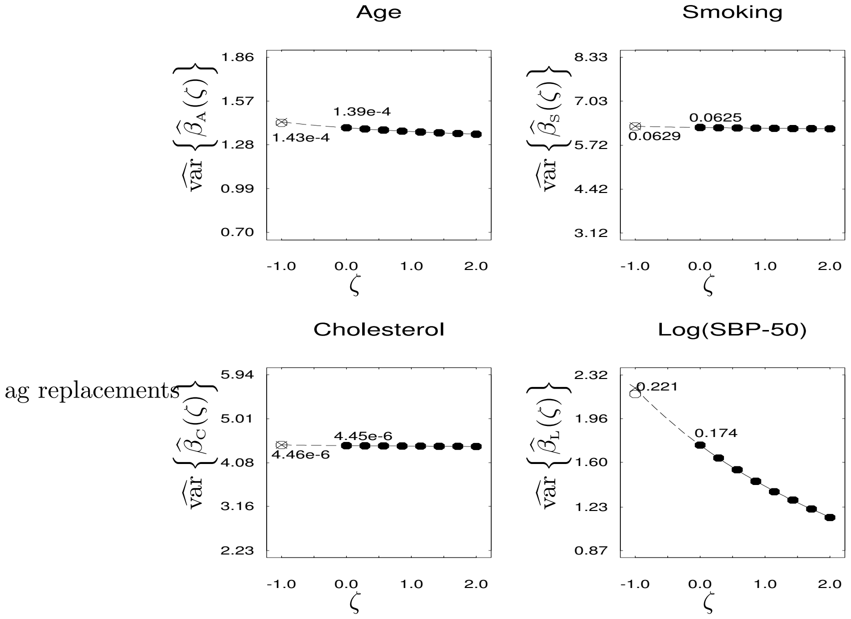


Figure 5.4 *Variance extrapolation functions for the Framingham logistic regression variance estimation. Values of* $\{(\widehat{\tau}^2(\zeta_m) - s_\Delta^2(\zeta_m)),\ \zeta_m\}_1^8$ *for each coefficient estimate (see Section 5.3.3 for definitions of* $\widehat{\tau}^2(\zeta_m)$ *and* $s_\Delta^2(\zeta_m)$*) are plotted (solid circles) and the fitted rational linear extrapolant (solid line) is extrapolated to* $\zeta = -1$ *(dashed line), resulting in the SIMEX variance estimate (cross). Open circles indicate SIMEX variance estimates obtained with the quadratic extrapolant. Naive variance estimates are obtained via the sandwich formula. The coefficient axis labels for Age are multiplied by* 10^4*, for Smoking by* 10^2*, for Cholesterol by* 10^6*, and for Log(SBP−50) by* 10^1*. Naive and SIMEX estimate values in the graphs are in original units of measurement.*

5.4.2.2 Empirical SIMEX and Heteroscedastic Measurement Error

For the analysis in this section, we use the same data as in the previous analysis. However, the model is now

$$\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j},$$

where the $\mathbf{U}_{i,j}$ have mean zero and variance $\sigma_{u,i}^2$. That is, the assumption of homogeneity of variances is not made, and we use empirical SIMEX as described in Section 5.3. The results of the analysis are reported in Table 5.1. For the empirical SIMEX estimator, standard errors were calculated as described in Devanarayan (1996) and are the empirical

SIMEX counterparts of the three versions of regular SIMEX standard errors.

5.4.3 Multiple Covariates Measured with Error

In this section, we consider a model with two predictors measured with error and use it to illustrate both the SIMEX method and the STATA computing environment for SIMEX estimation. The true-data model is similar to that considered in the first analysis in this section. The major difference is that we now regard serum cholesterol as measured with error and use the repeat measurements from Exams #2 and #3 to estimate the measurement error variance.

Preliminary analysis of the duplicate measures of cholesterol indicated that the measurement error is heteroscedastic with variation increasing with the mean. In the previous analyses, cholesterol was regarded as error free, and thus error modeling issues did not arise. Now that we regard cholesterol as measured with error, it makes sense to consider transformations to simplify the error model structure. In this case, simply taking logarithms homogenizes the error variance nicely. This changes our true-data model from the one considered in the preceding section to the logistic model with predictors $\mathbf{Z}_1 =$ age, $\mathbf{Z}_2 =$ smoking status, $\mathbf{X}_1 =$ log(cholesterol) at Exam #3, and $\mathbf{X}_2 =$ log(SBP$-$50) at Exam #3.

The assumed error model is $(\mathbf{W}_1, \mathbf{W}_2) = (\mathbf{X}_1, \mathbf{X}_2) + (\mathbf{U}_1, \mathbf{U}_2)$, where $(\mathbf{U}_1, \mathbf{U}_2)$ is bivariate normal with zero mean and covariance matrix Σ_u. The error covariance matrix was estimated by one-half the sample covariance matrix of the differences between the Exam #2 and Exam #3 measurements of $\mathbf{X}_1$ and $\mathbf{X}_2$, resulting in

$$\widehat{\Sigma}_u = \begin{pmatrix} 0.00846 & 0.000673 \\ 0.000673 & 0.0126 \end{pmatrix}. \tag{5.17}$$

The estimated correlation is small, .065, but significantly different from zero (p-value = .0088), so we do not assume independence of the measurement errors.

The two error variances correspond to marginal reliability ratios of $\lambda_1 = 0.73$ and $\lambda_2 = 0.76$, respectively, for $\mathbf{W}_1$ and $\mathbf{W}_2$. Thus, in the absence of strong multicollinearity, we expect the SIMEX estimates of the coefficients of log(cholesterol) and log(SBP$-$50) to be approximately $1/\lambda_1 = 1.37$ and $1/\lambda_2 = 1.32$ times as large as the corresponding naive estimates.

Following is the STATA code and output for the naive analysis:

```
. qvf firstchd age smoke lcholest3 lsbp3, family(binomial)

Generalized linear models                         No. of obs      =      1615
Optimization     : MQL Fisher scoring             Residual df     =      1610
```

```
                    (IRLS EIM)                    Scale param     =         1
Deviance         =   824.240423                   (1/df) Deviance = .5119506
Pearson          =   1458.82744                   (1/df) Pearson  =  .906104
Variance Function: V(u) = u(1-u)                  [Bernoulli]
Link Function    : g(u) = log(u/(1-u))            [Logit]
Standard Errors  : EIM Hessian

------------------------------------------------------------------------------
    firstchd |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .056446   .0117413     4.81   0.000      .0334334    .0794585
       smoke |    .572659   .2498046     2.29   0.022      .0830509    1.062267
   lcholest3 |   2.039176   .5435454     3.75   0.000      .9738468    3.104506
       lsbp3 |   1.518676   .3889605     3.90   0.000      .7563275    2.281025
       _cons |  -23.39799   3.413942    -6.85   0.000      -30.0892   -16.70679
------------------------------------------------------------------------------
```

The STATA code and output for the SIMEX analysis appear below. Prior to running this STATA code, the elements of the estimated error covariance matrix, $\widehat{\Sigma}_u$ in (5.17) were assigned to V and are input to the SIMEX procedure with the command suuinit(V). One subtlety in STATA is that the order of the predictor-variable variances along the diagonal of V, must correspond to the order of the variables measured with error listed in the STATA simex command. In this example, wcholest is the first variable measured with error and wlsbp is the second, and so their error variances are placed in the $(1, 1)$ and $(2, 2)$ components of V respectively.

```
. simex (firstchd = age smoke) (wcholest:lcholest3)
(wlsbp:lsbp3), family(binomial) suuinit(V) bstrap seed(10008)
Estimated time to perform bootstrap: 2.40 minutes.

Simulation extrapolation                       No. of obs         =      1615
                                               Bootstraps reps    =       199
Residual df  =       1610                      Wald F(4,1610)     =     24.10
                                               Prob > F           =    0.0000
Variance Function: V(u) = u(1-u)               [Bernoulli]
Link Function    : g(u) = log(u/(1-u))         [Logit]

------------------------------------------------------------------------------
             |              Bootstrap
    firstchd |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0545443   .0099631     5.47   0.000      .0350023    .0740863
       smoke |   .5803764   .2638591     2.20   0.028      .0628329     1.09792
    wcholest |     2.5346   .7278619     3.48   0.001      1.106944    3.962256
       wlsbp |    1.84699   .4529421     4.08   0.000      .9585718    2.735408
       _cons |  -27.44831   4.231603    -6.49   0.000     -35.74834   -19.14828
------------------------------------------------------------------------------
```

STATA also provides SIMEX plots for visually assessing the extrapolation step. The variables $\mathbf{Z}_1 =$ age and $\mathbf{Z}_2 =$ smoking status are not affected by measurement error much, so we only present the SIMEX plots for $\mathbf{X}_1 = \log(\text{cholesterol})$ and $\mathbf{X}_2 = \log(\text{SBP}-50)$ in Figure 5.5. Note that whereas we use ζ as the variance inflation factor for SIMEX remeasured data, the default in STATA is to identify this parameter as "Lambda."

Note that with the SIMEX analysis there is substantial bias correction in the coefficients for the coefficients of log(cholesterol) and log(SBP$-$50), but not quite as large as predicted from the inverse marginal reliability ratios, c.f., for log(cholesterol) $(1.37)(2.04) = 2.79$, for log(SBP$-$50) $(1.32)(1.52) = 2.01$. Three factors contribute to the differences. First, whereas the marginal reliability ratios provide a useful rule of thumb for determining bias corrections, they do not account for collinearity among the predictors or for correlation among the measurement errors. The proper rule-of-thumb multiplier in this case is the inverse of the reliability matrix (Gleser, 1992), but because it is matrix-valued it is not as readily computed and so not as useful. Second, there is variability in the SIMEX estimates associated with the choice of B. In STATA the default is $B = 199$, but this can be overridden using the breps() command. For these data, increasing B to 1,000 results in greater corrections for attenuation (we got estimated coefficients for log(cholesterol) and log(SBP$-$50) of 2.65 and 1.86, respectively). Recall that for multiple predictors measured with error, greater replication is necessary. Finally, the default extrapolant in STATA is the quadratic, which generally results in somewhat less correction for bias than the rational linear extrapolant. Using the rational-linear extrapolant, we got estimates for log(cholesterol) and log(SBP$-$50) of 2.76 and 1.89, respectively.

5.5 SIMEX in Some Important Special Cases

This section describes the bias-correction properties of SIMEX in four important special cases.

5.5.1 Multiple Linear Regression

Consider the multiple linear regression model

$$\mathbf{Y}_i = \beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x \mathbf{X}_i + \epsilon_i.$$

In the notation of Section 5.3, $\Theta = (\beta_0, \beta_z^t, \beta_x)^t$. If non-iid pseudo errors are employed in the SIMEX simulation step, it is readily seen that

$$\widehat{\Theta}(\zeta) \quad = \quad \left\{ \sum_{i=1}^{n} \begin{pmatrix} 1 & \mathbf{Z}_i^t & \mathbf{W}_i \\ \mathbf{Z}_i & \mathbf{Z}_i\mathbf{Z}_i^t & \mathbf{Z}_i\mathbf{W}_i \\ \mathbf{W}_i & \mathbf{W}_i\mathbf{Z}_i^t & \mathbf{W}_i^2 + \zeta\sigma_u^2 \end{pmatrix} \right\}^{-1}$$

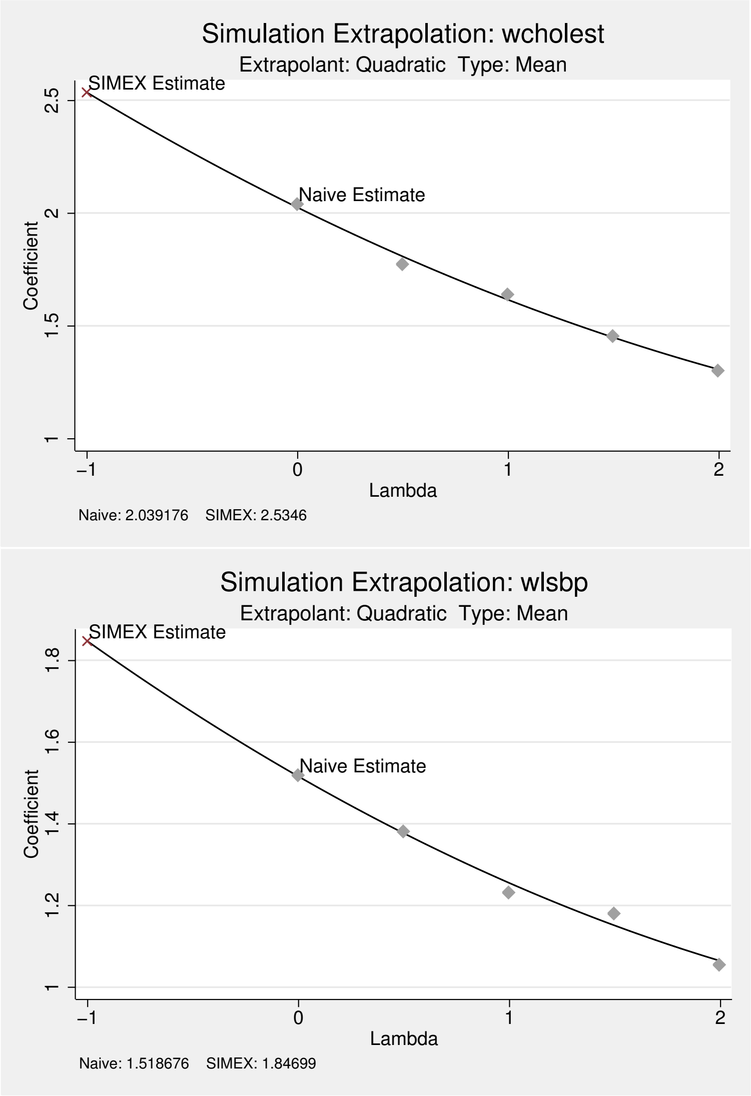


Figure 5.5 *STATA SIMEX plots for log(cholesterol) (top) and log(SBP$-$50) (bottom) for the Framingham logistic regression model with $\mathbf{Z}_1$ = age, $\mathbf{Z}_2$ = smoking status, $\mathbf{X}_1$ = log(cholesterol), and $\mathbf{X}_2$ = log(SBP$-$50). Note that where we use ζ, STATA uses "Lambda."*

$$\times \left\{ \sum_{i=1}^{n} \begin{pmatrix} \mathbf{Y}_i \\ \mathbf{Z}_i \mathbf{Y}_i \\ \mathbf{W}_i \mathbf{Y}_i \end{pmatrix} \right\}.$$

Solving this system of equations we find that

$$\widehat{\beta}_v(\zeta) = (\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{Y} - \frac{(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{W}\left(\mathbf{W}^t\mathbf{Y} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{Y}\right)}{\mathbf{W}^t\mathbf{W} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{W} + \zeta\sigma^2}, \tag{5.18}$$

$$\widehat{\beta}_x(\zeta) = \frac{\mathbf{W}^t\mathbf{Y} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{Y}}{\mathbf{W}^t\mathbf{W} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{W} + \zeta\sigma^2}, \tag{5.19}$$

where $\beta_v = (\beta_0, \beta_z^t)^t$, $\mathbf{V}^t = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n)$ with $\mathbf{V}_i = (1, \mathbf{Z}_i^t)^t$. All of the components of $\widehat{\Theta}(\zeta)$ are functions of ζ of the form $\mathcal{G}_{\mathrm{RL}}(\zeta,\ \Gamma)$ for suitably defined, component-dependent $\Gamma = (\gamma_1, \gamma_2, \gamma_3)^t$.

It follows that if the models fit in the SIMEX extrapolation step have the form $\mathcal{G}_{\mathrm{RL}}(\zeta, \Gamma)$, allowing different Γ for different components, then SIMEX results in the usual method-of-moments estimator of Θ.

5.5.2 *Loglinear Mean Models*

Suppose that $\mathbf{X}$ is a scalar and that $E(\mathbf{Y}|\mathbf{X}) = \exp(\beta_0 + \beta_x\mathbf{X})$, with variance function $\mathrm{var}(\mathbf{Y} \mid \mathbf{X}) = \sigma^2 \exp\{\theta(\beta_0 + \beta_x\mathbf{X})\}$ for some constants σ^2 and θ. It follows from the appendix in Stefanski (1989) that if $(\mathbf{W}, \mathbf{X})$ has a bivariate normal distribution and generalized least squares is the method of estimation, then $\widehat{\beta_0}(\zeta)$ and $\widehat{\beta_x}(\zeta)$ consistently estimate

$$\beta_0(\zeta) = \beta_0 + (1+\zeta)\frac{\mu_x\sigma_u^2\beta_x + \beta_x^2\sigma_x^2\sigma_u^2/2}{\sigma_x^2 + (1+\zeta)\sigma_u^2}$$

and

$$\beta_x(\zeta) = \frac{\beta_x\sigma_x^2}{\sigma_x^2 + (1+\zeta)\sigma_u^2},$$

respectively, where $\mu_x = E(\mathbf{X})$, $\sigma_x^2 = \mathrm{Var}(\mathbf{X})$ and $\sigma_u^2 = \mathrm{Var}(\mathbf{W} \mid \mathbf{X})$.

The rational linear extrapolant is asymptotically exact for estimating both β_0 and β_x.

5.5.3 *Quadratic Mean Models*

There is already a literature on polynomial regression with additive measurement error; see Wolter and Fuller (1982); Stefanski (1989); Cheng and Schneeweiss (1998); Iturria, Carroll, and Firth (1999); and Cheng, Schneeweiss, and Thamerus (2000). Thus, the use of SIMEX in this problem thus might not be considered in practice, but it is still interesting because it is an example where neither the quadratic nor the rational linear extrapolant provides exact answers.

Consider fitting a quadratic regression model using orthogonal polynomials and least square estimation. Components of the parameter vector $\Theta = (\beta_0, \beta_{x,1}, \beta_{x,2})^t$ are the coefficients in the linear model

$$\mathbf{Y}_i = \beta_0 + \beta_{x,1}(\mathbf{X}_i - \overline{\mathbf{X}}) + \beta_{x,2}(\mathbf{X}_i^2 - a - b\mathbf{X}_i) + \epsilon_i, \tag{5.20}$$

where $a = a\{(\mathbf{X}_i)_1^n\}$ and $b = b\{(\mathbf{X}_i)_1^n\}$ are the intercept and slope, respectively, of the least squares regression line of $\mathbf{X}_i^2$ on $\mathbf{X}_i$. Model (5.20) is a reparameterization of the usual quadratic regression model $\mathbf{Y}_i = \beta_0 + \beta_{x,1}\mathbf{X}_i + \beta_{x,2}\mathbf{X}_i^2 + \epsilon_i$. The usual model often has severe collinearity, but the reparameterized model is orthogonal. The so-called naive estimator for this model is obtained by fitting the quadratic regression to $(\mathbf{Y}_i,\ \mathbf{W}_i)_1^n$, noting that $\mathbf{W}_i$ replaces $\mathbf{X}_i$, $i = 1, \ldots, n$, in the definitions of a and b.

Let $\mu_{x,j} = E(\mathbf{X}^j)$, $j = 1, \ldots, 4$. We assume for simplicity that $\mu_{x,1} = 0$ and $\mu_{x,2} = 1$. The exact functional form of $\widehat{\Theta}_b(\zeta)$ is known for this model and is used to show that asymptotically, $\widehat{\Theta}(\zeta)$ converges in probability to $\Theta(\zeta)$ given by

$$\begin{aligned} \beta_0(\zeta) &= \beta_0, \\ \beta_{x,1}(\zeta) &= \frac{\beta_{x,1}\sigma_x^2}{\sigma_x^2 + \delta}, \\ \beta_{x,2}(\zeta) &= \frac{\mu_{x,3}\beta_{x,1}\delta + (1+\delta)\beta_{x,2}(\mu_{x,4} - 1) - \mu_{x,3}^2\beta_{x,2}}{(1+\delta)(\mu_{x,4} - 1 + 4\delta + 2\delta^2) - \mu_{x,3}^2}, \end{aligned}$$

where $\delta = (1+\zeta)\sigma_u^2$.

Note that both $\beta_0(\zeta)$ and $\beta_{x,1}(\zeta)$ are functions of ζ of the form $\mathcal{G}_{\mathrm{RL}}(\zeta, \Gamma)$, whereas $\beta_{x,2}(\zeta)$ is not. For arbitrary choices of σ_u^2, $\mu_{x,3}$, $\mu_{x,4}$, $\beta_{x,1}$, and $\beta_{x,2}$, the shape of $\beta_{x,2}(\zeta)$ can vary dramatically for $-1 \le \zeta \le 2$, thereby invalidating the extrapolation step employing an approximate extrapolant. However, in many practical cases, the quadratic extrapolant corrects for most of the bias, especially for σ_u^2 sufficiently small. When $\mathbf{X}$ is normally distributed, $\beta_{x,2}(\zeta) = \beta_{x,2}/(1+\delta)^2$, which is monotone for all $\zeta \ge -1$ and reasonably well approximated by either a quadratic or $\mathcal{G}_{\mathrm{RL}}(\zeta, \Gamma)$ for a limited but useful range of values of σ_u^2.

5.6 Extensions and Related Methods

5.6.1 *Mixture of Berkson and Classical Error*

We now consider the Berkson/classical mixed error model, which was discussed previously in Section 3.2.5 (see Table 3.1, and also Section 1.8.2 and Section 8.6 for log-scale versions of the model). Recall that the defining characteristic is that the error model contains both classical and Berkson components. Specifically, it is assumed that

$$\begin{aligned}\mathbf{X} &= \mathcal{L} + \mathbf{U}_b, &(5.21)\\ \mathbf{W} &= \mathcal{L} + \mathbf{U}_c. &(5.22)\end{aligned}$$

When $\mathbf{U}_b = 0$, the classical error model is obtained, whereas the Berkson error model results when $\mathbf{U}_c = 0$. The variances of the error terms are $\sigma_{u_c}^2$ and $\sigma_{u_b}^2$. Some features of this error model, when $(\mathbf{X}, \mathbf{W})$ is bivariate normal, that we will use later are:

$$E(\mathbf{W} \mid \mathbf{X}) = (1-\gamma_x)\mu_x + \gamma_x \mathbf{X}, \quad \text{where} \quad \gamma_x = \frac{\sigma_x^2 - \sigma_{u_b}^2}{\sigma_x^2};$$

$$\begin{aligned}\text{cov}(\mathbf{X}, \mathbf{W}) &= \gamma_x \sigma_x^2;\\ \text{var}(\mathbf{W} \mid \mathbf{X}) &= \sigma_w^2 - \gamma_x^2 \sigma_x^2;\\ \text{var}(\mathbf{W}) &= \sigma_x^2 - \sigma_{u_b}^2 + \sigma_{u_c}^2. &(5.23)\end{aligned}$$

Apart from Schafer, Stefanski, and Carroll (1999), SIMEX for this error model has not been considered and is not as well studied as classical-error SIMEX. We now show how to implement SIMEX estimation in this model, assuming that $\sigma_{u_c}^2$ and $\sigma_{u_b}^2$ are known. Define

$$\widehat{\sigma}_x^2 = s_w^2 - \sigma_{u_c}^2 + \sigma_{u_b}^2, \quad \text{and} \quad \widehat{\mu}_x = \overline{\mathbf{W}}, \tag{5.24}$$

where $\overline{W}$ and s_w^2 are the sample mean and variance of the $\mathbf{W}$ data. Then, for $0 \le \zeta \le \zeta_{\text{max}}$ where $\zeta_{\text{max}} \le \zeta_{\text{max}}^* = (\widehat{\sigma}_x^2 - \widehat{\sigma}_{u_b}^2)/\sigma_{u_b}^2$, set

$$\begin{aligned}\widehat{a}_2 &= \frac{\widehat{\sigma}_x^2 - \sigma_{u_b}^2(1+\zeta)}{(\widehat{\sigma}_x^2 - \sigma_{u_b}^2)},\\ \widehat{a}_3 &= +\sqrt{\widehat{\sigma}_x^2 + \sigma_{u_c}^2(1+\zeta) - \sigma_{u_b}^2(1+\zeta) - \widehat{a}_2^2(\widehat{\sigma}_x^2 + \sigma_{u_c}^2 - \sigma_{u_b}^2)}\ ,\\ \widehat{a}_1 &= (1-\widehat{a}_2)\widehat{\mu}_x. &(5.25)\end{aligned}$$

Note that when $\zeta \le \zeta_{\text{max}}^*$ the term under the radical sign is nonnegative, and hence $\widehat{a}_3$ is real. Then the b^{th} error-inflated set of pseudo measurements is defined as

$$\mathbf{W}_{b,i}(\zeta) = \widehat{a}_1 + \widehat{a}_2 \mathbf{W}_i + \widehat{a}_3 \mathbf{U}_{b,i}, \quad i = 1, \cdots, n. \tag{5.26}$$

With the one change that the upper bound of the grid of ζ values must not exceed ζ_{max}^*, the SIMEX algorithm works the same from this point on as it does for the case of classical measurement error.

We now show that under the assumption that $(\mathbf{X}, \mathbf{W})$ is bivariate normal, the remeasured data from (5.26) possess, asymptotically, the key property of remeasured data that we saw for other error models in equations (5.2), (5.5), (5.7), and (5.10). Let a_1, a_2, and a_3 denote quantities defined in (5.25) when s_w^2 and $\overline{\mathbf{W}}$ are replaced by their asymptotic limits to σ_w^2 and μ_x. Now consider the remeasured random variable

$$\mathbf{W}(\zeta) = a_1 + a_2 \mathbf{W} + a_3 \mathbf{U}. \tag{5.27}$$

Noting that $E\{\mathbf{W}(\zeta) \mid \mathbf{X}\} = a_1 + a_2 E(\mathbf{W} \mid \mathbf{X})$ and $\text{var}(\mathbf{W}(\zeta) \mid \mathbf{X}\} = a_2^2 \text{var}(\mathbf{W} \mid \mathbf{X}) + a_3^2$, and that as $\zeta \to -1$, $a_1 \to (1 - 1/\gamma_x)\mu_x$, $a2 \to 1/\gamma_x$ and $a_3 \to \sigma_x^2 - \sigma_w^2/\gamma_x^2$, it follows that as $\zeta \to -1$,

$$E\{\mathbf{W}(\zeta) \mid \mathbf{X}\} \to \mathbf{X} \quad \text{and} \quad \text{var}\{\mathbf{W}(\zeta) \mid \mathbf{X}\} \to 0. \tag{5.28}$$

Thus, just as in the other error models we considered in Section 5.3.1, the mean squared error $\text{MSE}\{\mathbf{W}(\zeta)|\mathbf{X}\} = E[\{\mathbf{W}(\zeta) - \mathbf{X}\}^2|\mathbf{X}]$ converges to zero as $\zeta \to -1$; see (5.10).

5.6.2 Misclassification SIMEX

Küchenhoff, Mwalili, and Lesaffre (2005) developed a general method of correcting for bias in regression and other estimators when discrete data are misclassified, called the *misclassification SIMEX* (MC-SIMEX). In broad strokes, the method works in much the same way as the SIMEX methods discussed previously. However, the details of the method differ, especially the simulation component, which could logically be called *reclassification* in the spirit of the term *remeasurement* used previously. The method requires that the misclassification matrix $\Pi = (\pi_{ij})$ be known or estimable where

$$\pi_{ij} = \text{pr}(\mathbf{W} = i \mid \mathbf{X} = j). \tag{5.29}$$

Note that the case of no misclassification corresponds to having $\Pi = I$, the identity matrix.

In Section 8.4, we discuss an example of misclassification using maximum likelihood methods, when $\mathbf{X}$ is binary. In such cases, maximum likelihood is relatively simple, and there would be little need to use MC-SIMEX.

In continuous-variable SIMEX, remeasured data are generated in the sense that $\mathbf{W}(\zeta)$ is constructed as a measurement of $\mathbf{W}$, in the same manner that $\mathbf{W}$ is a measurement of $\mathbf{X}$. With misclassification, all variables are discrete. Küchenhoff, Mwalili, and Lesaffre (2005) show how to generate reclassified data in the sense that $\mathbf{W}(\zeta)$ is constructed as a misclassified version $\mathbf{W}$, in the same manner that $\mathbf{W}$ is a misclassified version of $\mathbf{X}$. Suppose that Π has the spectral decomposition $\Pi = E\Lambda E^{-1}$, where Λ is the diagonal matrix of eigenvalues and E is the corresponding matrix of eigenvectors. We can now write symbolically $\mathbf{W} = MC[\Pi](\mathbf{X})$, where the misclassification operation, $MC[\Pi](\mathbf{X})$, denotes the generation of the misclassified variable $\mathbf{W}$ from the true variable $\mathbf{X}$ according to the probabilities (5.29). Define $\Pi^\zeta = E\Lambda^\zeta E^{-1}$. In MC-SIMEX reclassified data are generated as

$$\mathbf{W}_{b,i}(\zeta) = MC[\Pi^\zeta](\mathbf{W}_i), \tag{5.30}$$

where the random reclassification step is repeated $b = 1, \ldots, B$ times for each of the $i = 1, \ldots, n$ variables. As in SIMEX, the simulation step in (5.30) is repeated for a grid of ζ values, $0 \le \zeta_1 < \cdots < \zeta_M$. Once the reclassified data are generated, the rest of the SIMEX algorithm is similar to those discussed previously.

The key idea behind continuous-variable SIMEX is that if $\mathbf{W}|\mathbf{X} \sim \text{Normal}(\mathbf{X}, \sigma_u^2)$ and $\mathbf{W}(\zeta)|\mathbf{W} \sim \text{Normal}(\mathbf{W}, \zeta\sigma_u^2)$, then the conditional distribution $\mathbf{W}(\zeta)|\mathbf{X} \sim \text{Normal}(\mathbf{W}, (1+\zeta)\sigma_u^2)$. The analogous property for MC-SIMEX is if $\mathbf{W} = MC[\Pi](\mathbf{X})$ and $\mathbf{W}(\zeta) = MC[\Pi^\zeta](\mathbf{W})$, then $\mathbf{W}(\zeta) = MC[\Pi^{1+\zeta}](\mathbf{X})$, where the three preceding equalities denote equality in distribution. For continuous-variable SIMEX, $\zeta = -1$ corresponds to the case of no measurement error in the sense that $(1+\zeta)\sigma_u^2 = 0$. For MC-SIMEX, $\zeta = -1$ corresponds to the case of no misclassification in the sense that $\Pi^{1+\zeta} = \Pi^0 = I$.

The heuristic explanation of why MC-SIMEX works is similar to the explanation for SIMEX. A statistic calculated from the misclassified data, say $\widehat{\Theta} = \widehat{\Theta}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$, will converge asymptotically to a limiting value that depends on the matrix Π, say $\Theta(\Pi)$. In the case of no misclassification, $\Pi = I$ and the true-data statistic would be consistent, leading to the conclusion that $\Theta(I) = \Theta_0$, the true parameter value. The same statistic calculated from data generated according to (5.30) will converge to $\Theta(\Pi^{1+\zeta})$. Now, if we could model how $\Theta(\Pi^{1+\zeta})$ depends on ζ, then we could extrapolate the model to $\zeta = -1$, resulting in $\lim_{\zeta \to -1} \Theta(\Pi^{1+\zeta}) = \Theta(\Pi^0) = \Theta(I) = \Theta_0$, the true parameter. The extrapolation step does exactly this with the finite-sample data estimates. Küchenhoff, Mwalili, and Lesaffre (2005) investigated the asymptotic true extrapolant function for a number of representative models and concluded that a quadratic extrapolant function and a loglinear extrapolant function are adequate for a wide variety of models.

5.6.3 Checking Structural Model Robustness via Remeasurement

In this section, we briefly describe a useful remeasurement method that has its roots in SIMEX estimation. Huang, Stefanski, and Davidian (2006) show how to use remeasurement and SIMEX-like plots to check the robustness of certain model assumptions in structural measurement error models, such as those described in Chapter 8. The idea is simple, and we have already seen the essence of it in the weighted least squares example in Section 5.3.1. In that example, one weighted least

squares estimator ($p = 1.5$) was robust to measurement error, and the robustness was apparent from the horizontal SIMEX plot in Figure 5.2. The method of Huang, Stefanski, and Davidian (2006) is based on the fact that if an estimator is not biased by measurement error, then its SIMEX plot should be linear with zero slope. They developed this idea for checking robustness of parametric modeling assumptions in structural measurement error models.

We give an overview of the method for a simple structural model of the type in equation (8.7). In a structural model, the $\mathbf{X}_i$ are regarded as random variables. If a parametric model is assumed for the density of $\mathbf{X}$, say $f_X(x, \widetilde{\alpha}_2)$, then the density of the observed data is

$$f_{Y,W}(y, w|\Theta, \widetilde{\alpha}_2, \sigma_u^2) = \int f_{Y|X}(y|x, \Theta) f_{W|X}(w|x, \sigma_u^2) f_X(x, \widetilde{\alpha}_2) dx, \tag{5.31}$$

where $f_{W|X}(w|x, \sigma_u^2)$ is the $\text{Normal}(\mathbf{X}, \sigma_u^2)$ density. The corresponding likelihood for the case σ_u^2 is known is

$$L(\Theta, \widetilde{\alpha}_2) = \prod_{i=1}^{n} f_{Y,W}(\mathbf{Y}_i, \mathbf{W}_i|\Theta, \widetilde{\alpha}_2, \sigma_u^2). \tag{5.32}$$

The appealing features of structural modeling are that inference is based on the likelihood (5.32) and the estimators are consistent and asymptotically efficient as long as the model is correct. However, the Achilles' heel of structural model is specification of the model for $\mathbf{X}$. If this is not correct, then maximum likelihood estimators need not be consistent or efficient. Remeasurement provides a method of checking whether misspecification of the model for $\mathbf{X}$ is causing bias in the parameter of interest θ. The method is based on the observation that if estimators of Θ based on the model $f_{Y,W}(y, w|\Theta, \widetilde{\alpha}_2, \sigma_u^2)$ using data $\{\mathbf{Y}_i, \mathbf{W}_i\}$ are not biased by measurement error, then estimators of Θ based on the model $f_{Y,W}(y, w|\Theta, \widetilde{\alpha}_2, (1+\zeta)\sigma_u^2)$ using remeasured data $\{\mathbf{Y}_i, \mathbf{W}_i(\zeta)\}$ should not be biased by measurement error. Alternatively, if $f_{Y,W}(y, w|\Theta, \widetilde{\alpha}_2, \sigma_u^2)$ is a correct model for $(\mathbf{Y}, \mathbf{W})$, then $f_{Y,W}(y, w|\Theta, \widetilde{\alpha}_2, (1+\zeta)\sigma_u^2)$ is necessarily a correct model for $(\mathbf{Y}, \mathbf{W}(\zeta))$. And in this case, the SIMEX pseudodata estimators $\widehat{\Theta}(\zeta)$ are consistent for the true Θ for all $\zeta > 0$. Consequently, the plot of $\widehat{\Theta}(\zeta)$ versus ζ should be flat. Conversely, if the plot of $\widehat{\Theta}(\zeta)$ versus ζ is not flat, then the model for $\mathbf{X}$ is not correct, assuming the other components of the model are correct.

The suggested procedure is simple. For a given assumed model for $\mathbf{X}$, generate remeasured data sets and calculate $\widehat{\Theta}(\zeta)$, as described in Section 5.3. Then construct a SIMEX plot as in Figure 5.1. If the plot is a flat line, then the indicated conclusion is that the assumed model for

X is robust to bias from measurement error. If the plot is not flat, then the indicated conclusion is that the assumed model for **X** is not robust. Subjective determination from the plot is not necessary. Huang, Stefanski and Davidian (2006) proposed and studied a test statistic for making an objective determination of robustness. A more complete treatment of the robustness in structural measurement error models and details of the test statistic for testing robustness can be found in their paper.

Bibliographic Notes

In addition to STATA's implementation of continuous-variable SIMEX, an **R** implementation of MC-SIMEX (Section 5.6.2) and continuous-variable SIMEX has been written by Wolfgang Lederer, see http://cran.r-mirror.de/src/contrib/Descriptions/simex.html.

Since the original paper by Cook and Stefanski (1994), a number of papers have appeared that extend the basic SIMEX method, adapt it to a particular model, or shed light on its performance via comparisons to other methods; see Küchenhoff and Carroll (1997); Eckert, Carroll, and Wang (1997); Wang, Lin, Gutierrez, and Carroll (1998); Luo, Stokes, and Sager (1998); Fung and Krewski (1999); Lin and Carroll (1999, 2000); Polzehl and Zwanzig (2004); Staudenmayer and Ruppert (2004); Li and Lin (2003a,b); Devanarayan and Stefanski (2002); Kim and Gleser (2000); Kim, Hong, and Jeong (2000); Holcomb (1999); Carroll, Maca, and Ruppert (1999); Jeong and Kim (2003); Li and Lin (2003).

SIMEX has found applications in biostatistics and epidemiology (Marcus and Elias, 1998; Marschner, Emberson, Irwin, et al., 2004; Greene and Cai, 2004), ecology (Hwang and Huang, 2003; Gould, Stefanski, and Pollock, 1999; Solow, 1998; Kangas, 1998), and data confidentiality (Lechner and Pohlmeier, 2004).

Besides MC-SIMEX introduced in Section 5.6.2, there is a large literature on correcting the effects of misclassification of a discrete covariate. See Gustafson (2004) for an extensive discussion and Buonaccorsi, Laake, and Veirød, (2005) for some recent results.

CHAPTER 6

INSTRUMENTAL VARIABLES

6.1 Overview

The methods discussed thus far depend on knowing the measurement error variance, or estimating it, for example, with replicate measurements or validation data. However, it is not always possible to obtain replicates, and thus direct estimation of the measurement error variance is sometimes impossible. In the absence of information about the measurement error variance, estimation of the regression model parameters is still possible, provided the data contain an *instrumental variable* (IV), $\mathbf{T}$, in addition to the unbiased measurement, $\mathbf{W} = \mathbf{X} + \mathbf{U}$.

In later sections, we state more precisely the conditions required of an instrument, as they differ somewhat from one model to another. However, in all cases an instrument must possess three key properties: (i) $\mathbf{T}$ must not be independent of $\mathbf{X}$; (ii) $\mathbf{T}$ must be uncorrelated with the measurement error $\mathbf{U} = \mathbf{W} - \mathbf{X}$; (iii) $\mathbf{T}$ must be uncorrelated with $\mathbf{Y} - E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X})$. In summary, $\mathbf{T}$ must be uncorrelated with all the variability remaining after accounting for $(\mathbf{Z}, \mathbf{X})$. It is of some interest that in certain cases, especially linear regression, $\mathbf{U}$ can be correlated with the variability remaining after accounting for $(\mathbf{Z}, \mathbf{X})$, that is, $\mathbf{Y} - E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X})$, and thus differential measurement error sometimes can be allowed.

One possible source of an instrumental variable is a second, possibly biased, measurement of $\mathbf{X}$ obtained by an independent measuring method. Thus, the assumption that a variable is an instrument is weaker than the assumption that it is a replicate measurement. However, the added generality is gained at the expense of increased variability in bias-corrected estimators relative to cases where the measurement error variance is known or directly estimated. More important, if $\mathbf{T}$ is assumed to be an instrument when it is not, that is, if $\mathbf{T}$ *is* correlated with either $\mathbf{U} = \mathbf{W} - \mathbf{X}$ or $\mathbf{Y} - E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X})$, then instrumental variable estimators can be biased asymptotically, *regardless of the size of the measurement error variance*. So falsely assuming a variable is an instrument can lead to erroneous inferences even in the case of large sample size and small measurement error; see Sections 6.2.2.1 and 6.5.2.

In the Framingham data analysis of Chapter 5, it was assumed explicitly that transformed blood pressure measurements from successive exam periods were replicate measurements, even though a test of the replicate measurements assumption was found to be statistically (although not practically) significant. The same data can also be analyzed under the weaker assumption that the Exam #2 blood pressure measurements are instrumental variables. We do this in Section 6.5 to illustrate the instrumental variable methods.

In this chapter, we restrict attention to the important and common case in which there is a generalized linear model relating $\mathbf{Y}$ to $(\mathbf{Z}, \mathbf{X})$, that is, the mean and variance functions depend on a linear function of the covariates and predictors. Except for the linear model, we also assume that the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{T}, \mathbf{W})$ is linear, although in Section 6.6 other possibilities are considered. In other words, we assume a regression calibration model (Section 2.2), leading to a hybrid combination of classical additive error and regression calibration error. Instrumental variable estimation is introduced in the context of linear models in Section 6.2. We then describe an extension of the linear IV estimators to nonlinear models using regression calibration–like approximations in Section 6.3. An alternative generalization of linear model IV estimation due to Buzas (1997) is presented in Section 6.4. The methods are illustrated by example in Section 6.5. Section 6.6 discusses other approaches. The chapter concludes with some bibliographic notes. Additional technical details are in Appendix B.5.

6.1.1 *A Note on Notation*

In this chapter it is necessary to indicate numerous regression parameters and we adopt the notation used by Stefanski and Buzas (1995). Consider linear regression with mean $\beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t\mathbf{X}$. Then $\beta_{Y|\underline{1}ZX}$ is the coefficient of $\mathbf{1}$, that is, the intercept β_0, in the generalized linear regression of $\mathbf{Y}$ on $\mathbf{1}$, $\mathbf{Z}$ and $\mathbf{X}$. Also, $\beta^t_{Y|1\underline{Z}X} = \beta_z^t$ is the coefficient of $\mathbf{Z}$ in the regression of $\mathbf{Y}$ on $\mathbf{1}$, $\mathbf{Z}$ and $\mathbf{X}$. This notation allows representation of subsets of coefficient vectors, for example,

$$\beta^t_{Y|\underline{1Z}X} = (\beta_{Y|\underline{1}ZX},\ \beta^t_{Y|1\underline{Z}X}) = (\beta_0, \beta_z^t)$$

and, if the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{T})$ has mean $\alpha_0 + \alpha_z^t\mathbf{Z} + \alpha_t^t\mathbf{T}$, then

$$\beta^t_{X|\underline{1ZT}} = (\beta_{X|\underline{1}ZT},\ \beta^t_{X|1\underline{Z}T},\ \beta^t_{X|1Z\underline{T}}) = (\alpha_0, \alpha_z^t, \alpha_t^t).$$

Also, many of the results in this chapter are best described in terms of the composite vectors

$$\begin{aligned} \widetilde{\mathbf{X}} &= (\mathbf{1}, \mathbf{Z}^t, \mathbf{X}^t)^t, & \widetilde{\mathbf{W}} &= (\mathbf{1}, \mathbf{Z}^t, \mathbf{W}^t)^t, \\ \widetilde{\mathbf{T}} &= (\mathbf{1}, \mathbf{Z}^t, \mathbf{T}^t)^t, & \widetilde{\mathbf{U}} &= \widetilde{\mathbf{W}} - \widetilde{\mathbf{X}}. \end{aligned} \tag{6.1}$$

Note that those components of $\widetilde{\mathbf{U}}$ corresponding to the error-free variables $(\mathbf{1}, \mathbf{Z}^t)$ will equal zero.

6.2 Instrumental Variables in Linear Models

6.2.1 *Instrumental Variables via Differentiation*

Much intuition about the manner in which an instrumental variable is used can be obtained by considering the following equations

$$\begin{aligned} \mathbf{Y} &= f(\mathbf{X}) + \epsilon, \\ \mathbf{W} &= \mathbf{X} + \mathbf{U}, \end{aligned}$$

regarding the scalars $\mathbf{Y}$, $\mathbf{X}$, ϵ, $\mathbf{W}$, and $\mathbf{U}$ as *mathematical* variables. Differentiating both sides of the top equation with respect to $\mathbf{T}$, using the chain rule $\partial f/\partial\mathbf{T} = (\partial f/\partial\mathbf{X})(\partial\mathbf{X}/\partial\mathbf{T})$, noting that $\partial\mathbf{X}/\partial\mathbf{T} = \partial\mathbf{W}/\partial\mathbf{T} - \partial\mathbf{U}/\partial\partial\mathbf{T}$, and rearranging terms results in

$$\frac{\partial\mathbf{W}}{\partial\mathbf{T}}\frac{\partial f}{\partial\mathbf{X}} = \partial\mathbf{Y}/\partial\mathbf{T} + (\partial f/\partial\mathbf{X})(\partial\mathbf{U}/\partial\mathbf{T}) - \partial\epsilon/\partial\mathbf{T}.$$

Consequently if $\partial\mathbf{U}/\partial\mathbf{T} = \partial\epsilon/\partial\mathbf{T} = 0$ and $\partial\mathbf{W}/\partial\mathbf{T} \neq 0$, then

$$\frac{\partial f}{\partial\mathbf{X}} = \frac{\partial\mathbf{Y}/\partial\mathbf{T}}{\partial\mathbf{W}/\partial\mathbf{T}}. \tag{6.2}$$

That is, as long as we know how $\mathbf{Y}$ and $\mathbf{W}$ change with $\mathbf{T}$, we can determine the way that f changes with $\mathbf{X}$.

The suggestive analysis above explains the essential features and workings of instrumental variable estimation in linear measurement error models. If an instrument, $\mathbf{T}$, is such that it is not related to $\mathbf{U}$ or ϵ ($\partial\mathbf{U}/\partial\mathbf{T} = \partial\epsilon/\partial\mathbf{T} = 0$) but is related to $\mathbf{X}$ (note that when $\partial\mathbf{U}/\partial\mathbf{T} = 0$, then $\partial\mathbf{W}/\partial\mathbf{T} = \partial\mathbf{X}/\partial\mathbf{T}$ and $\partial\mathbf{X}/\partial\mathbf{T} \neq 0 \Longrightarrow \partial\mathbf{W}/\partial\mathbf{T} \neq 0$), then we can determine how f varies with $\mathbf{X}$ using only the observed variables $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{T}$. For linear models, the essential properties of an instrument are that $\mathbf{T}$ is uncorrelated with $\mathbf{U}$ and ϵ, and is correlated with $\mathbf{X}$.

The derivation of (6.2) depends critically on the lack of relationships between $\mathbf{T}$ and ϵ, and between $\mathbf{T}$ and $\mathbf{U}$, and also on the denominator on the right-hand side of (6.2) being nonzero. If either of the first two conditions is not met, (6.2) would not be an equality (the statistical analogue is bias). If the third condition is violated, we get indeterminacy because of division by zero (the statistical analogue is excessive variability).

6.2.2 Simple Linear Regression with One Instrument

We first consider simple linear regression with one instrument. Suppose that $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{T}$, are scalar random variables such that

$$\begin{aligned} \mathbf{Y} &= \beta_{Y|\underline{1}X} + \mathbf{X}\beta_{Y|1\underline{X}} + \epsilon, \\ \mathbf{W} &= \mathbf{X} + \mathbf{U}, \end{aligned} \tag{6.3}$$

where ϵ and $\mathbf{U}$ have mean zero, and all random variables have finite variances. Define covariances among these variables as follows: $\text{cov}(\mathbf{T}, \mathbf{Y}) = \sigma_{ty}$, $\text{cov}(\mathbf{T}, \mathbf{X}) = \sigma_{tx}$, etc. In order to estimate the slope in the regression of $\mathbf{Y}$ on $\mathbf{X}$, that is, $\beta_{Y|1\underline{X}}$, we only require that

$$\sigma_{t\epsilon} = \sigma_{tu} = 0, \quad \text{and} \quad \sigma_{tx} \neq 0. \tag{6.4}$$

To see this, note that (6.3) implies that

$$\begin{aligned} \text{cov}(\mathbf{T}, \mathbf{Y}) &= \sigma_{ty} = \sigma_{tx}\beta_{Y|1\underline{X}} + \sigma_{t\epsilon} \\ \text{cov}(\mathbf{T}, \mathbf{W}) &= \sigma_{tw} = \sigma_{tx} + \sigma_{tu}, \end{aligned} \tag{6.5}$$

so that if (6.4) holds, then

$$\frac{\text{cov}(\mathbf{T}, \mathbf{Y})}{\text{cov}(\mathbf{T}, \mathbf{W})} = \frac{\sigma_{tx}\beta_{Y|1\underline{X}} + \sigma_{t\epsilon}}{\sigma_{tx} + \sigma_{tu}} = \beta_{Y|1\underline{X}}.$$

Suppose now that $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{T}_i)$ are a sample with the structure (6.3), and that $\widehat{\sigma}_{ty}$ and $\widehat{\sigma}_{tw}$ are the sample covariances of σ_{ty} and σ_{tw}, respectively. Then the *instrumental variable* estimator,

$$\widehat{\beta}^{IV}_{Y|1\underline{X}} = \frac{\widehat{\sigma}_{ty}}{\widehat{\sigma}_{tw}}, \tag{6.6}$$

is a consistent estimator of $\beta_{Y|1\underline{X}}$.

6.2.2.1 IV Estimation Potential Pitfalls

We now use this simple model to illustrate our previous warnings about the potential pitfalls associated with instrumental variable estimation when the IV assumptions are not satisfied. Because sample covariances are consistent estimators, we know that as $n \to \infty$,

$$\widehat{\beta}^{IV}_{Y|1\underline{X}} \longrightarrow \frac{\text{cov}(\mathbf{T}, \mathbf{Y})}{\text{cov}(\mathbf{T}, \mathbf{W})} = \frac{\sigma_{tx}\beta_{Y|1\underline{X}} + \sigma_{t\epsilon}}{\sigma_{tx} + \sigma_{tu}}. \tag{6.7}$$

Consider the right-hand side of (6.7) for various combinations of assumption *violations*: $\sigma_{t\epsilon} \neq 0$; $\sigma_{tu} \neq 0$; $\sigma_{tx} = 0$. For example, if $\sigma_{tx} \neq 0$ and $\sigma_{tu} = 0$, but $\sigma_{t\epsilon} \neq 0$, then the IV estimator has an asymptotic bias given by $\sigma_{t\epsilon}/\sigma_{tx}$, which can have either sign ($\pm$) and can be of any magnitude, depending on how close $|\sigma_{tx}|$ is to zero. Clearly, there are other combinations of assumption violations that also lead to potentially significant biases. Note that such biases are possible *regardless of the size of the measurement error variance*. In fact, even when $\sigma_u^2 = 0$, IV estimation can lead to large biases when the IV assumptions are violated. So the possibility exists that in trying to correct for a small amount of bias due to measurement error, one could introduce a large amount of bias due to an erroneous IV assumption. The message should be clear: *Use IV estimation only when there is convincing evidence that the IV assumptions are reasonable.*

Relative to the asymptotic results in the previous paragraph, the potential pitfalls *can be even greater* with finite samples of data when σ_{tx} is not far from zero. This is because random variation in the denominator, $\widehat{\sigma}_{tw}$, of (6.6), can cause it to be arbitrarily close to zero, in which case the estimator in (6.6) is virtually worthless by dint of its excessive variability. Fortunately, we can gain insight into whether this is a likely problem by testing the null hypothesis $H_0: \ \sigma_{tw} = 0$. This is most easily done by testing for zero slope in the linear regression of $\mathbf{W}$ on $\mathbf{T}$. Instrumental variable estimation is contraindicated unless there is strong evidence that this slope is nonzero (Fuller, 1987, p. 54).

Problems similar to those noted above occur with multiple predictors measured with error and multiple instrumental variables, although the linear algebra for diagnosing and understanding them is more involved. For this more general setting, Fuller (1987, p. 150–154) describes a test analogous to the regression test described above, and also a small-sample modification of (6.6) that, in effect, controls the denominator so that it does not get too close to zero; see Section 6.2.3.1 for details.

6.2.2.2 Technical Generality of the Assumptions

When the IV assumptions (6.4) hold, the consistency of $\widehat{\beta}^{IV}_{Y|1\underline{X}}$ is noteworthy for the lack of other conditions under which it is obtained. Although the representation for $\mathbf{Y}$ in (6.3) is suggestive of the common linear model, consistency does not require the usual linear model assumption that ϵ and $\mathbf{X}$ are uncorrelated. Neither are any conditions required about the relationship between $\mathbf{X}$ and the instrument $\mathbf{T}$ other than that of nonzero covariance in (6.4); nor is it required that $\mathbf{U}$ and ϵ, or $\mathbf{U}$ and $\mathbf{X}$ be uncorrelated. Although very few assumptions are necessary, it does not mean that the various covariances can be arbitrary. The fact that the covariance matrix of the composite vector $(\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{T})$ must be nonnegative definite imposes restrictions on certain variances and covariances. For example, it is impossible to have $\text{corr}(\mathbf{T}, \mathbf{X}) = 0.99$, $\text{corr}(\mathbf{X}, \mathbf{U}) = 0.99$, and $\text{corr}(\mathbf{T}, \mathbf{U}) = 0$.

6.2.2.3 Practical Generality of the Assumptions

The technical generality of the assumptions seems impressive, but as a practical matter is much less so. The assumption is that $\mathbf{T}$ is uncorrelated with ϵ and $\mathbf{U}$, which in practice is often approximately equivalent to assuming that $\mathbf{T}$ is independent of $(\epsilon, \mathbf{U})$. However, $\mathbf{T}$ has to be related to $\mathbf{X}$, and so as a practical matter, $\mathbf{X}$ too has to be independent of $(\epsilon, \mathbf{U})$.

6.2.3 Linear Regression with Multiple Instruments

We now consider multiple linear regression with multiple instruments, starting with the case where the number of instruments is the same as the number of components of $\mathbf{X}$. The case where the number of instruments exceeds the number of predictors is presented at the end of this section. Suppose that the scalar $\mathbf{Y}$ and the composite vectors $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{T}}$ in (6.1) are such that

$$\begin{aligned} \mathbf{Y} &= \widetilde{\mathbf{X}}^t\beta_{Y|\underline{\widetilde{X}}} + \epsilon, \\ \widetilde{\mathbf{W}} &= \widetilde{\mathbf{X}} + \widetilde{\mathbf{U}}, \end{aligned} \tag{6.8}$$

where ϵ and $\mathbf{U}$ have mean zero, and all random variables have finite second moments.

In what follows, covariances are replaced by uncentered, expected crossproduct matrices, for example $\Omega_{\widetilde{\mathbf{T}}\mathbf{Y}} = E(\widetilde{\mathbf{T}}\mathbf{Y})$ in place of $\sigma_{ty} = \text{cov}(\mathbf{T}, \mathbf{Y})$, a consequence of the fact that a column of ones is included in each of $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{T}}$. Let $\dim(\cdot)$ denote the dimension of the argument. The multiple linear regression counterparts of assumptions (6.4) are

$$\Omega_{\widetilde{\mathbf{T}}\epsilon} = \mathbf{0}, \quad \Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{U}}} = \mathbf{0}, \quad \text{and} \quad \text{rank}(\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}) = \dim(\widetilde{\mathbf{X}}), \tag{6.9}$$

The last assumption requires that $\mathbf{T}$ and $\mathbf{X}$ not be independent. As we discussed in detail in Section 6.2.2 for the cases of simple linear regression with one instrument, violations of the key assumptions (assumptions (6.4) for simple linear regression and (6.9) for multiple linear regression), can have severe consequences. The case where the instrument is an independent measurement of $\mathbf{X}$ obtained using a second, independent, method of measurement (possibly biased or with different error variance) is one where the key assumptions (6.9) can be expected to hold a priori. Even in such cases, however, the declaration of independence is seldom infallible. For other cases, often subject matter expertise must be brought to bear on the problem of determining whether the assumptions in (6.9) are reasonable.

It follows from (6.8) that

$$E(\widetilde{\mathbf{T}}\mathbf{Y}) = \Omega_{\widetilde{\mathbf{T}}\mathbf{Y}} = \Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}\beta_{Y|\underline{\widetilde{X}}} + \Omega_{\widetilde{\mathbf{T}}\epsilon},$$

$$E(\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}^t) = \Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}} = \Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}} + \Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{U}}}. \tag{6.10}$$

Equation (6.10) is the multiple linear regression counterpart of (6.5). Note that

$$(\Omega^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}})^{-1}\Omega^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\Omega_{\widetilde{\mathbf{T}}\mathbf{Y}} = (\Omega^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}})^{-1}\Omega^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}\beta_{Y|\underline{\widetilde{X}}} = \beta_{Y|\underline{\widetilde{X}}}.$$

Consequently if we replace expectations by averages, so that for example $\widehat{\Omega}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}} = n^{-1}\sum_{i=1}^n \widetilde{\mathbf{T}}_i\widetilde{\mathbf{W}}_i^t$, then the *instrumental variable* estimator,

$$\widehat{\beta}^{IV}_{Y|\underline{\widetilde{X}}} = (\widehat{\Omega}^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\widehat{\Omega}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}})^{-1}\widehat{\Omega}^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\widehat{\Omega}_{\widetilde{\mathbf{T}}\mathbf{Y}}, \tag{6.11}$$

is a consistent estimator of $\beta_{Y|\underline{\widetilde{X}}}$.

6.2.3.1 Small Sample Modification

Fuller (1987, p. 150–154) described a small sample modification to handle the instability that can be caused by the matrix inversion in (6.11). It solves small sample problems arising with weak predictors, but has less of an ameliorative effect on problems due to violations of the assumptions that $\mathbf{T}$ and ϵ, and $\mathbf{T}$ and $\mathbf{U}$ are uncorrelated.

Let $\mathcal{V} = [\mathbf{Y}, \widetilde{\mathbf{W}}]$, and define $\mathcal{S} = [\mathcal{Y}, \mathcal{W}] = \widetilde{\mathbf{T}}(\widetilde{\mathbf{T}}^t\widetilde{\mathbf{T}})^{-1}\widetilde{\mathbf{T}}^t\mathcal{V}$. Let q be the number of components of $\widetilde{\mathbf{T}}$. Define

$$S_{aa} = \begin{bmatrix} S_{aa11} & S_{aa12} \\ S_{aa21} & S_{aa22} \end{bmatrix} = (n-q)^{-1}\mathcal{V}^t(\mathcal{V} - \mathcal{S}).$$

Let κ be the smallest root of the determinant equation $|\mathcal{S}^t\mathcal{S} - \kappa S_{aa}|$. Let $\alpha > 0$ be a fixed constant, for example, $\alpha = 4$. Fuller proposed the estimator

$$\widehat{\beta}^{IV}_{Y|\underline{\widetilde{X}}} = \{\mathcal{W}^t\mathcal{W} - (\kappa - \alpha)S_{aa22}\}^{-1}\{\mathcal{W}^t\mathcal{Y} - (\kappa - \alpha)S_{aa21}\}.$$

6.2.3.2 Technical Generality of the Result

The consistency of $\widehat{\beta}^{IV}_{Y|\underline{\widetilde{X}}}$ is again noteworthy for the lack of conditions under which it is obtained. The only conditions necessary are those in (6.9), the third of which requires at least as many instruments as variables measured with error. However, consistency does not require that any of the expected crossproduct matrices $\Omega_{\widetilde{\mathbf{X}}\epsilon}$, $\Omega_{\widetilde{\mathbf{X}}\widetilde{\mathbf{U}}}$, $\Omega_{\widetilde{\mathbf{U}}\epsilon}$ equal zero, although again, as for simple linear regression, the fact the covariance matrix of the composite vector $(\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{T})$ must be nonnegative definite imparts certain restriction on these crossproduct matrices. Also, even though certain instrumental variable estimators can be written as functions of the linear least squares regressions of $\mathbf{Y}$ and $\widetilde{\mathbf{W}}$ on $\widetilde{\mathbf{T}}$, the assumption that these regressions are linear in $\widetilde{\mathbf{T}}$ is not necessary.

6.2.3.3 Practical Generality of the Result

As in simple linear regression, for most practical purposes the assumption that $\mathbf{T}$ is uncorrelated with ϵ and $\mathbf{U}$ means that $\mathbf{X}$ is as well.

6.2.3.4 More Instruments than Predictors

Instrumental variable estimation differs somewhat when the number of instrumental variables exceeds the number of variables measured with error, the case we now consider. Our presentation parallels that of Section 6.2.3. We assume the model in (6.8) and pick up the discussion with (6.10). The key difference when $\dim(\widetilde{\mathbf{T}}) > \dim(\widetilde{\mathbf{X}})$ is that there are more equations in (6.10) than there are regression coefficients, and we use generalized inverses in place of ordinary matrix inverses. Considering (6.10), then for any generalized inverse of $\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}$, say

$$\Omega^{-(M)}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}} = (\Omega^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}} M \Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}})^{-1}\Omega^t_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}} M, \quad \text{with } M \text{ nonsingular},$$

it follows that

$$\Omega^{-(M)}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\Omega_{\widetilde{\mathbf{T}}\mathbf{Y}} = \Omega^{-(M)}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}\beta_{Y|\underline{\widetilde{X}}} = \beta_{Y|\underline{\widetilde{X}}}.$$

Consequently if $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{T}_i)$, $i = 1, \ldots, n$ is an iid sample satisfying (6.8), $\widehat{\Omega}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}$ and $\widehat{\Omega}_{\widetilde{\mathbf{T}}\mathbf{Y}}$ are any consistent estimators of $\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}$ and $\Omega_{\widetilde{\mathbf{T}}\mathbf{Y}}$, and $\widehat{M}$ converges in probability to a nonsingular matrix M, then the *instrumental variable* estimator,

$$\widehat{\beta}^{IV,(\widehat{M})}_{Y|\underline{\widetilde{X}}} = \widehat{\Omega}^{-(\widehat{M})}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\widehat{\Omega}_{\widetilde{\mathbf{T}}\mathbf{Y}}, \tag{6.12}$$

is a consistent estimator of $\beta_{Y|\underline{\widetilde{X}}}$.

One means of generating an estimator of the form (6.12) is to first do a multivariate regression of $\widetilde{\mathbf{W}}_i$ on $\widetilde{\mathbf{T}}_i$ and calculate the predicted values from it, $\widehat{\mathbf{X}}_i = \widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}_i$. These predicted values are denoted by $\widehat{\mathbf{X}}_i$ because under the model (6.8), $\widehat{\mathbf{X}}_i = \widehat{\beta}^t_{\widetilde{X}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}_i + \widetilde{\mathbf{U}}^*_i$, where $\widetilde{\mathbf{U}}^*_i = \widehat{\beta}^t_{\widetilde{U}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}_i$ has mean $\mathbf{0}$. Thus, apart from the addition of the mean zero vector $\widetilde{\mathbf{U}}^*_i$, $\widehat{\mathbf{X}}_i$ equals the predicted values that would be obtained from the regression of $\widetilde{\mathbf{X}}_i$ on $\widetilde{\mathbf{T}}_i$. With $\widehat{\mathbf{X}}_i$ so defined, the coefficient vector estimate from the least-squares regression of $\mathbf{Y}_i$ on $\widehat{\mathbf{X}}_i$ can be written as

$$\widehat{\beta}^{IV,(\widehat{M}_*)}_{Y|\underline{\widetilde{X}}} = \widehat{\Omega}^{-(\widehat{M}_*)}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}}\widehat{\Omega}_{\widetilde{\mathbf{T}}\mathbf{Y}},$$

where $\widehat{M}^{-1}_* = \widehat{\Omega}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{T}}} = (n^{-1}\sum\widetilde{\mathbf{T}}_i\widetilde{\mathbf{T}}^t_i)$, $\widehat{\Omega}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{W}}} = (n^{-1}\sum\widetilde{\mathbf{T}}_i\widetilde{\mathbf{W}}^t_i)$ and $\widehat{\Omega}_{\widetilde{\mathbf{T}}\mathbf{Y}} = (n^{-1}\sum\widetilde{\mathbf{T}}_i\mathbf{Y}_i)$.

Alternatively, if we replace $\widehat{M}$ with

$$\widehat{M}^\dagger = \widehat{\Omega}^{-1}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{T}}}\widehat{M}_1\widehat{\Omega}^{-1}_{\widetilde{\mathbf{T}}\widetilde{\mathbf{T}}}$$

in (6.12) for some other nonsingular matrix $\widehat{M}_1$, the resulting estimator can be written as

$$\widehat{\beta}^{IV,(\widehat{M}^\dagger)}_{Y|\underline{\widetilde{X}}} = \widehat{\beta}^{-(\widehat{M}_1)}_{\widetilde{W}|\underline{\widetilde{T}}}\widehat{\beta}_{Y|\underline{\widetilde{T}}}, \tag{6.13}$$

where $\widehat{\beta}^{-(M_1)}_{\widetilde{W}|\underline{\widetilde{T}}} = (\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}M_1\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})^{-1}\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}M_1$ and $\widehat{\beta}_{Y|\underline{\widetilde{T}}}$ is the least squares coefficient estimate in the regression of $\mathbf{Y}_i$ on $\widetilde{\mathbf{T}}_i$. Note the similarity of (6.13) to (6.6).

6.3 Approximate Instrumental Variable Estimation

6.3.1 IV Assumptions

We have taken care to explain the conditions required for instrumental variable estimation in linear models in order to make it easier to understand certain of the conditions we invoke for instrumental variable estimation in nonlinear models. Here, we continue to assume that a parametric model is correctly specified for $E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X})$ and that the measurement error $\mathbf{U}$ is independent of $\mathbf{Z}$.

Whereas lack of correlation is often sufficient when working with first- and second-moment estimators, for example, as in linear regression, it generally is replaced by independence when working with more complicated estimators. Thus, for the remainder of this chapter we work under a stronger set of assumptions than those required for instrumental variable estimation in linear models. These assumptions are the following:

1. $\mathbf{T}$ is correlated with $\mathbf{X}$;
2. $\mathbf{T}$ is independent of the measurement error $\mathbf{U} = \mathbf{W} - \mathbf{X}$ in the surrogate $\mathbf{W}$;
3. $(\mathbf{W}, \mathbf{T})$ is a surrogate for $\mathbf{X}$, in particular $E(\mathbf{Y}^k \mid \mathbf{Z}, \mathbf{X}, \mathbf{W}, \mathbf{T}) = E(\mathbf{Y}^k \mid \mathbf{Z}, \mathbf{X})$ for $k = 1, 2$. The key point here is that both $\mathbf{T}$ and the measurement error $\mathbf{U}$ in the surrogate $\mathbf{W}$ are independent of any variation in the response $\mathbf{Y}$ after accounting for $(\mathbf{Z}, \mathbf{X})$. In linear regression, this means that $\mathbf{T}$ and $\mathbf{U}$ are independent of the residual error ϵ.

With these assumptions, we can derive an alternative explanation of instrumental variable estimation in linear models. Note that

$$\begin{aligned}
E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{T}) &= E\{E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \mathbf{W}, \mathbf{T}) \mid \mathbf{Z}, \mathbf{T}\} \\
&= E\{E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}) \mid \mathbf{Z}, \mathbf{T}\} \\
&= E(\beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}\mathbf{X} \mid \mathbf{Z}, \mathbf{T}) \\
&= \beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}) \\
&= \beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}E(\mathbf{W} - \mathbf{U} \mid \mathbf{Z}, \mathbf{T})
\end{aligned}$$

$$= \quad \beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}E(\mathbf{W} \mid \mathbf{Z}, \mathbf{T}). \quad (6.14)$$

The key steps in this derivation require that $\mathbf{T}$ is a surrogate for $\mathbf{X}$ and that $\mathbf{U}$ is independent of $\mathbf{T}$ and $\mathbf{Z}$. It follows from (6.14) that if $E(\mathbf{W} \mid \mathbf{Z}, \mathbf{T})$ is linear $\mathbf{T}$, that is, $E(\mathbf{W} \mid \mathbf{Z}, \mathbf{T}) = \beta_{W|\underline{1}ZT} + \beta_{W|1\underline{Z}T}\mathbf{Z} + \beta_{W|1Z\underline{T}}\mathbf{T}$, then by equating coefficients of $\mathbf{T}$ we get that $\beta_{Y|1Z\underline{T}} = \beta_{W|1Z\underline{T}}\beta_{Y|1Z\underline{X}}$, and consequently that $\beta_{Y|1Z\underline{X}} = \beta^{(-)}_{W|1Z\underline{T}}\beta_{Y|1Z\underline{T}}$ when the $\beta_{W|1Z\underline{T}}$ has the required left inverse $\beta^{(-)}_{W|1Z\underline{T}}$.

6.3.2 Mean and Variance Function Models

We consider generalized linear models, and mean–variance models. Examples of these models are linear, logistic, and Poisson regression. As described more fully in Sections A.7 and A.8, such models depend on a linear combination of the predictors plus possibly a parameter θ that describes the variability in the response. The sections listed above give details for model fitting when there is no measurement error. It might be useful upon first reading to think of this chapter simply as dealing with a class of important models, the details of fitting of which are standard in many computer programs.

These models can be written in general form as

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = f(\beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}\mathbf{X}), \tag{6.15}$$

$$\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \sigma^2 g^2(\beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}\mathbf{X}, \theta), \tag{6.16}$$

and include homoscedastic linear regression, where $f(v) = v$ and $g \equiv 1$, and logistic regression where $\sigma^2 = 1$, f is the logistic distribution function and g^2 is the Bernoulli variance $f(1-f)$. The only notational change with other parts of the book is that the parameters β_0, β_z, and β_x have been replaced by $\beta_{Y|\underline{1}ZX}$, $\beta_{Y|1\underline{Z}X}$ and $\beta_{Y|1Z\underline{X}}$, respectively.

In terms of the composite vectors

$$\begin{aligned} \widetilde{\mathbf{X}} &= (\mathbf{1}, \mathbf{Z}^t, \mathbf{X}^t)^t, & \widetilde{\mathbf{W}} &= (\mathbf{1}, \mathbf{Z}^t, \mathbf{W}^t)^t, \\ \widetilde{\mathbf{T}} &= (\mathbf{1}, \mathbf{Z}^t, \mathbf{T}^t)^t, & \widetilde{\mathbf{U}} &= \widetilde{\mathbf{W}} - \widetilde{\mathbf{X}} \end{aligned}$$

defined in (6.1) and $\beta_{Y|\underline{\widetilde{X}}} = (\beta_{Y|\underline{1}ZX}, \beta^t_{Y|1\underline{Z}X}, \beta^t_{Y|1Z\underline{X}})^t$, the basic model (6.15)–(6.16) becomes

$$\begin{aligned} E(\mathbf{Y}|\widetilde{\mathbf{X}}) &= f(\beta^t_{Y|\underline{\widetilde{X}}}\widetilde{\mathbf{X}}), \\ \text{var}(\mathbf{Y}|\widetilde{\mathbf{X}}) &= \sigma^2 g^2(\beta^t_{Y|\underline{\widetilde{X}}}\widetilde{\mathbf{X}}, \theta). \end{aligned}$$

The goal is to estimate $\beta_{Y|\underline{\widetilde{X}}}$, θ and σ^2.

The assumptions that are necessary for our methods are stated more precisely in Section B.5.1, but we note here that in addition to the conditions stated in Section 6.1, we also assume that the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{T}, \mathbf{W})$ is approximately linear; see (B.27), that is, we assume a *regression calibration model*, see Section 2.2. This restricts the applicability of the methods below somewhat, but is sufficiently general to encompass many potential applications. Combined with the classical additive measurement error model for $\mathbf{W}$, these assumptions result in a hybrid of classical and regression calibration structures, a subject discussed in more detail in Section 6.6.

We now describe two regression-calibration approximations for use with instrumental variables.

6.3.3 First Regression Calibration IV Algorithm

This section describes a simple method when the number of instruments exactly equals the numbers of variables measured with error, and a second method for the case $\dim(\widetilde{\mathbf{T}}) > \dim(\widetilde{\mathbf{X}})$. In Section B.5.1.1, it is shown that, to the level of approximation of regression calibration,

$$E(\mathbf{Y}|\widetilde{\mathbf{T}}) = m_{\mathbf{Y}}\{\beta^t_{Y|\underline{\widetilde{X}}}E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}})\} = f(\beta^t_{Y|\underline{\widetilde{X}}}\beta^t_{\widetilde{X}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}).$$

Since $\mathbf{W} = \mathbf{X} + \mathbf{U}$ and since $\mathbf{U}$ is independent of $(\mathbf{Z}, \mathbf{T})$, the regression of $\widetilde{\mathbf{W}}$ on $\widetilde{\mathbf{T}}$ is the same as the regression of $\widetilde{\mathbf{X}}$ on $\widetilde{\mathbf{T}}$, so that $\beta_{\widetilde{W}|\underline{\widetilde{T}}} = \beta_{\widetilde{X}|\underline{\widetilde{T}}}$, and hence it follows that (approximately)

$$\beta_{Y|\underline{\widetilde{T}}} = \beta_{\widetilde{W}|\underline{\widetilde{T}}}\beta_{Y|\underline{\widetilde{X}}}. \tag{6.17}$$

That is, the coefficient of $\widetilde{\mathbf{T}}$ in the generalized linear regression of $\mathbf{Y}$ on $\widetilde{\mathbf{T}}$ is the product of $\beta^t_{Y|\underline{\widetilde{X}}}$ and $\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}$.

This leads to an extremely simple algorithm:

- Let d_z be the number of $\mathbf{Z}$ variables, d_x be the number of $\mathbf{X}$ variables, and d_t be the number of $\mathbf{T}$ variables. Perform a multivariate regression of $\widetilde{\mathbf{W}}$ on $\widetilde{\mathbf{T}}$ to obtain $\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}$, which is a matrix with $1+\dim(\mathbf{Z})+\dim(\mathbf{X})$ rows and $1+\dim(\mathbf{Z})+\dim(\mathbf{T})$ columns. For $j = 1, ..., 1+\dim(\mathbf{Z})$, the j^{th} row of $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}$ has a 1.0 in the j^{th} column and all other elements equal to zero, reflecting the fact that the regression of $\mathbf{Z}$ on $(\mathbf{Z}, \mathbf{T})$ has no error. For $k = 1, ..., \dim(\mathbf{X})$, row $1+\dim(\mathbf{Z})+k$ of $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}$ contains the regression coefficients when regressing the k^{th} element of $\mathbf{W}$ on $(\mathbf{Z}, \mathbf{T})$, including the intercept in this regression.
- Then perform a generalized linear regression of $\mathbf{Y}$ on the predicted values $\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}$ to obtain an estimator of $\beta_{Y|\underline{\widetilde{X}}}$, which we denote $\widehat{\beta}^{IV1,RC}_{Y|\underline{\widetilde{X}}}$.
- This estimator is easily computed, as it requires only linear regressions

of the components of $\widetilde{\mathbf{W}}$ on $\widetilde{\mathbf{T}}$, and then quasilikelihood and variance function estimation of $\mathbf{Y}$ on the "predictors" $\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}$.

The second means of exploiting the basic regression calibration approximation works directly from the identity (6.17). For a fixed nonsingular matrix M_1, let $\widehat{\beta}^{-(M_1)}_{\widetilde{W}|\underline{\widetilde{T}}} = (\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}} M_1 \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})^{-1}\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}} M_1$. The second estimator is

$$\widehat{\beta}^{IV1,(M_1)}_{Y|\underline{\widetilde{X}}} = \widehat{\beta}^{-(M_1)}_{\widetilde{W}|\underline{\widetilde{T}}}\widehat{\beta}_{Y|\underline{\widetilde{T}}}, \tag{6.18}$$

where $\widehat{\beta}_{Y|\underline{\widetilde{T}}}$ is the estimated regression coefficient when the generalized model is fit to the $(\mathbf{Y}, \widetilde{\mathbf{T}})$ data. Note that (6.18) makes evident the requirement that $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}$ be of full rank. When $\mathbf{T}$ and $\mathbf{W}$ are the same dimension, this estimator does not depend on M_1 and is identical to the first estimator, but not otherwise. When there are more instruments than variables measured with error the choice of M_1 matters. In Section B.5.2.1 we derive an estimate $\widehat{M}_1$ that minimizes the asymptotic variance of $\widehat{\beta}^{IV1,(M_1)}_{Y|\underline{\widetilde{X}}}$. Section B.5.2 gives the relevant asymptotic distribution, although of course the bootstrap can always be used.

6.3.4 Second Regression Calibration IV Algorithm

The second algorithm exploits the fact that both $\mathbf{W}$ and $\mathbf{T}$ are surrogates. The derivation of the estimator is involved (Section B.5.1.2), but the estimator is not difficult to compute.

Let $\dim(\mathbf{Z})$ be the number of components of $\mathbf{Z}$. Define

$$\begin{aligned} \beta_{Y|\underline{\widetilde{T}}\widetilde{W}} &= \beta_{Y|\underline{1ZT}W}, \\ \beta_{Y|\widetilde{T}\underline{\widetilde{W}}} &= (0_{1\times d},\ \beta^t_{Y|1ZT\underline{W}})^t, \end{aligned}$$

where $d = 1 + \dim(\mathbf{Z})$. Then, for a given matrix M_2, the second instrumental variables estimator is

$$\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\widetilde{X}}} = \widehat{\beta}^{-(M_2)}_{\widetilde{W}|\underline{\widetilde{T}}}(\widehat{\beta}_{Y|\underline{\widetilde{T}}\widetilde{W}} + \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}\widehat{\beta}_{Y|\widetilde{T}\underline{\widetilde{W}}}).$$

When $\mathbf{T}$ and $\mathbf{W}$ are the same dimension, $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\widetilde{X}}}$ does not depend on M_2. In Section B.5.2.1, we derive an estimate of M_2 that minimizes the asymptotic variance of $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\widetilde{X}}}$ for the case $\dim(\mathbf{T}) > \dim(\mathbf{W})$.

6.4 Adjusted Score Method

Buzas (1997) developed an approach to IV estimation that, unlike the approximate regression calibration approach in Section 6.3, actually produces fully consistent estimators in certain important generalized linear models with scalar predictor $\mathbf{X}$ subject to measurement error. The method is based upon the hybrid of classical and regression calibration models, a subject discussed in the regression calibration approximation in Section 6.3.2, and also discussed in more detail in Section 6.6.

In the hybrid approach, along with the measurement error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$, we have a regression calibration model for $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{T})$, where we write $E(\mathbf{X}|\mathbf{Z}, \mathbf{T}) = m_{\mathbf{X}}(\mathbf{Z}, \mathbf{T}, \gamma)$. Generally, the parameter γ will have to be estimated, but this is the beauty inherent in the assumptions of the hybrid approach, namely, that as long as the measurement error $\mathbf{U}$ is independent of $(\mathbf{Z}, \mathbf{T})$, then (possibly nonlinear) regression of $\mathbf{W}$ on $(\mathbf{Z}, \mathbf{T})$ will provide an estimate of γ. This is generally done by solving the least squares equation of the form

$$\sum_{i=1}^{n} \psi_{m_{\mathbf{X}}}(\mathbf{W}_i, \mathbf{Z}_i, \mathbf{T}, \widehat{\gamma}) = \mathbf{0}.$$

The starting point for Buzas' method is that along with the regression parameters $\beta_{Y|\underline{1ZX}}$, there may be additional parameters τ. He then supposes that there is a score function that produces consistent estimators in the absence of measurement error. In his framework, the mean function is denoted by $m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \beta_{Y|\underline{1ZX}}, \tau)$. The form of the mean functions of most interest here is where

$$m_{\mathbf{Y}}(Z, \mathbf{x}; \beta_{Y|\underline{1ZX}}) = \frac{a_1 + a_2\mathrm{exp}(a_5\mathbf{x})}{a_3 + a_4\mathrm{exp}(a_5\mathbf{x})},$$

where $a_1, \ldots, a_5$ are scalar functions of Z and $\beta_{Y|\underline{1ZX}}$, but not $\mathbf{x}$. Noteworthy in this class are

- The logistic mean model $m_{\mathbf{Y}}(Z, \mathbf{x}; \beta_{Y|\underline{1ZX}}) = 1/\{1 + \exp(-\beta_{Y|\underline{1}ZX} - \beta_{Y|1\underline{Z}X}Z - \beta_{Y|1Z\underline{X}}\mathbf{x})\}$, obtained when $a_1 = a_3 = 1$, $a_2 = 0$, $a_4 = \exp(-\beta_{Y|\underline{1}ZX} - \beta_{Y|1\underline{Z}X}Z)$ and $a_5 = -\beta_{Y|1Z\underline{X}}$.
- The Poisson loglinear mean model $m_{\mathbf{Y}}(Z, \mathbf{x}; \beta_{Y|\underline{1ZX}}) = \exp(\beta_{Y|\underline{1}ZX} + \beta_{Y|1\underline{Z}X}Z + \beta_{Y|1Z\underline{X}}\mathbf{x})$, obtained when $a_1 = a_4 = 0$, $a_2 = \exp(\beta_{Y|\underline{1}ZX} + \beta_{Y|1\underline{Z}X}Z)$, $a_3 = 1$ and $a_5 = \beta_{Y|1Z\underline{X}}$.

In these problems, the score function ψ has the form

$$\begin{aligned} &\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \beta_{Y|\underline{1ZX}}, \tau) = \\ &\qquad \{\mathbf{Y} - m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \beta_{Y|\underline{1ZX}}, \tau)\}g(\mathbf{Z}, \mathbf{X}, \beta_{Y|\underline{1ZX}}, \tau); \end{aligned}$$

and is such that the estimating equations

$$\textstyle\sum_{i=1}^{n} \psi(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \beta_{Y|\underline{1ZX}}, \tau) = \mathbf{0},$$

produce consistent estimators of $\beta_{Y|\underline{1ZX}}$ and τ in the absence of measurement error. The class of estimators covered by this setup includes nonlinear least squares (Gallant, 1987), quasilikelihood and variance function models (Carroll and Ruppert, 1988), and generalized linear models (McCullagh and Nelder, 1989), among others.

Buzas (1997) showed that when the measurement error $\mathbf{U}$ is symmetrically distributed about zero given $(\mathbf{Z}, \mathbf{X}, \mathbf{T})$, then a score function leading to consistent estimation is the following. Define

$$\phi(\mathbf{Z}, \mathbf{W}, \mathbf{T}, \beta_{Y|\underline{1ZX}}) = \left| \frac{m_{\mathbf{Y},x}(\mathbf{Z}, E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}); \beta_{Y|\underline{1ZX}})}{m_{\mathbf{Y},x}(\mathbf{Z}, \mathbf{W}; \beta_{Y|\underline{1ZX}})} \right|^{1/2}$$

with $m_{\mathbf{Y},x}(Z, \mathbf{x}; \beta_{Y|\underline{1ZX}}) = (\partial/\partial\mathbf{x}) m_{\mathbf{Y}}(Z, \mathbf{x}; \beta_{Y|\underline{1ZX}})$. Then the modified score leading to consistent estimation is

$$\begin{aligned} &\psi_{\text{IV}}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{T}, \beta_{Y|\underline{1ZX}}, \tau) = \\ &\qquad \{\mathbf{Y} - m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{W}; \beta_{Y|\underline{1ZX}})\}\phi(\mathbf{Z}, \mathbf{W}, \mathbf{T}, \beta_{Y|\underline{1ZX}}) \\ &\qquad\quad \times g(\mathbf{Z}, E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}), \beta_{Y|\underline{1ZX}}, \tau). \end{aligned}$$

Inference can be obtained either by the bootstrap or by using the method of stacking estimating equations in Section A.6.6, that is, stack $\psi_{m_{\mathbf{X}}}(\cdot)$ below $\psi_{\text{IV}}(\cdot)$.

	Age	Smoke	Chol	LSBP
Naive	.056	.573	.0078	1.524
Std. Err.	.010	.243	.0019	.364
IV1	.054	.577	.0076	2.002
Std. Err.	.011	.244	.0020	.517
IV2	.054	.579	.0077	1.935
Std. Err.	.011	.244	.0020	.513
Adj Score	.055	.597	.0082	1.930
Std. Err.	.011	.250	.0020	.494

Table 6.1 *Estimates and standard errors from the Framingham data instrumental variable logistic regression analysis. This analysis used the one-dimensional instrumental variable LSBP* $= log\{(SBP_{2,1} + SBP_{2,2})/2 - 50\}$. *"Smoke" is smoking status and "Chol" is cholesterol level. Standard errors calculated using the sandwich method.*

6.5 Examples

6.5.1 Framingham Data

We now illustrate the methods presented in this chapter using the Framingham heart study data from Section 5.4.1, wherein two systolic blood pressure measurements from each of two exams were used. It was assumed that the two transformed variates

$$\mathbf{W}_1 = \log\{(\text{SBP}_{3,1} + \text{SBP}_{3,2})/2 - 50\}$$

and

$$\mathbf{W}_2 = \log\{(\text{SBP}_{2,1} + \text{SBP}_{2,2})/2 - 50\},$$

where $\text{SBP}_{i,j}$ is the j^{th} measurement of SBP from the i^{th} exam, $j = 1, 2$, $i = 2, 3$, were replicate measurements of the long-term average transformed SBP.

Table 6.1 displays estimates of the same logistic regression model fit in Section 5.4.2.1 with the difference that $\mathbf{W}_2$ was employed as an instrumental variable, not as a replicate measurement, that is, in the notation of this section, $\mathbf{W} = \mathbf{W}_1$ and $\mathbf{T} = \mathbf{W}_2$.

Because $\mathbf{T}$ has the same dimension as $\mathbf{W}$, the estimate $\beta^{IV1,(M_1)}_{Y|\underline{\widetilde{X}}}$ does not depend on M_1 and is equivalent to $\beta^{IV1,RC}_{Y|\underline{\widetilde{X}}}$. This common estimate is listed under IV1 in Table 6.1. Also $\beta^{IV2,(M_2)}_{Y|\underline{\widetilde{X}}}$ does not depend on M_2 and is listed under IV2 in the table. For the Buzas estimate, a linear regression model for $E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T})$ was used, $M(\mathbf{Z}, \mathbf{T}, \gamma) = (1, \mathbf{Z}^t, \mathbf{T}^t)\gamma$ with $\widehat{\gamma}$ obtained by least squares, so that

$$\psi_{m_{\mathbf{X}}}(\mathbf{W}, \mathbf{Z}, \mathbf{T}, \gamma) = \{\mathbf{W} - (1, \mathbf{Z}^t, \mathbf{T}^t)\gamma\}(1, \mathbf{Z}^t, \mathbf{T}^t)^t.$$

Table 6.2 displays estimates of the same logistic regression model with the difference that the instrumental variable $\mathbf{T}$ was taken to be the two-dimensional variate

$$\mathbf{T} = \{\log(\text{SBP}_{2,1}),\ \log(\text{SBP}_{2,2})\}. \tag{6.19}$$

Note the similarity among the estimates in Tables 6.1 and 6.2.

The primary purpose of this second analysis is to illustrate the differences between the estimators when $\dim(\mathbf{T}) > \dim(\mathbf{X})$, and to emphasize that $\mathbf{T}$ need only be correlated with $\mathbf{X}$, and not a second measurement, for the methods to be applicable.

However, we also use this model to illustrate further the key assumptions (6.9) and to discuss the issues involved in verifying them.

Rewrite the IV model (6.19) as $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2)$. Given our previous discussions of the Framingham data, it follows that reasonable models for $\mathbf{W}$ and $\mathbf{T}$ are

$$\mathbf{W} = \mathbf{X} + \mathbf{U},$$

	Age	Smoke	Chol	LSBP
Naive	.056	.573	.0078	1.524
Std. Err.	.010	.243	.0019	.364
IV1,RC	.054	.577	.0076	1.877
Std. Err.	.011	.244	.0020	.481
IV1,(M_1)	.054	.577	.0076	1.884
Std Err.	.011	.244	.0020	.483
IV2,(M_2)	.054	.579	.0077	1.860
Std. Err.	.011	.244	.0020	.484
Adj Score	.055	.592	.0082	1.887
Std. Err.	.011	.250	.0020	.494

Table 6.2 *Estimates and standard errors from the Framingham data instrumental variable logistic regression analysis. This analysis used the two-dimensional instrumental variable* $\{log(SBP_{2,1}),\ log(SBP_{2,2})\}$*. "Smoke" is smoking status and "Chol" is cholesterol level. Standard errors calculated using the sandwich method.*

$$\begin{aligned}\mathbf{T}_1 &= a_1 + b_1\mathbf{X} + \mathbf{U}_1,\\ \mathbf{T}_2 &= a_2 + b_2\mathbf{X} + \mathbf{U}_2,\end{aligned} \tag{6.20}$$

where $\mathbf{U}$, $\mathbf{U}_1$, and $\mathbf{U}_2$ are mutually independent and independent of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$. These independence assumptions are comparable to those used previously to justify the various analyses of the Framingham data. With ϵ replaced by $\mathbf{Y} - E(\mathbf{Y}|\mathbf{Z}, \mathbf{X})$ (because the model for $\mathbf{Y}$ is logistic not linear), the aforementioned independence assumptions ensure the validity of the first two components of (6.9). Now for the model (6.20) the crossproduct matrix $\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}$ is the 6×5 matrix,

$$\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}} = E\begin{bmatrix} 1 & \mathbf{Z}^t & \mathbf{X} \\ \mathbf{Z} & \mathbf{Z}\mathbf{Z}^t & \mathbf{Z}\mathbf{X} \\ a_1 + b_1\mathbf{X} & (a_1 + b_1\mathbf{X})\mathbf{Z}^t & (a_1 + b_1\mathbf{X})\mathbf{X} \\ a_2 + b_2\mathbf{X} & (a_2 + b_2\mathbf{X})\mathbf{Z}^t & (a_2 + b_2\mathbf{X})\mathbf{X} \end{bmatrix}$$

(recall that $\dim(\mathbf{Z}) = 3$ and $\dim(\widetilde{\mathbf{X}}) = 5$ for the Framingham data). It should be obvious that $\mathrm{rank}(\Omega_{\widetilde{\mathbf{T}}\widetilde{\mathbf{X}}}) = 5$, if and only if at least one of b_1 and b_2 is nonzero. (Okay, we're kidding about the "obvious" part — you'll just have to take our word on this one. However, for those doubting Thomases among the readers, we suggest that you first consider the case where the components of $\mathbf{X}$ and $\mathbf{Z}$ are iid and standardized to mean zero and variance one, and then extend to the general case.)

This example illustrates the fact that not all instruments need to be correlated with $\mathbf{X}$, and that if multiple instruments are available (all satisfying (6.9) of course) there is no harm using them. In fact, as Fuller (1987, p. 154) notes, adding more instrumental variables can improve the quality of an IV estimator.

6.5.2 Simulated Data

The instrumental variable results in Table 6.2 are very close to what was obtained for regression calibration and SIMEX; see Table 5.1 in Section 5.4.1. The reader can then be forgiven for concluding that instrumental variable analyses are equivalent to corrections for attenuation. The Framingham data, though, are a special case, because the instrument is for all practical purposes an unbiased estimate of $\mathbf{X}$ with the same measurement error as that of $\mathbf{W}$. A simulation will help dispel the notion that instrumental variables are always equivalent to corrections for attenuation. First the simulation. Consider linear regression of $\mathbf{Y}$ on $\mathbf{X}$, with slope $\beta_x = 1$ and error about the line $\sigma_\epsilon^2 = 1$. Let the sample size be $n = 400$. Let $\mathbf{X} = \mathrm{Normal}(0, 1)$, and let the measurement error variance $\sigma_u^2 = 1$, so that the reliability ratio is $\lambda = 0.50$. We consider two cases. In the first, $\mathbf{W}$ is replicated in order to estimate σ_u^2, but only the first replicate is used. In the second, we have an instrument $\mathbf{T} = 0.2\mathbf{X} + \nu$, where $\nu = \mathrm{Normal}(0, 1)$. Here, the instrument does not have a high correlation with $\mathbf{W}$, so the division inherent in (6.6) is bound to cause a problem. Then, in 500 simulations, the naive estimator is biased as expected, and the correction for attenuation and instrumental variable estimators are nearly unbiased. However, the correction for attenuation estimator had much less variability than the instrumental variables estimator, either in its raw form or with the correction for small samples; see Figure 6.1. Even the corrected form has a variance more than four times greater than the correction for attenuation.

6.6 Other Methodologies

6.6.1 Hybrid Classical and Regression Calibration

We have seen two examples in which the classical additive measurement error model relating $(\mathbf{W}, \mathbf{X})$ is combined with a parametric regression

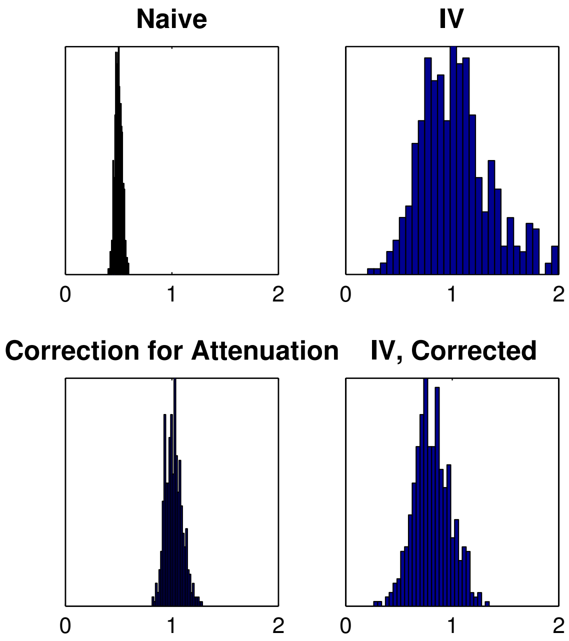


Figure 6.1 *Comparison of methods in simulated data when the instrument is weak. Top left: naive estimate. Bottom left: correction for attenuation. Top right: instrumental variables estimator. Bottom right: instrumental variables with small sample correction.*

calibration model relating $\mathbf{X}$ to $(\mathbf{Z},\mathbf{T})$; see Section 6.3.2 and 6.4. Several papers use the same basic modeling strategy.

Hausman, Newey, Ichimura, et al. (1991) consider the polynomial regression model in which the unobserved true $\mathbf{X}$ is measured with classical additive error, while it is related to the instrument though a regression calibration model (Section 2.2). Specifically, their model is that $\mathbf{W}=\mathbf{X}+\mathbf{U}$, and that

$$\begin{aligned}\mathbf{Y} &= \beta_0+\beta_z^t\mathbf{Z}+\textstyle\sum_{j=1}^{p}\beta_{x,j}\mathbf{X}^j+\epsilon;\\ \mathbf{X} &= \alpha_0+\alpha_z^t\mathbf{Z}+\alpha_t^t\mathbf{T}+\nu. \end{aligned}\qquad(6.21)$$

Of course, (6.21) is a regression calibration model. Hausman, Newey, Ichimura, et al. (1991) assume, in effect, that ϵ and $\mathbf{U}$ are each independent of $(\mathbf{Z},\mathbf{T})$ but that they need not be independent of one another. They also assume that ν is independent of $(\mathbf{Z},\mathbf{T})$, but they allow ν and ϵ to be correlated. Effectively, their method is to compute higher-order moments of the observed data to show that all parameters can be identified, and then to estimate these moments.

Schennach (2006) also considers a hybrid version of the classical and regression calibration approaches when $\mathbf{X}$ is scalar and there are no covariates $\mathbf{Z}$. Her general model also has $\mathbf{W}=\mathbf{X}+\mathbf{U}$ and takes the form

$$\begin{aligned}\mathbf{Y} &= m_{\mathbf{Y}}(\mathbf{X},\mathcal{B})+\epsilon;\\ \mathbf{X} &= m_{\mathbf{X}}(\mathbf{T},\gamma)+\nu. \end{aligned}\qquad(6.22)$$

She assumes, in effect, that ϵ is independent of $(\nu,\mathbf{T})$, that $\mathbf{U}$ is independent of $(\epsilon,\nu,\mathbf{T})$, and that ν is independent of $\mathbf{T}$. She notes that it is possible to extend the model to include $\mathbf{Z}$. The functional forms of $m_{\mathbf{Y}}(\cdot)$ and $m_{\mathbf{X}}(\cdot)$ are assumed known, with the unknowns being the parameters $(\mathcal{B},\gamma)$. Her method is more complex than in the polynomial case.

6.6.2 Error Model Approaches

In the instrumental variable context, hybrid models such as (6.21) and (6.22) are appealing because their means as a function of $(\mathbf{Z},\mathbf{T})$ can be estimated simply by regressing $\mathbf{W}$ on $(\mathbf{Z},\mathbf{T})$. This has led us to regression calibration as an approximate device (Section 6.3), the adjusted score method for certain special problems (Section 6.4), and other modeling approaches (Section 6.6.1). All these methods are intrinsically different from the approaches to measurement error modeling when the error variance is known or can be estimated in other ways, for example, replication.

Under stronger technical, although not practical, conditions that were previously discussed, however, it is possible to achieve identifiability of estimation and also to employ methods from previous and succeeding chapters, for example, SIMEX; see Carroll, Ruppert, Crainiceanu, et al. (2004). First consider the case when there is no $\mathbf{Z}$. For scalar $\mathbf{X}$, Carroll, Ruppert, Crainiceanu, et al. (2004) start with a model that relates the response to covariates and random error as

$$\mathbf{Y}=\mathcal{G}(\mathbf{X},\mathcal{B},\epsilon).$$

This is a completely general model including generalized linear models, nonlinear models, etc. These authors assume the usual classical additive error model $\mathbf{W}=\mathbf{X}+\mathbf{U}$, and they relate the instrument via a generalization of the biased classical error model in (2.1) discussed in Section 2.2.1. Specifically, to begin with they assume that

$$\mathbf{T}=\alpha_0+\alpha_1\mathbf{X}+\nu,\qquad(6.23)$$

and that $(\epsilon, \mathbf{U}, \nu, \mathbf{X})$ are all mutually independent. Then, it follows that

$$\text{var}(\mathbf{U}) = \sigma_u^2 = \text{var}(\mathbf{W}) - \frac{\text{cov}(\mathbf{W}, \mathbf{T})\text{cov}(\mathbf{Y}, \mathbf{W})}{\text{cov}(\mathbf{Y}, \mathbf{T})}.$$

One can thus estimate σ_u^2 by replacing the variance and covariances by their sample versions and then using one's favorite estimation method tuned to the case where an estimate of σ_u^2 is available. If the model $\mathcal{G}(\mathbf{X}, \mathcal{B}, \epsilon)$ is simple linear regression, this algorithm produces the usual instrumental variable estimator. It is worth pointing out that this method-of-moments estimate need not be positive, or smaller than var$(\mathbf{W})$, and Carroll, Ruppert, Crainiceanu, et al. (2004) suggest placing bounds on the attenuation.

More generally, we can include $\mathbf{Z}$ by writing the model

$$\begin{aligned} \mathbf{Y} &= \mathcal{G}(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \epsilon); \\ \mathbf{T} &= f_0(\mathbf{Z}) + f_1(\mathbf{Z})\mathbf{X} + \nu, \end{aligned}$$

the latter being a varying coefficient model, with function f_0 and f_1. Carroll, Ruppert, Crainiceanu, et al. (2004) show how to estimate σ_u^2 in this general model, even when f_0 and f_1 are modeled nonparametrically: Once σ_u^2 is estimated, estimating the model parameter $\mathcal{B}$ can be done by any of the methods discussed elsewhere in this book.

For example, in the Framingham data, it makes sense to use (6.23), since the instrument is a second, possibly biased, measurement of true blood pressure. When we applied the method to estimate σ_u^2 and then used regression calibration, we obtained estimates and standard errors that were essentially the same as in Table 6.1.

Bibliographic Notes

The literature on instrumental variable estimation in linear models is extensive. For readers wanting more than the introduction in Section 6.2, a good place to start is Fuller (1987). Interest in instrumental variables in nonlinear measurement error models is more recent, and a number of methods have been proposed that are generally either more specialized, or more involved than the methods described in this chapter. See, for example, Amemiya (1985, 1990a, 1990b) for general nonlinear measurement error models; Stefanski and Buzas (1995), Buzas and Stefanski (1996b), and Thoresen and Laake (1999) for binary measurement error models; Buzas and Stefanski (1996c) for certain generalized linear models; and Hausman, Newey, Ichimura, et al. (1991) for polynomial models. Carroll, Ruppert, Crainiceanu, et al. (2004) provided identifiability results for very general measurement error models with instrumental variables, and develop methods for estimating nonlinear measurement error

models nonparametrically by combining an estimate of the measurement error variance derived from the instrument with methods of nonparametric estimation for measurement error models with known error variance. Other papers of general interest are Carroll and Stefanski (1994) and Greenland (2000).

CHAPTER 7

SCORE FUNCTION METHODS

7.1 Overview

Regression calibration (Chapter 4) and SIMEX (Chapter 5) are widely applicable, general methods for eliminating or reducing measurement error bias. These methods result in estimators that are consistent in important special cases, such as linear regression and loglinear mean models, but that are only approximately consistent in general.

In this chapter, we describe methods that are almost as widely applicable, but that result in fully consistent estimators more generally. Consistency is achieved by virtue of the fact that the estimators are M-estimators whose score functions are unbiased in the presence of measurement error. This property is also true of structural model maximum likelihood and quasilikelihood estimates, as discussed in Chapter 8. The lack of assumptions about the unknown $\mathbf{X}_i$ distinguishes the methods in this chapter from those in Chapter 8. The methods are functional methods, as defined in Section 2.1.

However, we do not deal with functional modeling as it is used in the linear models measurement error setting, for it is not a viable option for nonlinear measurement error models. Suppose for the sake of discussion that the measurement error covariance matrix Σ_{uu} is known. In the old classical functional model, the unobservable $\mathbf{X}_i$ are fixed constants and are regarded as parameters. With additive, normally distributed measurement error, functional maximum likelihood maximizes the joint density of the observed data with respect to all of the unknown parameters, including the $\mathbf{X}_i$. While this works for linear regression (Gleser, 1981), it fails for more complex models such as logistic regression (Stefanski and Carroll, 1985). Indeed, the functional estimator in most nonlinear models is both extremely difficult to compute and not even consistent or valid. The methods in this chapter make no assumptions about the $\mathbf{X}_i$, are often easier computationally, and lead to valid estimation and inference.

We focus on the case of additive, normally distributed measurement error, so that $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with $\mathbf{U}$ distributed as a normal random vector with mean zero and covariance matrix Σ_{uu}, and two broad classes of score function methods that have frequent application.

- The *conditional-score* method of Stefanski and Carroll (1987) exploits special structures in important models such as linear, logistic, Poisson loglinear, and gamma-inverse regression, using a traditional statistical device, conditioning on sufficient statistics, to obtain estimators.
- The *corrected-score* method effectively estimates the estimator that one would use if there were no measurement error.

We start with linear and logistic regression, using these important special cases both as motivation for and explanation of the general methods. Next, the conditional- and corrected-score methods are illustrated with a logistic regression example in Section 7.2.3. Then, in successive sections, we describe the conditional-score and corrected-score methods in detail, covering the basic theory and giving examples for each method.

We warn the reader that the mathematical notation of conditional and corrected scores is more complex than that of regression calibration and SIMEX. However, the formulae are simple to program and implement, with the possible exception of Monte Carlo corrected scores (MCCS), which require complex variable computation that may not be available in all programming languages. More important, the methods in this chapter result in fully consistent estimators under the conditions stated on the true-data model and the error model, not just approximately consistent, as is often the case for regression calibration and SIMEX.

7.2 Linear and Logistic Regression

This section introduces the ideas of corrected and conditional scores in two important problems, namely, linear regression and logistic regression. In linear regression, of course, we already know how to construct valid estimation and inferential methods, as described in Section 3.4, so nothing really new is being done here: The calculations are simply easier to follow for this case, and those wishing to understand the ideas, especially the new ideas for corrected scores, will find the linear regression calculations give useful insight. For logistic regression, these methods produce consistent, and not just approximately consistent methods.

7.2.1 Linear Regression Corrected and Conditional Scores

Consider the multiple linear regression model with mean $E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}) = \beta_0 + \beta_z^t \mathbf{Z} + \beta_x^t \mathbf{X}$, variance $\mathrm{var}(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}) = \sigma^2$, and the classical additive, nondifferential error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with $\mathbf{U} = \mathrm{Normal}(\mathbf{0}, \Sigma_{uu})$ where Σ_{uu} is known. Write the unknown regression parameter as $\Theta_1 = (\beta_0, \beta_z^t, \beta_x^t)^t$ and $\Theta = (\Theta_1^t, \Theta_2)^t$ with $\Theta_2 = \sigma^2$.

The ordinary least squares score function for multiple linear regression

in the absence of measurement error is

$$\Psi_{\mathrm{LS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta) = \begin{bmatrix} \{\mathbf{Y}_i - (1, \mathbf{Z}_i^t, \mathbf{X}_i^t)\Theta_1\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{X}_i \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right) \sigma^2 - \{\mathbf{Y}_i - (1, \mathbf{Z}_i^t, \mathbf{X}_i^t)\Theta_1\}^2 \end{bmatrix}. \qquad (7.1)$$

The upper equation is the least squares score function (the so-called normal equations) for Θ_1, the regression parameters. The factor $(n-p)/n$ with $p = \dim(\Theta_1)$ in the lower equation implements the usual degrees-of-freedom correction for the estimator of σ^2.

7.2.1.1 Linear Regression Conditional Score

We now describe an approach to consistent estimation that requires no assumptions about the $\mathbf{X}$-variables. The *derivation* of the method, but not its validity, assumes normality of the true-regression equation error, ϵ_i, as well as the measurement errors $\mathbf{U}_i$. Define

$$\Delta_i = \mathbf{W}_i + \mathbf{Y}_i \Sigma_{uu} \beta_x / \sigma^2. \qquad (7.2)$$

Given $\mathbf{Z}_i$ and $\mathbf{X}_i$, the random variables $\mathbf{Y}_i$ and Δ_i are linear functions of jointly normal random vectors and thus are jointly normal, conditionally on $(\mathbf{Z}_i, \mathbf{X}_i)$. Consequently, the conditional distribution of $\mathbf{Y}_i$ given $(\mathbf{Z}_i, \mathbf{X}_i, \Delta_i)$ is also normal, and standard multivariate-normal calculations show that

$$\begin{aligned} E(\mathbf{Y}_i \mid \mathbf{Z}_i, \mathbf{X}_i, \Delta_i) = E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) &= \frac{\beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \Delta_i}{1 + \beta_x^t \Sigma_{uu} \beta_x / \sigma^2}, \\ \mathrm{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \mathbf{X}_i, \Delta_i) = \mathrm{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) &= \frac{\sigma^2}{1 + \beta_x^t \Sigma_{uu} \beta_x / \sigma^2}. \end{aligned} \qquad (7.3)$$

These conditional moments are noteworthy for their lack of dependence on $\mathbf{X}_i$. We will show in Section 7.3 that this is by design, that is, the manner in which Δ_i is defined ensures that these moments depend only on the observed data and not on $\mathbf{X}_i$.

It follows from (7.3) that the *conditional score*,

$$\Psi_{\mathrm{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \begin{bmatrix} \{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix} \\ \sigma^2 - \dfrac{\{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\}^2}{\mathrm{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)/\sigma^2} \end{bmatrix},$$

has the property that

$$E\left\{\Psi_{\mathrm{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) \mid \mathbf{Z}_i, \Delta_i\right\} = \mathbf{0},$$

so its unconditional mean also vanishes. Thus, Ψ_{Cond} can be used to form unbiased estimating equations,

$$\textstyle\sum_{i=1}^{n}\Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0}. \tag{7.4}$$

However, in practice we estimate the parameters by solving the small-sample modified estimating equations

$$\textstyle\sum_{i=1}^{n}\left[\begin{array}{c} \{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\}\begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\sigma^2 - \dfrac{\{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\}^2}{\text{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)/\sigma^2} \end{array}\right] = \mathbf{0}. \tag{7.5}$$

The factor $(n-p)/n$ in the equation for σ^2 implements the degrees-of-freedom correction for the estimator of σ^2. The asymptotic theory of M-estimators in Section A.6 can be applied to approximate the distribution of $\widehat{\Theta}$.

7.2.1.2 Linear Regression Corrected Score

We now derive the *corrected score* for linear regression using the general method of construction described in Section 7.4. The corrected score for linear regression is readily obtained using other approaches, hence the general method of construction is overkill for this case. However, the development is instructive and readily generalizes to problems of greater interest.

The general method of constructing corrected scores uses complex variables and complex-valued functions. Although familiarity with complex variables is helpful to understand how the method works, it is not essential to using, or even implementing the methods, provided one uses a programming language with complex number capabilities (GAUSS and MATLAB have such capabilities, for example). We use the bold Greek letter iota ($\boldsymbol{\iota}$) to denote the unit imaginary number, $\boldsymbol{\iota} = \sqrt{-1}$, to distinguish it from the observation index i. Only a few facts about complex numbers are used in this section: a) $\boldsymbol{\iota}^2 = -1$; b) the real part of complex number is $\text{Re}(z_1 + \boldsymbol{\iota} z_2) = z_1$; and c) if $f(z_1 + \boldsymbol{\iota} z_2)$ is a function of a complex variable, then $f(z_1 + \boldsymbol{\iota} z_2) = g(z_1, z_2) + \boldsymbol{\iota} h(z_1, z_2)$, where g and h are both real-valued functions and g is the real part of f, that is, $g(z_1, z_2) = \text{Re}\{f(z_1 + \boldsymbol{\iota} z_2)\}$.

The general method of construction has a similar feel to SIMEX, in that we use the computer to generate random variables to help in defining an estimator. In the case of corrected scores, these random variables are defined as follows.

Now, for $b = 1, ..., B$, generate random variables $\mathbf{U}_{b,i}$ that are independent normal random vectors with mean zero and covariance matrix

Σ_{uu}. Consider the complex-valued random variate,

$$\widetilde{\mathbf{W}}_{b,i} = \mathbf{W}_i + \boldsymbol{\iota}\mathbf{U}_{b,i}. \tag{7.6}$$

The Monte Carlo corrected score (MCCS) is obtained in three steps:

1. Replace $\mathbf{X}_i$ with $\widetilde{\mathbf{W}}_{b,i}$ in a score function that is unbiased in the absence of measurement error — for linear least squares regression this is (7.1).
2. Take the real part, $\text{Re}(\cdot)$, of the resulting expression to eliminate the imaginary part.
3. Average over multiple sets of pseudorandom vectors, $b = 1, \ldots, B$.

For linear regression, these steps result in

$$\begin{aligned}\widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) &= B^{-1}\textstyle\sum_{b=1}^{B}\text{Re}\left\{\Psi_{\text{LS}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_{b,i}, \Theta)\right\} \\ &= \left[\begin{array}{c} \{\mathbf{Y}_i - (1, \mathbf{Z}_i^t, \mathbf{W}_i^t)\Theta_1\}\begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{W}_i \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \widehat{M}_{u,i}\beta_x \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\sigma^2 - \{\mathbf{Y}_i - (1, \mathbf{Z}_i^t, \mathbf{W}_i^t)\Theta_1\}^2 + \beta_x^t\widehat{M}_{u,i}\beta_x \end{array}\right],\end{aligned}$$

where $\widehat{M}_{u,i} = B^{-1}\sum_{b=1}^{B}\mathbf{U}_{b,i}\mathbf{U}_{b,i}^T$. Because $E(\widehat{M}_{u,i}) = \Sigma_{uu}$, it follows that for all i and B,

$$E\left\{\widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) \mid \mathbf{Z}_i, \mathbf{X}_i\right\} = \Psi_{\text{LS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta), \tag{7.7}$$

and consequently that, if we ignore the degrees-of-freedom correction factor $(n-p)/n$ in (7.1),

$$E\left\{\widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta)\right\} = \mathbf{0}. \tag{7.8}$$

Equation (7.7) provides insight into how corrected scores work. The corrected score, $\widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta)$, is an unbiased estimator of the score that would have been be used, $\Psi_{\text{LS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta)$, if measurement error were not present.

It follows from general M-estimator theory (Section A.6) that under regularity conditions the estimating equations,

$$\textstyle\sum_{i=1}^{n}\widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0}, \tag{7.9}$$

admit a consistent and asymptotically normal sequence of solutions.

We can gain further insight into the workings of corrected scores by solving (7.9) for the case of linear regression, resulting in

$$\begin{aligned}\widehat{\Theta}_1 &= \left(\widehat{M}_{1zw,1zw} - \widetilde{\Omega}\right)^{-1}\widehat{M}_{y,1zw}, \\ \widehat{\sigma}^2 &= (n-p)^{-1}\textstyle\sum_{i=1}^{n}\left\{\left(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i\right)^2 - \widehat{\beta}_x^t\widehat{\Sigma}_{uu}\widehat{\beta}_x\right\},\end{aligned}$$

where

$$\widehat{M}_{1zw,1zw} = n^{-1}\sum_{i=1}^{n} \begin{pmatrix} 1 & \mathbf{Z}_i^t & \mathbf{W}_i^t \\ \mathbf{Z}_i & \mathbf{Z}_i\mathbf{Z}_i^t & \mathbf{Z}_i\mathbf{W}_i^t \\ \mathbf{W}_i & \mathbf{W}_i\mathbf{Z}_i^t & \mathbf{W}_i\mathbf{W}_i^t \end{pmatrix},$$

$$\widetilde{\Omega} = \begin{pmatrix} 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\Sigma}_{uu} \end{pmatrix}, \quad \widehat{\Sigma}_{uu} = \left(n^{-1}\sum_{i=1}^{n}\widehat{M}_{u,i}\right),$$

$$\widehat{M}_{y,1zw} = n^{-1}\sum_{i=1}^{n}\mathbf{Y}_i \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{W}_i \end{pmatrix}, \quad \widehat{\mathbf{Y}}_i = (1, \mathbf{Z}_i^t, \mathbf{W}_i^t)\widehat{\Theta}_1.$$

Because we are working under the assumption that Σ_{uu} is known, it probably seems odd, and it is certainly inefficient, that these estimators depend on the random matrix $\widehat{\Sigma}_{uu}$. The sensible strategy is to replace $\widehat{\Sigma}_{uu}$ with Σ_{uu}. Doing so yields, apart from degrees-of-freedom corrections on the relevant covariance matrices, the usual linear models, method-of-moments correction for measurement error bias (Fuller, 1987). Practically, the substitution of Σ_{uu} for $\widehat{\Sigma}_{uu}$ can also be accomplished by taking B large, because $\widehat{\Sigma}_{uu}$ converges to Σ_{uu} as $B \to \infty$. Usually, in practice B does not need to be very large to obtain good results. This is because the randomness introduced in the construction of the Monte Carlo corrected scores is subject to double averaging over n and B. This is apparent in the linear regression corrected-score estimator, as it depends on the $\mathbf{U}_{b,i}$ only via

$$\widehat{\Sigma}_{uu} = (nB)^{-1}\sum_{i=1}^{n}\sum_{b=1}^{B} U_{b,i}U_{b,i}^t,$$

and the variances of the components of this random matrix are on the order of $(nB)^{-1}$.

Herein lies the advantage of the general theory in Section 7.4. For many measurement error models, substituting the complex variate $\widetilde{\mathbf{W}}_{b,i}$ defined in (7.6) for $\mathbf{X}_i$ into a score function that is unbiased in the absence of measurement error, taking the real part, and averaging over $b = 1, \ldots, B$ results in an unbiased score that is a function of the observed data. In the linear model we can shortcut the pseudorandom number generation and averaging, because all of the expressions involved depend only on first- and second-order sample moments. The corrected score in this case is

$$\widetilde{\Psi}_{\text{MCCS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) =$$

$$\begin{bmatrix} \{\mathbf{Y}_i - (1, \mathbf{Z}_i^t, \mathbf{W}_i^t)\Theta_1\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{W}_i \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \Sigma_{uu}\beta_x \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\sigma^2 - \{\mathbf{Y}_i - (1, \mathbf{Z}_i^t, \mathbf{W}_i^t)\Theta_1\}^2 + \beta_x^t\Sigma_{uu}\beta_x \end{bmatrix}.$$

Note $E\{\widetilde{\Psi}_{\text{MCCS,B}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) \mid \mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i\} = \widetilde{\Psi}_{\text{CS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta)$. We call $\widetilde{\Psi}_{\text{CS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta)$ a corrected score to distinguish it from the Monte Carlo corrected score $\widetilde{\Psi}_{\text{MCCS,B}}$. Corrected scores for certain other simple common statistical models can be found without using Monte Carlo averaging, and some of these are given in Section 7.4.3. However, whenever a corrected score exists, the Monte Carlo corrected score estimates it precisely for B large and avoids the mathematical problem of finding it, although, of course, at the cost of complex variable computation.

Note that in the above discussion, no assumptions were made about the true-regression equation error, $\epsilon_i = \mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \mathbf{X}_i)$, either in practice or in the derivation. A final important point to note about the corrected-score method is that no assumptions are made about the unobserved $\mathbf{X}$ variables other than those assumptions that would be needed to ensure consistent estimation in the absence of measurement error. This fact follows from the key property (7.7).

7.2.2 Logistic Regression Corrected and Conditional Scores

Now we consider the multiple logistic regression model, $\text{pr}(\mathbf{Y} = 1 \mid \mathbf{Z}, \mathbf{X}) = H(\beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t\mathbf{X})$, where $H(t) = 1/\{1 + \exp(-t)\}$ is the logistic distribution function, and the classical additive, nondifferential error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with $\mathbf{U} = \text{Normal}(\mathbf{0}, \Sigma_{uu})$ where Σ_{uu} is known. Write the unknown regression parameter as $\Theta = (\beta_0, \beta_z^t, \beta_x^t)^t$.

The maximum likelihood score function for multiple logistic regression in the absence of measurement error is

$$\Psi_{\text{ML}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta) = [\mathbf{Y}_i - H\{(1, \mathbf{Z}_i^t, \mathbf{X}_i^t)\Theta\}] \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{X}_i \end{pmatrix}. \tag{7.10}$$

7.2.2.1 Logistic Regression Conditional Score

The conditional-score method for logistic regression is similar to that for linear regression. We again start by defining

$$\Delta_i = \mathbf{W}_i + \mathbf{Y}_i\Sigma_{uu}\beta_x. \tag{7.11}$$

Note that the definition in (7.11) differs slightly from that in (7.2), due to the absence of a variance parameter in logistic regression. Conditioned on $(\mathbf{Z}_i, \mathbf{X}_i)$, both $\mathbf{Y}_i$ and $\mathbf{W}_i$ have exponential family densities. Standard exponential family calculations (a good exercise) show that

$$\begin{aligned} E(\mathbf{Y}_i \mid \mathbf{Z}_i, \mathbf{X}_i, \Delta_i) &= H\left(\beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \Delta_i - \beta_x^t \Sigma_{uu}\beta_x/2\right) \\ &= E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) \\ &= \mathrm{pr}(\mathbf{Y}_i = 1 \mid \mathbf{Z}_i, \Delta_i). \end{aligned} \tag{7.12}$$

As in linear regression, the conditional distribution of $\mathbf{Y}_i$ given $(\mathbf{Z}_i, \Delta)$ does not depend on $\mathbf{X}_i$. It follows from (7.12) that the *conditional score*,

$$\Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix},$$

has the property that

$$E\left\{\Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) \mid \mathbf{Z}_i, \Delta_i\right\} = \mathbf{0},$$

so its unconditional mean also vanishes. Thus Ψ_{Cond} can be used to form unbiased estimating equations,

$$\textstyle\sum_{i=1}^n \Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0}, \tag{7.13}$$

to which the standard asymptotic theory on M-estimators in Section A.6 can be applied to approximate the distribution of $\widehat{\Theta}$. For issues of computation, see Section 7.5.

7.2.2.2 Logistic Regression Corrected Score

We now derive the corrected score for logistic regression using the general method of construction described in Section 7.4. The logistic model does not satisfy the smoothness conditions required by the corrected-score theory. However, Novick and Stefanski (2002) showed that even though the logistic score does not have the requisite smoothness properties, the corrected-score method can still be applied, and as long as the measurement error variance is not large, it produces nearly consistent estimators. In other words, when applied to logistic regression, the corrected-score method is approximate in the sense of reducing measurement error bias, but the quality of the approximation is so remarkably good that the bias is negligible in practice.

The method of construction is identical to that for the linear model with the one exception that Ψ_{ML} in (7.10) replaces Ψ_{LS} in (7.1). With $\widetilde{\mathbf{W}}_{b,i}$ defined as in (7.6), the corrected score for logistic regression is

$$\widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = B^{-1}\textstyle\sum_{b=1}^{B} \mathrm{Re}\left\{\Psi_{\text{ML}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_i, \Theta)\right\}.$$

Just as we did for the linear model in (7.2.1), it is possible to expand and simplify the expressions $\mathrm{Re}\{\Psi_{\text{ML}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_i, \Theta)\}$ to obtain an expression for $\widetilde{\Psi}_{\text{MCCS},B}$ in terms of standard functions. However, unlike the linear model, the resulting expression is not very enlightening, does not have a closed-form solution, and its limit as $B \to \infty$ is not easy to obtain. Expanding and simplifying are also not necessary for computing purposes, provided the programming software has complex number capabilities. Because the logistic model is not covered by the mathematical theory of corrected scores, analogues of neither (7.7) or (7.8) hold exactly, but both hold to a high degree of approximation.

As with the linear model, estimating equations are formed in the usual fashion, that is,

$$\textstyle\sum_{i=1}^n \widetilde{\Psi}_{\text{MCCS},B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0},$$

and large-sample inference uses the standard M-estimation methods in Section A.6.

7.2.3 Framingham Data Example

We illustrate the corrected- and conditional-score methods for logistic regression with the Framingham data used in the example of Section 4.3. All of the replicate measurements were used, and thus our variance estimate is based on 1,614 degrees of freedom and we proceed under the assumption that the sampling variability in this estimate is negligible, that is, the case of known measurement error.

Estimates and standard errors are in Table 7.1, for the conditional-score estimator (7.13), the corrected-score estimators (7.13) with $B = 10$ and $B = 10,000$, and the naive estimates for comparison. These estimates should be compared with those in Table 5.1, where almost the same answers are obtained. As explained in Section (7.2.2), conditional-score estimators are fully consistent as long as the logistic model and normal error model hold, and possess certain asymptotic variance optimality properties. The standard errors in Table 7.1 were computed from the sandwich-formula variance estimates in Section 7.5.1.

The difference among the three measurement error model estimates is clearly negligible. The equivalence, to the number of significant digits presented, of the corrected-score estimators for $B = 10$ and $B = 10,000$ supports the claim made in Section 7.2.1 that B does not need to be very large to obtain satisfactory results. The similarity between the conditional-score estimates and the corrected-score estimates supports the claim that the corrected-score procedure is, for most practical purposes, comparable to consistent methods, even though it is not covered by the theory in Section 7.4.

	Age	Smoke	Chol	LSBP
Naive	.055	.59	.0078	1.71
Std. Err.	.010	.24	.0019	.39
Conditional	.053	.60	.0078	1.94
Std. Err.	.011	.24	.0020	.44
Corrected ($B = 10$)	.054	.60	.0078	1.94
Std. Err.	.011	.24	.0020	.44
Corrected ($B = 10^4$)	.054	.60	.0078	1.94
Std. Err.	.011	.24	.0020	.44

Table 7.1 *Conditional-score and corrected-score estimates and sandwich standard errors from the Framingham data logistic regression analyses. Here "Smoke" is smoking status, "Chol" is cholesterol, and "LSBP" is log(SBP−50). Two sets of corrected-score estimates were calculated using different levels of Monte Carlo averaging,* $B = 10$ *and* $B = 10,000$.

7.2.3.1 Two Predictors Measured with Error

An appealing feature of the conditional- and corrected-score methods is the ease with which multiple predictors measured with error are handled. We now consider the Framingham logistic model for the case in which both systolic blood pressure and serum cholesterol are measured with error, first analyzed in Section 5.4.3 using SIMEX.

Recall that when serum cholesterol entered the model as a predictor measured with error, error modeling considerations indicated that a log transformation was appropriate to homogenize error variances. Thus, as in Section 5.4.3, the true-data model includes predictors $\mathbf{Z}_1 =$ age, $\mathbf{Z}_2 =$ smoking status, $\mathbf{X}_1 =$ log(cholesterol) at Exam #3 and $\mathbf{X}_2 =$ log(SBP−50) at Exam #3. The error model is $(\mathbf{W}_1, \mathbf{W}_2) = (\mathbf{X}_1, \mathbf{X}_2) + (\mathbf{U}_1, \mathbf{U}_2)$, where $(\mathbf{U}_1, \mathbf{U}_2)$ is bivariate normal with zero mean and covariance matrix Σ_u, with error covariance matrix estimate,

$$\widehat{\Sigma}_u = \begin{pmatrix} 0.00846 & 0.000673 \\ 0.000673 & 0.0126 \end{pmatrix}.$$

	Age	Smoke	LChol	LSBP
Naive	.056	.57	2.04	1.52
Std. Err.	.011	.24	.52	.37
SIMEX (STATA)	.055	.58	2.53	1.85
Std. Err.	.010	.26	.73	.45
Conditional	.054	.60	2.84	1.93
Std. Err.	.011	.25	.72	.47
Corrected ($B = 10^2$)	.054	.59	2.83	1.92
Std. Err.	.011	.25	.72	.47
Corrected ($B = 10^4$)	.054	.59	2.82	1.92
Std. Err.	.011	.25	.72	.47

Table 7.2 *Conditional-score and corrected-score estimates and sandwich standard errors from the Framingham data logistic regression analyses with both SBP and cholesterol measured with error. Here "Smoke" is smoking status, "LChol" is log(cholesterol), and "LSBP" is log(SBP−50). Two sets of corrected-score estimates were calculated using different levels of Monte Carlo averaging,* $B = 100$ *and* $B = 10,000$.

The two error variances result in marginal reliability ratios of $\lambda_1 = 0.73$ and $\lambda_2 = 0.76$, respectively for $\mathbf{W}_1$ and $\mathbf{W}_2$, with linear model corrections for attenuation of $1/\lambda_1 = 1.37$ and $1/\lambda_2 = 1.32$. So in the absence of strong multicollinearity the conditional- and corrected-score estimators of the coefficients of log(cholesterol) and log(SBP−50) should be inflated by approximately 37% and 32%, compared to the naive estimates.

The results of the analysis displayed in Table 7.2 are consistent with expectations. Difference between the conditional- and corrected-score estimates are negligible, and the bias correction in the estimates is consistent with the reliability ratios reported above, for log(cholesterol), $2.82/2.04 = 1.38 \approx 1.37$ and for log(SBP−50), $1.92/1.52 = 1.26 \approx 1.32$. For comparison, we include the naive and SIMEX estimates from Section 5.4.3. Recall that the SIMEX estimates are somewhat undercorrected for bias, as explained in Section 5.4.3.

As in Table 7.1 for the analysis assuming only ln(SBP−50) is measured with error, we calculated the Monte Carlo corrected score estimates using two different levels of Monte Carlo averaging. However, for the present model we took the lower level equal to $B = 100$, not $B = 10$ used in Table 7.1. The averaging required in the Monte Carlo corrected score, see (7.27), is effectively calculating an integral. As with any numerical method of integration, higher-dimensional integration require greater computational effort. Thus, the more variables measured with error, the larger one should take B.

7.3 Conditional Score Functions

In this section, we describe the conditional-score estimators of Stefanski and Carroll (1987) for an important class of generalized linear models. We first present the basic theory, followed by conditional scores for specific models. Finally, certain extensions are presented to describe the range of applications of the conditional-score approach. Once again, we note that this section, like this chapter as a whole, is heavy with formulae and algebra, but exhibiting the formulae makes the methods usable.

7.3.1 Conditional Score Basic Theory

7.3.1.1 Generalized Linear Models (GLIM)

Canonical generalized linear models (McCullagh and Nelder, 1989) for $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ have density or mass function

$$f(y|z, x, \Theta) = \exp\left\{\frac{y\eta - \mathcal{D}(\eta)}{\phi} + c(y, \phi)\right\}, \tag{7.14}$$

where $\eta = \beta_0 + \beta_z^t z + \beta_x^t x$ is called the *natural parameter*, and $\Theta = (\beta_0, \beta_z^t, \beta_x^t, \phi)$ is the unknown parameter to be estimated. The mean and variance of $\mathbf{Y}$ are $\mathcal{D}'(\eta)$ and $\phi\mathcal{D}''(\eta)$. This class of models includes:

- linear regression: mean $= \eta$, variance $= \phi$, $\mathcal{D}(\eta) = \eta^2/2$, $c(y, \phi) = -y^2/(2\phi) - \log(\sqrt{2\pi\phi}\ )$;
- logistic regression: mean $= H(\eta)$, variance $= H'(\eta)$, $\phi \equiv 1$, $\mathcal{D}(\eta) = -\log\{1 - H(\eta)\}$, $c(y, \phi) = 0$, where $H(x) = 1/\{1 + \exp(-x)\}$;
- Poisson loglinear regression: mean $= \exp(\eta)$, variance $= \exp(\eta)$, $\phi \equiv 1$, $\mathcal{D}(\eta) = \exp(\eta)$, $c(y, \phi) = -\log(y!)$;
- Gamma inverse regression: mean $= -1/\eta$, variance $= -\phi/\eta$, $\mathcal{D}(\eta) = -\log(-\eta)$, $c(y, \phi) = \phi^{-1}\log(y/\phi) - \log\{y\Gamma(1/\phi)\}$.

If the $\mathbf{X}_i$ were observed, then Θ is estimated by solving

$$\textstyle\sum_{i=1}^n \Psi_{\rm QL}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i) = \mathbf{0}, \tag{7.15}$$

where

$$\Psi_{\rm QL}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i) = \begin{pmatrix} \{\mathbf{Y}_i - \mathcal{D}'(\eta_i)\}\begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{X}_i \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\phi - \dfrac{\{\mathbf{Y}_i - \mathcal{D}'(\eta_i)\}^2}{\mathcal{D}''(\eta_i)} \end{pmatrix}, \tag{7.16}$$

where $\eta_i = \beta_0 + \beta_z^t\mathbf{Z}_i + \beta_x^t\mathbf{X}_i$.

For certain models (7.15) produces maximum likelihood estimators apart from the degrees of freedom correction $(n - p)/n$. However, in general it results in quasilikelihood estimators; see Section A.

7.3.1.2 GLIM MEMs and Conditional Scores

Assume now that the measurement error is additive and normally distributed, with error covariance matrix Σ_{uu}. If $\mathbf{X}$ is regarded as an unknown parameter and all other parameters are assumed known, then

$$\Delta = \mathbf{W} + \mathbf{Y}\Sigma_{uu}\beta_x/\phi \tag{7.17}$$

is a sufficient statistic for $\mathbf{X}$ (Stefanski and Carroll, 1987). Furthermore, the conditional distribution of $\mathbf{Y}$ given $(\mathbf{Z}, \Delta) = (z, \delta)$ is a canonical generalized linear model of the same form as (7.14) with certain changes. With $(\mathbf{Y}, \mathbf{Z}, \Delta) = (y, z, \delta)$, replace x with δ, and η, c, and $\mathcal{D}$ with

$$\begin{aligned} &\eta_* = \beta_0 + \beta_z^t z + \beta_x^t\delta; \\ &c_*(y, \phi, \beta_x^t\Sigma_{uu}\beta_x) = c(y, \phi) - (1/2)(y/\phi)^2\beta_x^t\Sigma_{uu}\beta_x; \\ &\mathcal{D}_*(\eta_*, \phi, \beta_x^t\Sigma_{uu}\beta_x) \\ &\quad = \phi\log\left[\int \exp\left\{y\eta_*/\phi + c_*(y, \phi, \beta_x^t\Sigma_{uu}\beta_x)\right\}\, d\mu(y)\right], \end{aligned}$$

where the last term is a sum if $\mathbf{Y}$ is discrete and an integral otherwise. This means that the conditional density or mass function is

$$\begin{aligned} &f(y|z, \delta, \Theta, \Sigma_{uu}) = \\ &\exp\left\{\frac{y\eta_* - \mathcal{D}_*(\eta_*, \phi, \beta_x^t\Sigma_{uu}\beta_x)}{\phi} + c_*(y, \phi, \beta_x^t\Sigma_{uu}\beta_x)\right\}, \end{aligned} \tag{7.18}$$

where $\eta_* = \beta_0 + \beta_z^t z + \beta_x^t\delta$.

The correspondence between (7.14) and (7.18) suggests simply substituting $\mathcal{D}_*(\eta_*, \phi, \beta_x^t\Sigma_{uu}\beta_x)$ for $\mathcal{D}(\eta)$ into (7.15)–(7.16), and solving the resulting equations replacing η_i by $\eta_{*,i} = \beta_0 + \beta_x^t\Delta_i + \beta_z^t\mathbf{Z}_i$, noting that Δ_i depends on β_x and ϕ. This simple idea is easily implemented and produces consistent estimators.

The conditional mean and variance of $\mathbf{Y}$ given $(\mathbf{Z}, \Delta)$ are determined by the derivatives of $\mathcal{D}_*$ with respect to η_*, that is,

$$\begin{aligned} E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) &= m(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \frac{\partial}{\partial \eta_*} \mathcal{D}_*; \\ \text{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) &= \phi v(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \phi \frac{\partial^2}{\partial \eta_*^2} \mathcal{D}_*. \qquad (7.19) \end{aligned}$$

The estimates of $\Theta = (\beta_0, \beta_x, \beta_z, \phi)$ are obtained by solving

$$\textstyle\sum_{i=1}^n \Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0},$$

where

$$\Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \left[\begin{array}{c} \{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\phi - \dfrac{\{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\}^2}{\text{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)/\phi} \end{array} \right] \qquad (7.20)$$

with $\eta_{*,i} = \beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \Delta_i$, with $\Delta_i = \mathbf{W}_i + \mathbf{Y}_i \Sigma_{uu} \beta_x / \phi$. Stefanski and Carroll (1987) discuss a number of ways of deriving unbiased estimating equations from (7.18) and (7.19). The approach described here is the simplest to implement.

7.3.2 Conditional Scores for Basic Models

In Sections 7.2.1 and 7.2.2, we presented the conditional scores for linear and logistic regression, respectively. It is an informative exercise to derive those formulae from the general theory, and we leave it to the reader to do so. In this section, we show how to derive the conditional scores in more complex models.

7.3.2.1 Poisson Regression

Linear and logistic regression are the only common canonical models for which $\mathcal{D}_*'$ and ${\mathcal{D}_*}''$ have closed-form expressions. In general, either numerical integration or summation is required to determine the moments (7.19). For example, for Poisson regression (for which $\phi \equiv 1$),

$$\mathcal{D}_*(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \log \left\{ \sum_{y=0}^{\infty} (y!)^{-1} \exp(y\eta_* - y^2 \beta_x^t \Sigma_{uu} \beta_x / 2) \right\}.$$

For this model, $\eta_* = \beta_0 + \beta_z^t \mathbf{Z} + \beta_x^t \Delta$ and

$$E(\mathbf{Y}^j \mid \mathbf{Z}, \Delta) = \frac{\sum_{y=0}^{\infty} y^j (y!)^{-1} \exp\{y(\eta_*) - y^2 \beta_x^t \Sigma_{uu} \beta_x / 2\}}{\sum_{y=0}^{\infty} (y!)^{-1} \exp\{y(\eta_*) - y^2 \beta^t \Sigma_{uu} \beta_x / 2\}}, \qquad (7.21)$$

and computation of the mean and variance functions requires numerical summation unless $\beta_x^t \Sigma_{uu} \beta_x = 0$.

7.3.2.2 Linear and Logistic Models with Interactions

Consider the usual form of the generalized linear model (7.14) with the difference that $\eta = \beta_0 + \beta_z^t z + \beta_x^t x + x^t \beta_{xz} z$ where β_{xz} is a $\dim(\mathbf{X}) \times \dim(\mathbf{Z})$ matrix of interaction parameters. Conditional-score estimation for this model was studied by Dagalp (2001). The model allows for interactions between the variables measured with error and those measured without error. In particular, it allows for analysis of covariance models with some of the covariates measured with error by having $\mathbf{Z}$ indicate group membership. The appropriate elements of β_{xz} can be constrained to equal zero if the model does not contain all possible interactions. The full parameter vector is denoted by $\Theta = (\beta_0, \beta_z^t, \beta_x^t, \text{vec}_*(\beta_{xz})^t, \phi)$ where vec$_*$ denotes the operator that maps the non-zero-constrained elements of the parameter matrix reading left to right, and top to bottom to a column vector. Assuming the normal measurement error model, $\mathbf{W} = \text{Normal}(\mathbf{X}, \Sigma_{uu})$, the distribution of the observed data again admits a sufficient statistic for $\mathbf{X}$,

$$\Delta = \mathbf{W} + \mathbf{Y} \Sigma_{uu} (\beta_x + \beta_{xz} \mathbf{Z}) / \phi.$$

This means that we can obtain unbiased score functions in the same fashion as with previous models, taking care to include components for the interaction components. For this model,

$$\Psi_{\text{Cond}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \left[\begin{array}{c} \{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \\ \mathbf{Z}_i \otimes \Delta \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\phi - \dfrac{\{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\}^2}{\text{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)/\phi} \end{array} \right], \qquad (7.22)$$

where $\mathbf{Z}_i \otimes \Delta$ represents a column vector of length $\dim\{\text{vec}_*(\beta_{xz})\}$ containing the product of the kth element of $\mathbf{Z}_i$ and the rth element of Δ_i if and only if the (r, k) element of β_{xz} in not constrained to equal zero.

Define

$$\xi = \beta_x + \beta_{xz} \mathbf{Z}.$$

For linear regression with $\phi = \sigma^2$ the required conditional expectations are

$$E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) = \frac{\beta_0 + \beta_z^t \mathbf{Z} + \xi^t \Delta}{1 + \xi^t \Sigma_{uu} \xi / \sigma^2},$$

$$\text{var}(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) = \frac{\sigma^2}{1 + \xi^t \Sigma_{uu} \xi / \sigma^2}.$$

For logistic regression, $\phi \equiv 1$, only the top component of (7.22) is relevant, and we need only the first conditional moment,

$$E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i) = \text{pr}(\mathbf{Y} = 1 \mid \mathbf{Z}_i, \Delta_i) = H(\beta_0 + \beta_z^t \mathbf{Z} + \xi^t \Delta - \xi^t \Sigma_{uu} \xi / 2),$$

where $H(\cdot)$ is, as usual, the logistic distribution function.

7.3.3 Conditional Scores for More Complicated Models

The following examples provide a sample of models for which conditional-score methods have been derived and studied since the first edition in 1995. The models, and the technical details of the derivations and the score functions, are generally more complicated than those considered previously. Our intent is to illustrate the range of application of the conditional-score approach, and we omit many of the mathematical details, describing only the models and the relevant conditioning sufficient statistic.

7.3.3.1 Conditional Scores with Instrumental Variables

Buzas and Stefanski (1996c) studied conditional-score estimation for the generalized linear model in (7.14) with observed predictor following the additive error model, $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U} = \text{Normal}(\mathbf{0}, \Sigma_{uu})$, for the case that Σ_{uu} is unknown but an instrument is observed,

$$\mathbf{T} = \text{Normal}(\gamma_1 + \gamma_z \mathbf{Z} + \gamma_x \mathbf{X}, \Omega), \tag{7.23}$$

where the parameters in (7.23) are also unknown. This is a version of the model studied in Section 6.3.2, with the additional structure of (7.14) imposed on the primary model relating $\mathbf{Y}$ to $\mathbf{X}$ and the multivariate linear model structure of (7.23). Note that in the most general case with $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{T}$ vector-valued, the regression in (7.23) is multivariate and γ_1, γ_z, and γ_x are matrices of the appropriate dimensions. For this model Buzas and Stefanski (1996c) derive conditional-score functions under the assumptions that $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{T}$ are conditionally independent given $\mathbf{Z}$ and $\mathbf{X}$, and $\text{rank}(\gamma_x) = \dim(\mathbf{X})$. The latter assumption requires at least as many instruments as variables measured with error. Under these assumptions

$$\Delta = \mathbf{W} + \mathbf{Y}\Sigma_{uu}\beta_x/\phi + \Sigma_{uu}\gamma_x^t \Omega^{-1}\mathbf{T}$$

is a sufficient statistic for $\mathbf{X}$ when all other parameters are assumed known. Conditional scores are obtained by conditioning on this statistic.

7.3.3.2 Proportional Hazards Model with Longitudinal Covariates Measured with Error

Tsiatis and Davidian (2001) used conditional score techniques to eliminate subject-specific, time-dependent covariate process parameters when the time-dependent covariate process is measured with error. In their model, the observed data for each subject includes the time on study $\mathbf{V}_i$, failure indicator F_i, error-free time-independent covariate $\mathbf{Z}_i$, and longitudinal measurements $\mathbf{W}_i(t_{ij}) = \mathbf{X}_i(t_{ij}) + \epsilon_{ij}$, $t_{i1} < \cdots < t_{i,k_i}$, where the unobserved time-dependent covariate process is modeled as $\mathbf{X}_i(u) = \alpha_{oi} + \alpha_{1i}u$ and the errors ϵ_{ij} are independent $\text{Normal}(0, \sigma^2)$. The survival model assumes that the hazard of failure is $\lambda_i(u) = \lambda_0(u)\exp\{\gamma \mathbf{X}_i(u) + \eta^t \mathbf{Z}_i\}$.

Defining $\widehat{\mathbf{X}}_i(u)$ to be the ordinary least squares estimator of $\mathbf{X}_i(u)$ using all of the longitudinal data up to and including time u, the counting process increment, $dN_i(u) = I(u \le V_i < u + du, F_i = 1, t_{i2} \le u)$, and the at-risk process $Y_i(u)I(V_i \ge u, t_{i2} \le u)$, Tsiatis and and Davidian's assumptions are such that conditioned on $\{\alpha_i, Y_i(u) = 1, \mathbf{Z}_i\}$, $\widehat{\mathbf{X}}_i(u) = \text{Normal}\{\alpha_{oi} + \alpha_{1i}u, \sigma^2\theta_i(u)\}$, where $\theta_i(u)$ is known. It follows that up to order du the conditional likelihood of $\{dN_i(u), \widehat{\mathbf{X}}_i(u)\}$ given $\{Y_i(u) = 1, \alpha_i, \mathbf{Z}_i\}$ admits a sufficient statistic for $\mathbf{X}_i(u)$ of the form

$$\Delta_i(u) = \widehat{\mathbf{X}}_i(u) + \gamma\sigma^2\theta_i(u)dN_i(u). \tag{7.24}$$

The statistic $\Delta_i(u)$ is used to derive conditional estimating equations free of the α_i by conditioning on Δ_i. Note the similarity of (7.24) to (7.17). Because of the formal equivalence between proportional hazard partial likelihood and logistic regression likelihood, the technical details of the corrected score for the proportional hazard model are similar to those for logistic regression.

7.3.3.3 Matched Case-Control Studies with Covariate Measurement Error

McShane, Midthune, Dorgan, et al. (2001) used the conditional-score method to derive estimators for matched case-control studies when covariates are measured with error. Their study design was a $1 : M$ matched case-control study with K strata, where in the absence of measurement error the preferred method of inference is based on the conditional prospective likelihood,

$$\text{pr}\left[\mathbf{Y}_1, \ldots, \mathbf{Y}_k \mid \{\mathbf{X}_k, \mathbf{Z}_k, (T_k = 1)\}_{k=1}^K\right] = \prod_{k=1}^K \frac{\exp\left\{\sum_{j=1}^{M+1} Y_{kj}(\mathbf{X}_{kj}^t\beta_x + \mathbf{Z}_{kj}^t\beta_z)\right\}}{\sum_{j=1}^{M+1}\exp(\mathbf{X}_{kj}^t\beta_x + \mathbf{Z}_{kj}^t\beta_z)},$$

where $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{k,M+1})$ is the vector of binary responses for the $M+1$ subjects in the kth stratum, $T_k = \sum_{j=1}^{M+1} Y_{kj}$, $(\mathbf{X}_{kj}^t, \mathbf{Z}_{kj}^t)^t$ is the error-free covariate for the jth subject in the kth stratum, and $\mathbf{X}_k = (\mathbf{X}_{k1}^t, \ldots, \mathbf{X}_{k,M+1}^t)^t$, $\mathbf{Z}_k = (\mathbf{Z}_{k1}^t, \ldots, \mathbf{Z}_{k,M+1}^t)^t$. The measurement error model is a Gaussian, nondifferential additive model with $\mathbf{W}_{kj} = \mathbf{X}_{kj} + \mathbf{U}_{kj}$, where the model for the errors $\mathbf{U}_{kj}$ allows for multiple additive components subject to certain restrictions.

With $\mathcal{B}_{k,x} = (Y_{k2}\beta_x^t, \ldots, Y_{k,M+1}\beta_x^t)^t$, $\mathbf{D}_{kz} = (\mathbf{Z}_{k2}^t, \ldots, \mathbf{Z}_{k,M+1}^t)^t - \mathbf{Z}_{k1}^t$, $\mathbf{D}_{kx} = (\mathbf{X}_{k2}^t, \ldots, \mathbf{X}_{k,M+1}^t)^t - \mathbf{X}_{k1}^t$, $\mathbf{D}_{kw} = (\mathbf{W}_{k2}^t, \ldots, \mathbf{W}_{k,M+1}^t)^t - \mathbf{W}_{k1}^t$, and $\mathbf{D}_{ku} = \mathbf{D}_{kw} - \mathbf{D}_{kx}$, where $\Sigma_{d_u,d_u} = \mathrm{cov}(\mathbf{D}_{ku}, \mathbf{D}_{ku})$, McShane et al. (2001) showed that

$$\Delta_k = \mathbf{D}_{kw} + \Sigma_{d_u,d_u}\mathcal{B}_{k,x}$$

is sufficient for $\mathbf{D}_{kx}$ when $\mathbf{D}_{kx}$ is regarded as a parameter and all other parameters are assumed known, $k = 1, \ldots, K$. Thus by conditioning on the Δ_k, estimating equations can be derived that do not depend on the unobserved $\mathbf{X}_{kj}$.

7.3.3.4 Joint Models with Subject-Specific Parameters

Joint models are discussed in greater detail in Section 11.7. Here, we consider a particular joint model that is amenable to the conditional score method. Li, Zhang, and Davidian (2004) adapted the conditional-score method to joint models with subject-specific random effects. Rather than model the distribution of the subject-specific effects, they showed how to derive conditional scores that are free of the subject-specific effects. In their model the i^{th} subject has observed data: $\mathbf{Y}_i$, the primary response; $\mathbf{Z}_i$, the error-free predictors; and longitudinal measurements $\mathbf{W}_i = (W_{i1}, \ldots, W_{ik_i})^t$ with W_{ij} measured at time t_{ij}. The longitudinal data are assumed to follow the model $\mathbf{W}_i = \mathbf{D}_i\mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{D}_i$ is a $k_i \times q$ full-rank design matrix depending on t_{ij}, $\mathbf{X}_i$ is a random, subject-specific effect modeling features of the i^{th} longitudinal profile, and $\mathbf{U}_i$ are Normal$(\mathbf{0}, \sigma_u^2\boldsymbol{I})$, independent of $\mathbf{X}_i$ and across i.

Li, Zhang, and Davidian (2004) assumed that conditioned on $(\mathbf{Z}_i, \mathbf{X}_i)$ the primary endpoint $\mathbf{Y}_i$ follows a generalized linear model of the form (7.14). It follows that

$$\Delta_i = \mathbf{D}_i^t\mathbf{W}_i + \sigma_u^2\mathbf{Y}_i\beta_x/\phi \tag{7.25}$$

is sufficient for $\mathbf{X}_i$ when all other parameters are assumed known. They then derived and studied conditional score estimators, as well as another conditional estimator described by Stefanski and Carroll (1987).

It is instructive to reconcile the statistic in (7.25) with the form of the statistic given in (7.17) for the general model. Starting with the linear model $\mathbf{W}_i = \mathbf{D}_i\mathbf{X}_i + \mathbf{U}_i$, multiplication by $(\mathbf{D}_i^t\mathbf{D}_i)^{-1}\mathbf{D}_i^t$ results in the unbiased error model $\mathbf{W}_i^* = \mathbf{X}_i + \mathbf{U}_i^*$ where $\mathbf{W}_i^* = (\mathbf{D}_i^t\mathbf{D}_i)^{-1}\mathbf{D}_i^t\mathbf{W}_i$ and $\mathbf{U}_i^* = (\mathbf{D}_i^t\mathbf{D}_i)^{-1}\mathbf{D}_i^t\mathbf{U}_i$ is Normal$\{\mathbf{0}, (\mathbf{D}_i^t\mathbf{D}_i)^{-1}\sigma_u^2\}$. If we now substitute $\mathbf{W}_i^*$ for $\mathbf{W}$ and $(\mathbf{D}_i^t\mathbf{D}_i)^{-1}\sigma_u^2$ for Σ_{uu} into the expression for the sufficient statistic given in (7.17), we get $\Delta_i^* = \mathbf{W}_i^* + \mathbf{Y}(\mathbf{D}_i^t\mathbf{D}_i)^{-1}\sigma_u^2\beta_x/\phi = (\mathbf{D}_i^t\mathbf{D}_i)^{-1}\left\{\mathbf{D}_i^t\mathbf{W}_i + \sigma_u^2\mathbf{Y}_i\beta_x/\phi\right\}$ In other words, after transforming to an unbiased error model, the sufficient statistic in (7.25) is a matrix multiple of the general model sufficient statistic in (7.17). The facts that the matrix, $(\mathbf{D}_i^t\mathbf{D}_i)^{-1}$, is known, and that a known, full-rank multiple of a sufficient statistic is also sufficient, establish the link between the general theory statistic (7.17) and the form of the statistic (7.25) used by Li, Zhang, and Davidian (2004).

7.4 Corrected Score Functions

This section gives the basic theory and the algorithm of corrected score functions. It applies to any model for which the usual estimator in the absence of measurement error is an M-estimator. The basic idea is very simple:

- Let $\Psi_{\text{True}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta)$, where Θ is the collection of all unknown parameters, denote the score function that would be used for estimation if $\mathbf{X}$ were observed. This could be a nonlinear least-squares score, a likelihood score (derivative of the loglikelihood), etc.
- Because $\mathbf{X}$ is not observed and hence $\Psi_{\text{True}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta)$ cannot be used for estimation, we do the next best thing and construct an unbiased estimator of $\Psi_{\text{True}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta)$ based on the observed data. This new score function is $\Psi_{\text{CS}}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)$. It has the property that $E\{\Psi_{\text{CS}}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)\} = \Psi_{\text{True}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta)$, and thus is also unbiased.
- The *corrected score function*, $\Psi_{\text{CS}}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)$, is used for estimation of Θ, the calculation of standard errors, inference, etc., using the M-estimation techniques in Section (A.6).

There are basically two ways to find the corrected score function:

1. Be clever! In some cases, one can intuit the corrected score function exactly. Some examples where this is possible are given in Section 7.4.3.
2. When intuition is lacking, or the corrected score is prohibitively complicated, an alternative is to use the complex variable theory, as we did in Section 7.2, and let the computer calculate the score function and solve it. Obviously, if we can be clever, we would not use the Monte Carlo approach, but the Monte Carlo approach expands the possible applications of the methodology. We discuss this approach in detail in Section 7.4.2.

7.4.1 Corrected Score Basic Theory

The method of corrected scores does not assume a model for the observed data per se. Rather, it starts with the assumption that there exists an unbiased score function that produces consistent estimators with error-free data. This is the true-data score function described above and called $\Psi_{\rm True}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta)$. The true-data score function should have the property that if $\mathbf{X}$ were observable, the score function would be unbiased, that is,

$$E\{\Psi_{\rm True}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \theta)|\mathbf{Z}, \mathbf{X}\} = \mathbf{0}.$$

For the linear and logistic regression models in Sections 7.2.1 and 7.2.2, $\Psi_{\rm True}$ was the least squares and maximum likelihood score in (7.1) and (7.10), respectively.

A corrected score is a function, $\Psi_{\rm CS}$, of the observed data having the property that it is unbiased for the true-data score function. In symbols, this means that

$$E\left\{\Psi_{\rm CS}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)|\mathbf{Y}, \mathbf{Z}, \mathbf{X})\right\} = \Psi_{\rm True}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta).$$

It follows from (7.26) and (7.26) that $\Psi_{\rm CS}$ is also conditionally unbiased, that is, $E\{\Psi_{\rm CS}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta) = \mathbf{0}$. Thus, by the general theory of M-estimation, the estimating equations,

$$\textstyle\sum_{i=1}^n \Psi_{\rm CS}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0},$$

possess a consistent, asymptotically normally sequence of solutions (Nakamura, 1990), whose asymptotic distribution is readily approximated using the M-estimation techniques in Section A.6.

Note that no assumptions about the $\mathbf{X}_i$ are made. Thus, the corrected-score method provides an attractive approach to consistent estimation when data are measured with error. The key technical problem is finding a corrected score satisfying (7.26) for a given $\Psi_{\rm True}$. Corrected scores have been identified for particular models by Nakamura (1990) and Stefanski (1989), and the results in Gray, Watkins, and Schucany (1973) provide a means of obtaining corrected scores via infinite series. These calculations are described in Section 7.4.3. In the absence of such exact results, Novick and Stefanski (2002) describe a general method of constructing corrected scores based on simple Monte Carlo averaging. We now outline their method.

7.4.2 Monte Carlo Corrected Scores

7.4.2.1 The Algorithm

The algorithm is simple, although perhaps with complex numbers *simple* is not the most appropriate word. The method is as follows.

- For $b = 1, ..., B$, generate random numbers $\mathbf{U}_{b,i}$ that are normally distributed with mean zero and covariance matrix Σ_{uu}.
- Form the complex-valued random variables

$$\widetilde{\mathbf{W}}_{b,i} = \mathbf{W}_i + \boldsymbol{\iota}\mathbf{U}_{b,i}. \tag{7.26}$$

  where $\boldsymbol{\iota} = \sqrt{-1}$.
- Define the Monte Carlo corrected score

$$\Psi_{\rm MCCS,B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = B^{-1}\textstyle\sum_{b=1}^B {\rm Re}\{\Psi_{\rm True}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_{b,i}, \Theta)\}. \tag{7.27}$$

- Get an estimator for Θ by solving the corrected-score estimating equations

$$\textstyle\sum_{i=1}^n \Psi_{\rm MCCS,B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \mathbf{0}.$$

- As $B \to \infty$, $\Psi_{\rm MCCS,B}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) \to \Psi_{\rm CS}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta)$. The number of generated random variables per subject, B, needs to be large enough to make this limit approximately correct. Often, however, rather small values of B suffice.
- The resulting corrected-score estimators are M-estimators to which the standard asymptotic results in Section A.6 apply.

7.4.2.2 The Theory

A mathematical result on which the corrected-score theory is based is that for suitably smooth, integrable functions, $\mathbf{f}(\cdot)$, the function defined by

$$\widetilde{\mathbf{f}}(\mathbf{W}_i) = E\left[{\rm Re}\left\{\mathbf{f}(\widetilde{\mathbf{W}}_{b,i})\right\}|\mathbf{X}_i, \mathbf{W}_i\right] \tag{7.28}$$

does not depend on $\mathbf{X}_i$ and is an unbiased estimator of $\mathbf{f}(\mathbf{X}_i)$, where Re{} denotes the real part of its argument, that is,

$$E\left\{\widetilde{\mathbf{f}}(\mathbf{W}_i)|\mathbf{X}_i\right\} = \mathbf{f}(\mathbf{X}_i) \tag{7.29}$$

(Stefanski, 1989; Stefanski and Cook, 1995). We will not prove the general result here. However, verification of (7.29) for the function $g(\mathbf{x}) = \exp(\mathbf{c}^t\mathbf{x})$ is instructive and also provides results that are used later in this section. First, note that by independence of $\mathbf{U}_{b,i}$ and $(\mathbf{X}_i, \mathbf{W}_i)$ and properties of the normal distribution characteristic function,

$$\begin{aligned} \widetilde{g}(\mathbf{W}_i) &= E\left[{\rm Re}\left\{g(\widetilde{\mathbf{W}}_{b,i})\right\}|\mathbf{X}_i, \mathbf{W}_i\right] \\ &= \exp(\mathbf{c}^t\mathbf{W}_i)\exp(-\mathbf{c}^t\Sigma_{uu}\mathbf{c}/2). \end{aligned}$$

Now the fact that $E\{\widetilde{g}(\mathbf{W}_i)|\mathbf{X}_i\} = g(\mathbf{X}_i)$ follows immediately from the normal moment generating function identity, $E\{\exp(\mathbf{c}^t\mathbf{W}_i)|\mathbf{X}_i\} = \exp(\mathbf{c}^t\mathbf{X}_i + \mathbf{c}^t\Sigma_{uu}\mathbf{c}/2)$.

The special case $g(\mathbf{X}_i) = \exp(\mathbf{c}^t\mathbf{X}_i)$ is very useful. It follows from the result for the exponential function that (7.29) also holds for the partial derivative $(\partial/\partial c)g(\mathbf{X}_i) = \mathbf{X}_i\exp(\mathbf{c}^t\mathbf{X}_i)$ and for higher-order derivatives, as well. Also, it is clear that (7.29) holds for linear combinations of exponentials, $\sum_j \exp(\mathbf{c}_j^t\mathbf{X}_i)$, and their partial derivatives with respect to $\mathbf{c}_j$. These results can be used to find the exact corrected scores in Section (7.4.3).

It follows from (7.29), that for score functions with components that are sufficiently smooth and integrable functions of their third argument,

$$E\left[\mathrm{Re}\{\Psi_{\text{True}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_{b,i}, \Theta)\}|(\mathbf{Y}_i, \mathbf{X}_i)\right] = \Psi_{\text{True}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta),$$

that is, $\mathrm{Re}\{\Psi_{\text{True}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_{b,i}, \Theta)\}$ is a corrected score.

The corrected score $\mathrm{Re}\{\Psi_{\text{True}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_{b,i}, \Theta)\}$ in (7.30) depends on the particular generated random vector $\mathbf{Z}_{b,i}$. The preferred corrected score is $E\{\Psi_{\text{True}}(\mathbf{Y}_i, \mathbf{Z}_i, \widetilde{\mathbf{W}}_{b,i}, \Theta)|(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)\}$, which eliminates variability due to $\mathbf{U}_{b,i}$. The conditional expectation is not always easy to determine mathematically. However, Monte Carlo integration provides a simple solution, resulting in the Monte Carlo corrected score (7.27). The Monte Carlo corrected score possesses the key property,

$$E\{\Psi_{\text{MCCS,B}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) \mid \mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i\} = \Psi_{\text{True}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta),$$

from (7.30) and converges to the exact conditional expectation desired as $B \to \infty$, that is,

$$\Psi_{\text{CS}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta) = \lim_{B\to\infty} \Psi_{\text{MCCS,B}}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta).$$

Corrected-score estimating equations are formed in the usual fashion as described above.

7.4.3 Some Exact Corrected Scores

The exponential function $g(\mathbf{X}) = \exp(\mathbf{c}^t\mathbf{X})$ is a special case of (7.29) studied in Section 7.4, and extensions derived from it are useful for finding exact corrected scores when the true-data score functions are linear combinations of products of powers and exponential.

7.4.3.1 Likelihoods with Exponentials and Powers

One useful class of models that admit corrected scores contains those models with log-likelihoods of the form

$$\log\{f(y|z, x, \Theta)\} =$$
$$\textstyle\sum_{k=0}^{2}\left\{c_k(y, z, \Theta)(\beta_x^t x)^k\right\} + c_3(y, z, \Theta)\exp(\beta_x^t x);$$

see the examples given below. Then, using normal distribution moment generating function identities, the required function is

$$\begin{aligned}&\Psi_*(y, z, w, \Theta, \Sigma_{uu}) = \\ &\quad\frac{\partial}{\partial\Theta^t}\left[\sum_{k=0}^{2}\left\{c_k(y, z, \Theta)(\beta_x^t w)^k\right\} - c_2(y, z, \Theta)\beta_x^t\Sigma_{uu}\beta_x\right. \\ &\qquad\qquad \left. + c_3(y, z, \Theta)\exp(\beta_x^t w - .5\beta_x^t\Sigma_{uu}\beta_x)\right].\end{aligned}$$

Regression models in this class include:

- Normal linear with mean $= \eta$, variance $= \phi$, $c_0 = -(y - \beta_0 - \beta_z^t z)^2/(2\phi) - \log(\sqrt{\phi})$, $c_1 = (y - \beta_0 - \beta_z^t z)/\phi$, $c_2 = -(2\phi)^{-1}$, $c_3 = 0$.
- Poisson with mean $= \exp(\eta)$, variance $= \exp(\eta)$, $c_0 = y(\beta_0 + \beta_z^t z) - \log(y!)$, $c_1 = y$, $c_2 = 0$, $c_3 = -\exp(\beta_0 + \beta_z^t z)$.
- Gamma with mean $= \exp(\eta)$, variance $= \phi\exp(2\eta)$, $c_0 = -\phi^{-1}(\beta_0 + \beta_z^t z) + (\phi^{-1} - 1)\log(y) + \phi^{-1}\log(\phi^{-1}) - \log\left\{\Gamma(\phi^{-1})\right\}$, $c_1 = \phi^{-1}$, $c_2 = 0$, $c_3 = -\phi^{-1}y\exp(-\beta_0 - \beta_z^t z)$.

7.4.4 SIMEX Connection

There is a connection between SIMEX and the Monte Carlo corrected-score method. SIMEX adds measurement error multiple times, computes the new estimator over each generated data set, and then extrapolates back to the case of no measurement error. The sequence of operations in simulation extrapolation is 1) generate pseudo-random, (real-valued) remeasured data sets as $\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \sqrt{\zeta}\Sigma_{uu}^{1/2}\mathbf{U}_{b,i}$; 2) calculate average estimates from the remeasured data sets; 3) determine the dependence of the averaged estimates on ζ; and 4) extrapolate to $\zeta = -1$.

The Monte Carlo corrected-score method is obtained more or less by reordering these operations. It starts with the complex-valued, pseudo-random data sets. Note that these are obtained by letting $\zeta \to -1$ in the SIMEX remeasured data, $\lim_{\zeta\to-1}\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \iota\Sigma_{uu}^{1/2}\mathbf{U}_{b,i}$. Rather than calculate an estimate from each complex pseudodata set and averaging, the complex, pseudodata estimating equations are averaged, the imaginary part is removed, and then the averaged equations are solved, resulting in a single estimate.

7.4.5 Corrected Scores with Replicate Measurements

The connection to SIMEX described in the previous section extends even further. In Section 5.3.1, we described a version of the SIMEX algorithm that automatically accommodates heteroscedastic measurement error with unknown variances, provided $k_i \geq 2$ replicate measurements are available for each true $\mathbf{X}_i$. The key innovation there was that pseudo data are generated as random linear contrasts of the replicate measurements. Similar methods of generating pseudo errors, with the key change that $\zeta = -1$, as described in the preceding section, can be used to construct corrected scores from replicate measurements that automatically handle heteroscedastic measurement error. Here, we present the approach described in Stefanski, Novick and Devanarayan (2005) for the case that $\mathbf{X}_i$ is a scalar.

Assume that the error model is $\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j}$, where $\mathbf{U}_{i,j}$, $j = 1,\ldots,k_i$, $i = 1,\ldots,n$ are independent Normal$(0,\sigma^2_{u,i})$, independent of $\mathbf{X}_i$, $\mathbf{Z}_i$, and $\mathbf{Y}_i$, with all $\sigma^2_{u,i}$ unknown. Let $\overline{\mathbf{W}}_i$ and $\widehat{\sigma}^2_i$ denote the sample mean and sample variance of the i^{th} set of replicates, and define

$$\widetilde{\mathbf{W}}_{b,i} = \overline{\mathbf{W}}_i + \boldsymbol{\iota}(k_i-1)^{1/2}\widehat{\sigma}_i T_{b,i}, \tag{7.30}$$

where $T_{b,i} = V_{b,1}(V^2_{b,1} + \cdots + V^2_{b,k_i-1})^{-1/2}$ and the $V_{b,i}$ are generated as independent Normal$(0,1)$ independent of the data.

Stefanski, Novick and Devanarayan (2005) proved a result comparable to that in (7.28) and (7.29). They showed that if $f(\cdot)$ is a suitably smooth, integrable function, then

$$\widetilde{\mathbf{f}}(\mathbf{W}_i) = E\left[\mathrm{Re}\left\{\mathbf{f}(\widetilde{\mathbf{W}}_{b,i})\right\}|\mathbf{X}_i, \overline{\mathbf{W}}_i\right] \tag{7.31}$$

does not depend on $\mathbf{X}_i$ and is an unbiased estimator of $\mathbf{f}(\mathbf{X}_i)$, that is,

$$E\left\{\widetilde{\mathbf{f}}(\mathbf{W}_i)|\mathbf{X}_i\right\} = \mathbf{f}(\mathbf{X}_i).$$

This result is used to construct corrected scores for the replicate-data, unknown-heteroscedastic-error-variance model in the same manner that the result in (7.28) and (7.29) was used to construct them for the known-error-variance case in Section (7.4.1). The Monte Carlo corrected score has exactly the same form,

$$\Psi_{\text{MCCS,B}}(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\Theta) = B^{-1}\textstyle\sum_{b=1}^{B}\mathrm{Re}\{\Psi_{\text{True}}(\mathbf{Y}_i,\mathbf{Z}_i,\widetilde{\mathbf{W}}_{b,i},\Theta)\};$$

the only difference is that $\widetilde{\mathbf{W}}_{b,i}$ is defined in (7.30), as opposed to (7.26).

7.5 Computation and Asymptotic Approximations

Conditional-score and corrected-score estimators are M-estimators, and thus the usual numerical methods and asymptotic theory for M-estimators apply to both. We outline the computation and distribution approximations here for the case where Σ_{uu} is known, and when it is estimated from independent data.

7.5.1 Known Measurement Error Variance

Let $\Psi_*(Y,\mathbf{Z},\mathbf{W},\Theta,\Sigma_{uu})$ denote either a conditional or corrected score. Showing the dependence of the score on Σ_{uu} will be useful when we deal with the case of estimated measurement error variance. Suppose that $\widehat{\Theta}_*$ is a solution to the estimating equations,

$$\textstyle\sum_{i=1}^{n}\Psi_*(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\widehat{\Theta}_*,\Sigma_{uu}) = \mathbf{0}. \tag{7.32}$$

Then, generally, $n^{1/2}(\widehat{\Theta}_* - \Theta)$ is asymptotically Normal $\left\{\mathbf{0},\ A^{-1}B(A^{-1})^t\right\}$, where A and B are consistently estimated by

$$\begin{aligned}\widehat{A} &= n^{-1}\textstyle\sum_{i=1}^{n}\Psi_{*\Theta}(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\widehat{\Theta}_*,\Sigma_{uu})\\ \widehat{B} &= n^{-1}\textstyle\sum_{i=1}^{n}\Psi_*(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\widehat{\Theta}_*,\Sigma_{uu})\Psi_*^t(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\widehat{\Theta}_*,\Sigma_{uu}),\end{aligned}$$

with $\Psi_{*\Theta}(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta,\Sigma_{uu}) = (\partial/\partial\Theta^t)\Psi_*(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta,\Sigma_{uu})$.

The matrix $\widehat{A}$ also appears in the Newton–Raphson iteration for solving (7.32). Starting with an initial estimate $\widehat{\Theta}^{(0)}$, successive iterates are obtained via

$$\widehat{\Theta}_*^{(k+1)} = \widehat{\Theta}_*^{(k)} - \widehat{A}^{-1}n^{-1}\textstyle\sum_{i=1}^{n}\Psi_*\left(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\widehat{\Theta}_*^{(k)},\Sigma_{uu}\right).$$

Estimating equations for both conditional- and corrected-score estimates can have multiple solutions, and thus Newton–Raphson iteration can be sensitive to starting values. Although the naive estimate is often a reasonable initial estimate, it is sometimes necessary to use a measurement-error bias-corrected estimate such as regression calibration or SIMEX estimates; see Small, Wang, and Yang (2000) for a discussion of the multiple-root problem.

Different estimators of A and B are sometimes used for both the conditional- and corrected-score methods. For the conditional-score method, define

$$\begin{aligned}a\left\{\mathbf{Z},\Delta(\Theta,\Sigma_{uu}),\Theta,\Sigma_{uu}\right\} &= E\left\{\Psi_{*\Theta}\left(\cdot\right)|\mathbf{Z},\Delta(\Theta,\Sigma_{uu})\right\},\\ b\left\{\mathbf{Z},\Delta(\Theta,\Sigma_{uu}),\Theta,\Sigma_{uu}\right\} &= \mathrm{cov}\left\{\Psi_*\left(\cdot\right)|\mathbf{Z},\Delta(\Theta,\Sigma_{uu})\right\}.\end{aligned}$$

Then the alternate estimators are

$$\begin{aligned}\widehat{A}_2 &= n^{-1}\textstyle\sum_{i=1}^{n}a\left\{\mathbf{Z}_i,\Delta_i(\widehat{\Theta},\Sigma_{uu}),\widehat{\Theta},\Sigma_{uu}\right\},\\ \widehat{B}_2 &= n^{-1}\textstyle\sum_{i=1}^{n}b\left\{\mathbf{Z}_i,\Delta_i(\widehat{\Theta},\Sigma_{uu}),\widehat{\Theta},\Sigma_{uu}\right\}.\end{aligned} \tag{7.33}$$

Comparable estimates of A and B for corrected scores are substantially more involved; see Novick and Stefanski (2002).

7.5.2 Estimated Measurement Error Variance

When Σ_{uu} is unknown, additional data are required to estimate it consistently and the asymptotic variance-covariance matrix of the estimators is altered. Let $\Psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu})$ denote either a conditional score or a corrected score and $\widehat{\Theta}_*$ the estimator obtained by solving the estimating equations with Σ_{uu} replaced by $\widehat{\Sigma}_{uu}$. Define $\gamma = \mathrm{vech}(\Sigma_{uu})$, where *vech* is the vector-half of a symmetric matrix, that is, its distinct elements.

When an independent estimate of the error covariance matrix is available, the following method applies. Let $\widehat{\gamma}$ be an estimate of γ that is assumed to be independent of $\widehat{\Theta}_*$, with asymptotic covariance matrix $C_n(\Sigma_{uu})$. If we define

$$D_n(\Theta, \Sigma_{uu}) = \textstyle\sum_{i=1}^n \dfrac{\partial}{\partial \gamma^t} \Psi\left(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}\right),$$

then a consistent estimate of the covariance matrix of $\widehat{\Theta}$ is

$$n^{-1}\widehat{A}^{-1}\left(\widehat{\Theta}_*, \widehat{\Sigma}_{uu}\right)\left\{\widehat{B}\left(\widehat{\Theta}_*, \widehat{\Sigma}_{uu}\right) + D_n(\widehat{\Theta}_*, \widehat{\Sigma}_{uu}) C_n(\widehat{\Sigma}_{uu}) D_n^t(\widehat{\Theta}_*, \widehat{\Sigma}_{uu})\right\}\widehat{A}^{-t}\left(\widehat{\Theta}_*, \widehat{\Sigma}_{uu}\right),$$

where $\widehat{A}$ and $\widehat{B}$ are the matrices estimated in the construction of sandwich-formulae variance estimates. We have shown their dependence on Θ and Σ_{uu} to emphasize that they are to be computed at the estimated values $\widehat{\Theta}$ and $\widehat{\Sigma}_{uu}$.

Finally, a problem of considerable importance occurs when there are k_i independent replicate measurements of $\mathbf{X}_i$, $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$, $j = 1, \ldots, k_i$. A common situation is when $k_i = 1$ for most i, and the remainder of the data have a single replicate ($k_i = 2$). Constructing estimated standard errors for this problem has not been done previously, and the justification for our results is given in Appendix B.6. The necessary changes are as follows. In computing the estimates, in the previous definitions, replace Σ_{uu} with Σ_{uu}/k_i and $\mathbf{W}_i$ with $\overline{\mathbf{W}}_{i\cdot}$, the sample mean of the replicates. The estimate of Σ_{uu} is the usual components of variance estimator,

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t}{\sum_{i=1}^n (k_i - 1)}.$$

While the above variance estimator has a known asymptotic distribution (based on the Wishart distribution), it is easier in practice to use the sandwich estimator of its variance,

$$C_n(\widehat{\Sigma}_{uu}) = \frac{\sum_{i=1}^n d_i d_i^t}{\left\{\sum_{i=1}^n (k_i - 1)\right\}^2},$$

where

$$d_i = \mathrm{vech}\left\{\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t\right\} - (k_i - 1)\mathrm{vech}\left(\widehat{\Sigma}_{uu}\right).$$

7.6 Comparison of Conditional and Corrected Scores

Conditional-score and corrected-score methods are both functional methods, and thus they have in common the fact that neither one requires assumptions on the $\mathbf{X}_i$ for consistency to hold in general. However, they differ in other important ways, such as underlying assumptions and ease of computation.

In general, conditional scores are derived under specific assumptions about both the model for $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ and the error model for $\mathbf{W}$ given $\mathbf{X}$, whereas corrected scores assume only a correct estimating function if $\mathbf{X}$ were observed, and sufficient assumptions on the error model to enable unbiased estimation of the true-data estimating function. Consequently, when the assumptions underlying the conditional-score method are satisfied, it will usually be more efficient. Some conditional scores require numerical summation or integration. In principle, exact corrected scores also require integration; however, the Monte Carlo corrected score methods come with a simple, built-in solution to this computational problem when the required integrals are not analytically tractable, although the simplicity requires complex-variable computation (which not all will find simple).

A comparison of the two approaches has been made for Poisson regression, which is one of the few models where both methods apply. For this model, the corrected-score estimator is more convenient because the corrected score has a closed form expression, whereas the conditional-score estimator requires numerical summation; see (7.21). However, the conditional-score estimator is more efficient in some practical cases (Stefanski, 1989). We also note that, for the Poisson model, Kukush, Schneeweiss, and Wolf (2004) compared the corrected-score estimator to a structural estimator and conclude that the former, while more variable, is preferred on the basis of insensitivity to structural-model assumptions, except when the error variance is large.

The conditional-score method and certain extensions thereof have a theoretical advantage in terms of efficiency. For the canonical generalized linear models of Section 7.3, Stefanski and Carroll (1987) showed that any unbiased estimating equation for $(\beta_0, \beta_z^t, \beta_x^t)^t$ must be conditionally unbiased given $(\mathbf{Z}, \Delta)$, and from this they deduce that the asymptotically efficient estimating equations for structural models are based on score functions of the form

$$\{\mathbf{Y}_i - E(\mathbf{Y}_i \mid \mathbf{Z}_i, \Delta_i)\} \left\{ \begin{array}{c} 1 \\ \mathbf{Z}_i \\ E(\mathbf{X}_i \mid \mathbf{Z}_i, \Delta_i) \end{array} \right\}.$$

This result shows that, in general, none of the methods we have proposed previously is asymptotically efficient in structural models, except when $E(\mathbf{X} \mid \mathbf{Z}, \Delta)$ is linear in $(\mathbf{Z}, \Delta)$. This is the case in linear regression with $(\mathbf{Z}, \mathbf{X})$ marginally normally distributed, and in logistic regression when $(\mathbf{Z}, \mathbf{X})$ given $\mathbf{Y}$ is normally distributed, that is, the linear discriminant model. The problem of constructing fully efficient conditional-score estimators based on simultaneous estimation of $E(\mathbf{X}_i \mid \mathbf{Z}_i, \Delta_i)$ has been studied (Lindsay, 1985; Bickel and Ritov, 1987; van der Vaart, 1988), although the methods are generally too specialized or too difficult to implement in practice routinely.

Both methods have further extensions not mentioned previously. The conditional-score method is easily extended to the case that the model for $\mathbf{W}$ given $\mathbf{X}$ is a canonical generalized linear model with natural parameter $\mathbf{X}$. Buzas and Stefanski (1996a) described a simple extension of corrected-score methods to additive nonnormal error models when the true-data score function depends on $\mathbf{X}$ only through $\exp(\beta_x^t \mathbf{X})$ and the measurement error possesses a moment-generating function. Extensions to nonadditive models are also possible in some cases. For example, Li, Palta, and Shao, (2004) studied a corrected score for linear regression with a Poisson surrogate. Nakamura (1990) showed how to construct a corrected estimating equation for linear regression with multiplicative lognormal errors. He also suggested different methods of estimating standard errors.

7.7 Bibliographic Notes

Conditioning to remove nuisance parameters is a standard technique in statistics. The first systematic application of the technique to generalized linear measurement error models appeared in Stefanski and Carroll (1987), which was based on earlier work by Lindsay (1982). Related methods and approaches can be found in Liang and Tsou (1992), Liang and Zeger (1995), Hanfelt and Liang (1995), Hanfelt and Liang (1997), Rathouz and Liang (1999), and Hanfelt (2003).

The systematic development of corrected-score methods started with Nakamura (1990) and Stefanski (1989). Further developments and applications of this method can be found in Buzas and Stefanski (1996a), Augustin (2004), Kukush, Schneeweiss, and Wolf (2004), Li, Palta, and Shao (2004), and Song and Huang (2005). A related technique, but one that does not possess the same functional-modeling properties as corrected scores, is presented by Wang and Pepe (2000).

Both conditional-score and corrected-score estimating equations can have multiple solutions. In simpler models we have not found this to be a problem, but it can be with more complicated models. Small, Wang, and Yang (2000) discussed the multiple root problem and proposed some solutions.

CHAPTER 8

LIKELIHOOD AND QUASILIKELIHOOD

8.1 Introduction

This chapter describes the use of likelihood methods in nonlinear measurement error models. Prior to the first edition of this text, there were only a handful of applications of likelihood methods in our context. Since that time, largely inspired by the revolution in Bayesian computing, construction of likelihood methods with computation by either frequentist or Bayesian means has become fairly common.

There are a number of important differences between the likelihood methods in this chapter and the methods described in previous chapters:

- The previous methods are based on additive or multiplicative measurement error models, possibly after a transformation. Typically, few if any distributional assumptions about the distribution of $\mathbf{X}$ are required. Likelihood methods require stronger distributional assumptions, but they can be applied to more general problems, including those with discrete covariates subject to misclassification.

- The likelihood for a fully specified parametric model can be used to obtain likelihood ratio confidence intervals. In methods not based on likelihoods, inference is based on bootstrapping or on normal approximations. In highly nonlinear problems, likelihood-based confidence intervals are generally more reliable than those derived from normal approximations and less computationally intensive than bootstrapping.

- Whereas the previous methods require little more than the use of standard statistical packages, likelihood methods are often computationally more demanding.

- Robustness to modeling assumptions is a concern for both approaches, but is generally more difficult to understand with likelihood methods.

- When the simpler methods described previously are applicable, a likelihood analysis will generally buy one increased efficiency, that is, smaller standard errors, albeit at the cost of extra modeling assumptions, the old “no free lunch” phenomenon. Sometimes, the gains in

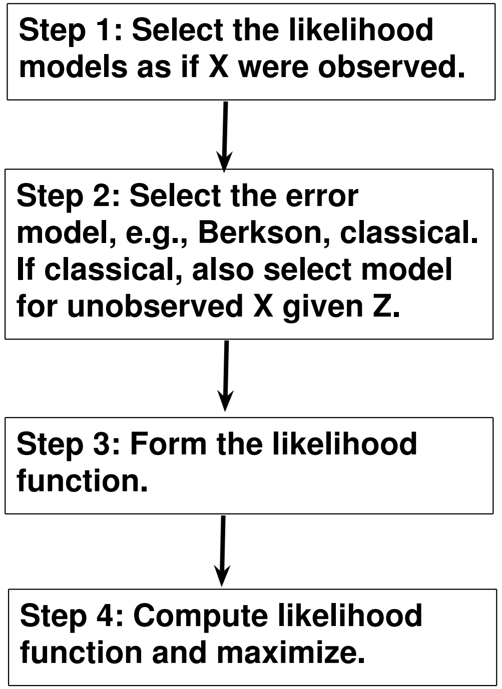

Step 1: Select the likelihood models as if X were observed.
Step 2: Select the error model, e.g., Berkson, classical. If classical, also select model for unobserved X given Z.
Step 3: Form the likelihood function.
Step 4: Compute likelihood function and maximize.

Figure 8.1 *Flowchart for the steps in a likelihood analysis.*

efficiency are very minor, as in typical logistic regression; see Stefanski and Carroll (1990b), who contrasted the maximum likelihood estimate and the conditional scores estimate of Chapter 7. They found that the conditional score estimates are usually fairly efficient relative to the maximum likelihood estimate, unless the measurement error is “large” or the logistic coefficient is “large,” where the definition of *large* is somewhat vague. One should be aware, though, that their calculations indicate that situations exist where *properly parameterized* maximum likelihood estimates are considerably more efficient than estimates derived from functional modeling considerations.

Figure 8.1 illustrates the steps in doing a likelihood analysis of a measurement error model. These steps are as follows:

- **Step 1**: To perform a likelihood analysis, one must specify a parametric model for every component the data. Any likelihood analysis begins with the model one would use if $\mathbf{X}$ were observable.
- **Step 2**: The next crucial decision is the error model that is to be chosen. This could be a classical error model, a Berkson model, a combination of the two, etc. If one has classical components in the measurement error model, then typically one also needs to specify a distribution for the unobserved $\mathbf{X}$ given the observable covariates $\mathbf{Z}$; see Section 2.2.3. Much of the grief of a likelihood analysis revolves around this step.

- **Step 3**: The likelihood function is constructed using the building blocks obtained in previous steps.

- **Step 4**: Now one has to do the sometimes hard slogging to compute the likelihood function, obtain parameter estimates, do inferences, etc. Because $\mathbf{X}$ is latent, that is, unobservable, this step can be difficult or time-consuming, because one must integrate out the possibly high dimensional latent variable.

We organize our discussion of likelihood methods around these four steps. Except where noted, we assume nondifferential measurement error (Section 2.5). For a review of maximum likelihood methods in general, see Appendix A.5.

Fully specified likelihood problems, including problems where $\mathbf{X}$ is not observable or is observable for only a subset of the data, are discussed in Sections 8.2 and 8.3. The use of likelihood ideas in quasilikelihood and variance function models (QVF) (Section A.7) is covered in Section 8.8.

8.1.1 *Step 1: The Likelihood If* $\mathbf{X}$ *Were Observable*

Likelihood analysis starts with an “outcome model” for the distribution of the response given the true predictors. The likelihood (density or mass) function of $\mathbf{Y}$ given $(\mathbf{Z},\mathbf{X})$ will be called $f_{Y|Z,X}(y|z,x,\mathcal{B})$ here, and interest lies in estimating $\mathcal{B}$.

The form of the likelihood function can generally be specified by reference to any standard statistics text. For example, if $\mathbf{Y}$ is normally distributed with mean $\beta_0+\beta_x^t\mathbf{X}+\beta_z^t\mathbf{Z}$ and variance σ^2, then $\mathcal{B}=(\beta_0,\beta_x,\beta_z,\sigma^2)$ and

$$f_{Y|Z,X}(y|z,x,\mathcal{B})=\sigma^{-1}\phi\left[\{(y-(\beta_0+\beta_x^tx+\beta_z^tz)\}/\sigma\right],$$

where $\phi(v)$ is the standard normal density function. If $\mathbf{Y}$ follows a logistic regression model with mean $H(\beta_0+\beta_x^t\mathbf{X}+\beta_z^t\mathbf{Z})$, then $\mathcal{B}=(\beta_0,\beta_x,\beta_z)$ and

$$\begin{aligned} f_{Y|Z,X}(y|z,x,\mathcal{B}) &= H^y\left(\beta_0+\beta_x^tx+\beta_z^tz\right) \\ &\quad\times\left\{1-H\left(\beta_0+\beta_x^tx+\beta_z^tz\right)\right\}^{1-y}. \end{aligned}$$

8.1.2 A General Concern: Identifiable Models

The concept of identifiability means that if one actually had an infinite number of observed data values, then one would know the parameters exactly, that is, they are identified. When a problem is not identifiable, it means that a key piece of information is unavailable. For example, in linear regression with $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$ all normally distributed, as described in Section 3.2.1, the parameters are not identifiable because a key piece of information is absent, namely, the measurement error variance. For this reason, replication data is needed to help estimate the measurement error variance.

In nonlinear measurement error models, sometimes the parameters are identifiable without any extra information other than measures of $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$, that is, without validation or replications. Brown and Mariano (1993) discuss this issue, considering both likelihood and quasilikelihood techniques.

A word of warning: One should not be overly impressed by all claims of identifiability. Many problems of practical importance actually are identifiable, but only barely so, and estimation without additional data is not practical. For instance, in linear regression it is known that the regression parameters can be identified without validation or replication as long as $\mathbf{X}$ is *not* normally distributed (Fuller, 1987, pp. 72–73). However, this means that the parameter estimates will be very unstable if $\mathbf{X}$ is at all close to being normally distributed. In binary regression with a normally distributed calibration, it is known that the probit model is not identified (Carroll, Spiegelman, Lan, et al., 1984) but that the logistic model is (Küchenhoff, 1990). The difference between these two models is so slight (Figure 4.9) that there is really no useful information about the parameters without some additional validation or replication data.

However, lest we leave you with a glum picture, there are indeed cases that the nonlinearity in the model helps make identifiability practical; for example, Rudemo, Ruppert, and Streibig, et al. (1989) describe a highly nonlinear model that is both identifiable and informative; see Section 4.7.3.

This discussion highlights the Sherlock Holmes phenomenon: Data are good, but more and different types of data are better.

8.2 Steps 2 and 3: Constructing Likelihoods

Having specified the likelihood function as if $\mathbf{X}$ were observable, we now turn to constructing the form of the likelihood function. This consists of 1 or 2 steps, depending on whether the error model is Berkson or classical. In this section, we allow for general error models and for the possibility that a second measure $\mathbf{T}$ is available. A likelihood analysis starts with determination of the joint distribution of $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{T}$ given $\mathbf{Z}$, as these are the observed variates.

To understand what is going on, we first describe the discrete case when there is neither a covariate $\mathbf{Z}$ measured without error nor a second measure $\mathbf{T}$. We then describe in detail the classical and Berkson models in turn.

8.2.1 The Discrete Case

First consider a simple problem wherein $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{X}$ are discrete random variables; no second measure $\mathbf{T}$ is observed; and there are no other covariates $\mathbf{Z}$. From basic probability, we know that

$$\begin{aligned}\mathrm{pr}(\mathbf{Y}=y,\mathbf{W}=w) &= \sum_x \mathrm{pr}(\mathbf{Y}=y,\mathbf{W}=w,\mathbf{X}=x)\\ &= \sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{W}=w,\mathbf{X}=x)\mathrm{pr}(\mathbf{W}=w,\mathbf{X}=x). \qquad (8.1)\end{aligned}$$

When $\mathbf{W}$ is a surrogate (nondifferential measurement error, see Section 2.5), it provides no additional information about $\mathbf{Y}$ when $\mathbf{X}$ is known, so (8.1) is

$$\begin{aligned}&\mathrm{pr}\,(\mathbf{Y}=y,\mathbf{W}=w)\\ &= \sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x,\mathcal{B})\mathrm{pr}(\mathbf{W}=w,\mathbf{X}=x), \qquad (8.2)\end{aligned}$$

where, to achieve a parsimonious model, we have (1) replaced $\mathrm{pr}(\mathbf{Y}=y|\mathbf{W}=w,\mathbf{X}=x)$ by $\mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x)$ and (2) indicated that typically one would use a parametric model $\mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x,\mathcal{B})$ for the latter. Thus, in addition to the underlying model, we must specify a model for the joint distribution of $\mathbf{W}$ and $\mathbf{X}$. How we do this depends on the model relating $\mathbf{W}$ and $\mathbf{X}$.

8.2.1.1 Classical Case, Discrete Data

For example, in the context of classical measurement error and, for simplicity, assuming no $\mathbf{Z}$, we would specify a model for $\mathbf{W}$ given $\mathbf{X}$, and then a model for $\mathbf{X}$. In other words,

$$\sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x,\mathcal{B})\mathrm{pr}(\mathbf{W}=w|\mathbf{X}=x)\mathrm{pr}(\mathbf{X}=x). \qquad (8.3)$$

Equation (8.3) has three components: (a) the underlying "outcome model" of primary interest; (b) the error model for $\mathbf{W}$ given the true covariates; and (c) the distribution of the true covariates, sometimes called the *exposure model* in epidemiology. Both (a) and (b) are expected; almost all the methods we have discussed so far require an underlying model and an error model. However, (c) is unexpected, in fact a bit disconcerting, because it requires a model for the distribution of the unobservable **X**. It is (c) that causes almost all the practical problems of implementation and model selection with maximum likelihood methods.

8.2.1.2 Berkson Case, Discrete Data

The likelihood of the observed data is (8.2) because **W** is a surrogate. At this point, however, the analysis changes. When the Berkson model holds, we write

$$\begin{aligned} &\mathrm{pr}(\mathbf{Y}=y,\mathbf{W}=w) \qquad (8.4)\\ &=\sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x,\mathcal{B})\mathrm{pr}(\mathbf{X}=x|\mathbf{W}=w)\mathrm{pr}(\mathbf{W}=w). \end{aligned}$$

The third component of (8.4) is the distribution of **W** and conveys no information about the critical parameter $\mathcal{B}$. Thus, we will divide both sides of (8.4) by pr($\mathbf{W}=w$) to get likelihoods conditional on **W**, namely,

$$\mathrm{pr}(\mathbf{Y}=y|\mathbf{W}=w)=\sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x,\mathcal{B})\mathrm{pr}(\mathbf{X}=x|\mathbf{W}=w). \qquad (8.5)$$

8.2.2 Likelihood Construction for General Error Models

We now describe the form of the likelihood function for general error models.

When there are covariates **Z** measured without error, or when there are second measures **T**, (8.3) changes in two ways. The second measure is appended to **W**, and all probabilities are conditional on **Z**. Thus, (8.3) is generalized to

$$\begin{aligned} &\sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{Z}=z,\mathbf{X}=x,\mathcal{B})\\ &\quad\times\mathrm{pr}(\mathbf{W}=w|\mathbf{Z}=z,\mathbf{X}=x)\mathrm{pr}(\mathbf{X}=x|\mathbf{Z}=z). \end{aligned}$$

In general, in problems where **X** is not observed but there is a natural error model, then in addition to specifying the underlying model and the error model, we must hypothesize a distribution for **X** given **Z**. In summary:

- The error model has a density or mass function which we will denote by $f_{W,T|Z,X}(w,t|z,x,\widetilde{\alpha}_1)$.
  - In many applications, the error model does not depend on z. For example, in the classical additive measurement error model (1.1) with normally distributed measurement error, σ_u^2 is the only component of $\widetilde{\alpha}_1$, there is no second measure **T**, and the error model density is $\sigma_u^{-1}\phi\{(w-x)/\sigma_u\}$, where $\phi(\cdot)$ is the standard normal density function.
  - If **W** is binary, a natural error model is the logistic where, for example, $\widetilde{\alpha}_1=(\alpha_{11},\alpha_{12},\alpha_{13})$ and $\mathrm{pr}(\mathbf{W}=1|\mathbf{X}=x,\mathbf{Z}=z)=H(\alpha_{11}+\alpha_{12}x+\alpha_{13}^t z)$.
  - Multiplicative models occur when $\mathbf{W}=\mathbf{XU}$, where typically **U** has a lognormal or gamma distribution with $E(\mathbf{U})=1$.
- We use $f_{X|Z}(x|z,\widetilde{\alpha}_2)$ to denote the density or mass function of **X** given **Z**. As might be expected, the latent nature of **X** makes specifying this distribution a matter of art. Nonetheless, there are some general guidelines:
  - When **X** is univariate, generalized linear models (Section A.8) are natural and useful. For example, one might hypothesize that **X** is normally distributed in a linear regression on **Z**, or that it follows a gamma or lognormal loglinear model in **Z**. If **X** were binary, a linear logistic regression on **Z** would be a natural candidate.
  - Some model robustness can be gained by specifying flexible distributions for **X** given **Z**. One class is to suppose that, depending on the context, **X** or log(**X**) follows a linear regression in **Z**, but that the regression errors have a mixture of normals density. Mixtures of normals can be difficult to work with, and an obvious alternative is to use the so-called seminonparametric family (SNP) of Davidian and Gallant (1993, p. 478): see Zhang and Davidian (2001) for a computationally convenient form of this approach. Davidian and Gallant's mixture model generalizes easily to the case that all components of **X** are continuous. We point out that Bayesians often use Dirichlet process mixtures to achieve a seminonparametric modeling approach.
  - For mixtures of discrete and continuous variables, the models of Zhao, Prentice, and Self (1992) hold considerable promise. Otherwise, one can proceed on a case-by-case basis. For example, one can split **X** into discrete and continuous components. The distribution of the continuous component given the discrete components might be modeled by multivariate normal linear regression, while that of the discrete component given **Z** could be any multivariate discrete random variable. We would be remiss in not pointing out that multivariate discrete models can be difficult to specify.

Having hypothesized the various models, the likelihood that $(\mathbf{Y}=y,\mathbf{W}=w,T=t)$ given that $\mathbf{Z}=z$ is then

$$f_{Y,W,T|Z}(y,w,t|z,\mathcal{B},\widetilde{\alpha}_1,\widetilde{\alpha}_2)$$

$$= \int f_{Y|Z,X,W,T}(y|z,x,w,t,\mathcal{B}) f_{W,T|Z,X}(w,t|z,x,\widetilde{\alpha}_1) \times f_{X|Z}(x|z,\widetilde{\alpha}_2) d\mu(x) \qquad (8.6)$$

$$= \int f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{W,T|Z,X}(w,t|z,x,\widetilde{\alpha}_1) \times f_{X|Z}(x|z,\widetilde{\alpha}_2) d\mu(x). \qquad (8.7)$$

The notation $d\mu(x)$ indicates that the integrals are sums if $\mathbf{X}$ is discrete and integrals if $\mathbf{X}$ is continuous. The assumption of nondifferential measurement error (Section 2.5), which is equivalent to assuming that $\mathbf{W}$ and $\mathbf{T}$ are surrogates for $\mathbf{X}$, was used in going from (8.6) to (8.7), and will be used without mention elsewhere in this chapter. The likelihood for the problem is just the product over the sample of the terms (8.7) evaluated at the data.

Of interest in applications is the density function of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$, which is (8.7) divided by its integral or, in the discrete case, sum over y. This density is an important tool in the process of model criticism, because it allows us to compute such diagnostics as the conditional mean and variance of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$, so that standard model verification techniques from regression analysis can be used.

8.2.3 The Berkson Model

In the Berkson model, a univariate $\mathbf{X}$ is not observed, but it is related to a univariate $\mathbf{W}$ by $\mathbf{X} = \mathbf{W} + \mathbf{U}$, perhaps after a transformation. There are no other covariates. Usually, $\mathbf{U}$ is taken to be independent of $\mathbf{W}$ and normally distributed with mean zero and variance σ_u^2, but more complex models are possible. For example, in the bioassay data of Chapter 4, the variance might be modeled as $\sigma_u^2 \mathbf{W}^{2\theta}$.

The additive model is not a requirement. In some cases, it might be more reasonable to assume that $\mathbf{X} = \mathbf{WU}$, where $\mathbf{U}$ has mean 1.0 and is either lognormal or gamma.

The Berkson additive model has an unusual feature in the linear regression problem. Suppose the regression has true mean $\beta_0 + \mathbf{X}\beta_x + \mathbf{Z}^t\beta_z$ and residual variance σ_ϵ^2. Then in the pure Berkson model, we replace $\mathbf{X}$ with $\mathbf{W} + \mathbf{U}$, with the following consequences.

- The good news is the following: The observed data have regression mean $\beta_0 + \mathbf{W}^t\beta_x + \mathbf{Z}^t\beta_z$, the observed data have the correct regression line, so that the naive analysis yields valid estimates of the regression line.

- The bad news is that the observed data have residual variance $\sigma_\epsilon^2 + \beta_x^2\sigma_u^2$. This means that the observed data cannot disentangle the true residual variance σ_ϵ^2 from the Berkson error variance σ_u^2, so that neither is identified; see Section 8.1.2.

It can also be shown that the additive Berkson model with homoscedastic errors leads to consistent estimates of nonintercept parameters in loglinear models and often to nearly consistent estimates in logistic regression. In the latter case, the exceptions occur with severe measurement error and a strong predictive effect; see Burr (1988).

In general problems, we must specify the conditional density or mass function of $\mathbf{X}$ given $\mathbf{W}$, which we denote by $f_{X|W}(x|w,\widetilde{\gamma})$. In the usual Berkson model, $\widetilde{\gamma}$ is σ_u^2, and the density is $\sigma_u^{-1}\phi\{(x-w)/\sigma_u\}$. In a Berkson model where the variance is proportional to $\mathbf{W}^{2\theta}$, the density is $(w^\theta\sigma_u)^{-1}\phi\left\{(x-w)/(w^\theta\sigma_u)\right\}$. The likelihood function then becomes

$$f_{Y|Z,W}(y|z,w,\mathcal{B},\widetilde{\gamma}) = \int f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{X|W}(x|w,\widetilde{\gamma}) d\mu(x). \qquad (8.8)$$

The likelihood for the problem is the product over the sample of the terms (8.8) evaluated at the data.

As a practical matter, there is rarely a direct “second measure” in the Berkson additive or multiplicative models. This means that the parameters in the Berkson model can be estimated only through the likelihood (8.8). In some cases, such as homoscedastic linear regression described above, not all of the parameters can be identified (estimated). For nonlinear models, identification usually is possible.

In classical generalized linear models, a likelihood analysis of a homoscedastic, additive Berkson model can be shown to be equivalent to a random coefficients analysis with random intercept for each study participant.

8.2.4 Error Model Choice

Modeling always has options. For example, there is nothing illegal in simply specifying a model for $\mathbf{X}$ given $\mathbf{W}$, as in equation (8.8), or a model for $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$, in which case $\mathbf{Z}$ is added to the error distribution in (8.8). One could then simply ignore the detailed aspects of the measurement error models, including such inconvenient things as whether the measurement error is additive and homoscedastic, etc. One can even specify reasonably flexible models for such distributions, for example, by using the Davidian and Gallant models. Mallick and Gelfand (1996) do just this in a Bayesian context.

This purely empirical approach has the attraction of its empiricism, but it almost forces one to write down very general models for $\mathbf{X}$ given $(\mathbf{W}, \mathbf{Z})$ in order to achieve sensible answers. In addition, there is the potential for a loss of information, because the real likelihood is the likelihood of $\mathbf{Y}$ *and* $\mathbf{W}$ given $\mathbf{Z}$, not simply the likelihood of $\mathbf{Y}$ given $(\mathbf{W},\mathbf{Z})$. This may seem like a minor difference, but we suspect that the difference is not minor. There is little to no literature on whether such an approach can yield sensible answers when additive–multiplicative error models hold.

8.3 Step 4: Numerical Computation of Likelihoods

The overall likelihood based on a sample of size n is the product over the sample of (8.6) when $\mathbf{X}$ is unobserved or the product over the sample of (8.8) in the Berkson model. Typically, one maximizes the logarithm of the overall likelihood in the unknown parameters. There are two ways one can maximize the likelihood function. The most direct is to compute the likelihood function itself, and then use numerical optimization techniques to maximize the likelihood. Below we provide a few details about computing the likelihood function. The second general approach is to view the problem as a missing-data problem, and then use missing-data techniques; see for example Little and Rubin (2002), Tanner (1993), and Geyer and Thompson (1992).

Computing the likelihoods (8.7) and (8.8) analytically is easy if $\mathbf{X}$ is discrete, as the conditional expectations are simply sums of terms. Likelihoods in which $\mathbf{X}$ has some continuous components can be computed using a number of different approaches. In some problems, the loglikelihood can be computed or very well approximated analytically, for example, linear, probit, and logistic regression with $(\mathbf{W},\mathbf{X})$ normally distributed; see Section B.7.2. In most problems that we have encountered, $\mathbf{X}$ is a scalar or a 2×1 vector. In these cases, standard numerical methods, such as Gaussian quadrature, can be applied, although they are not always very good. When sufficient computing resources are available, the likelihood can be computed using Monte Carlo techniques (Section B.7.1). One of the advantages of a Bayesian analysis by simulation methods is that $\mathbf{X}$ can be integrated out as part of the processing of sampling from the posterior; see Chapter 9.

8.4 Cervical Cancer and Herpes

In the cervical cancer example of Section 1.6.10, $(\mathbf{W},\mathbf{Y},\mathbf{X})$ are all binary, $\mathbf{W}$ is a possibly misclassified version of $\mathbf{X}$, and there is no variable $\mathbf{Z}$. In principle, MC-SIMEX (Section 5.6.2) could be used, but maximum likelihood is so simple with binary $\mathbf{X}$ that there seems little reason to use MC-SIMEX here. It would obviously be of interest to compare maximum likelihood approaches to misclassification of $\mathbf{X}$ with MC-SIMEX.

As mentioned in the introduction to this chapter, maximum likelihood is a particularly useful technique for treating the problem of misclassification of a discrete covariate. One reason for this is that numerical integration of $\mathbf{X}$ out of the joint density of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{W}$ is replaced by an easy summation. Another reason is that, for a discrete covariate, one can use a structural model without the need to make strong structural assumptions, since, for example, a binary random variable must have a Bernoulli distribution.

The log odds-ratio β is defined by

$$\begin{aligned}\exp(\beta) &= \frac{\mathrm{pr}(\mathbf{X}=1|\mathbf{Y}=1)/\mathrm{pr}(\mathbf{X}=1|\mathbf{Y}=0)}{\mathrm{pr}(\mathbf{X}=0|\mathbf{Y}=1)/\mathrm{pr}(\mathbf{X}=0|\mathbf{Y}=0)} && (8.9)\\ &= \frac{\mathrm{pr}(\mathbf{Y}=1|\mathbf{X}=1)/\mathrm{pr}(\mathbf{Y}=0|\mathbf{X}=1)}{\mathrm{pr}(\mathbf{Y}=1|\mathbf{X}=0)/\mathrm{pr}(\mathbf{Y}=0|\mathbf{X}=0)}. && (8.10)\end{aligned}$$

Here (8.9) is the odds-ratio for a retrospective study where the disease status $\mathbf{Y}$ is fixed, while (8.10) is the log-odds for a prospective study where the risk factor $\mathbf{X}$ is fixed. The equality of the two odds-ratios allows one to parameterize either prospectively (in terms of $\mathbf{Y}$ given $\mathbf{X}$ and $\mathbf{W}$) or retrospectively (in terms of $\mathbf{X}$ and $\mathbf{W}$ given $\mathbf{Y}$) and is the theoretical basis for case-control studies.

This problem is particularly easy to parameterize retrospectively by specifying the distributions of $\mathbf{X}$ given $\mathbf{Y}$, and $\mathbf{W}$ given $(\mathbf{X},\mathbf{Y})$. With differential measurement error, the six free parameters are $\alpha_{xd} = \mathrm{Pr}(\mathbf{W} = 1|\mathbf{X}=x,\mathbf{Y}=d)$ and $\gamma_d = \mathrm{Pr}(\mathbf{X}=1|\mathbf{Y}=d)$, $x=0,1$ and $d=0,1$.

For the validation data where $\mathbf{X}$ is observed, the likelihood is

$$\prod_{i=1}^{n_v}\prod_{\mathbf{y}=0}^{1}\prod_{\mathbf{x}=0}^{1}\prod_{\mathbf{w}=0}^{1}\mathrm{pr}(\mathbf{W}_i=\mathbf{w},\mathbf{X}_i=\mathbf{x}|\mathbf{Y}_i=\mathbf{y})^{I(\mathbf{W}_i=\mathbf{w},\mathbf{X}_i=\mathbf{x},\mathbf{Y}_i=\mathbf{y})}, \quad (8.11)$$

where n_v is the size of the validation data set. For the nonvalidation data we integrate out $\mathbf{X}$ by a simple summation $\mathrm{pr}(\mathbf{W}=w|\mathbf{Y}=\mathbf{y}) = \mathrm{pr}(\mathbf{W}=\mathbf{w},\mathbf{X}=0|\mathbf{Y}=\mathbf{y}) + \mathrm{pr}(\mathbf{W}=\mathbf{w},\mathbf{X}=1|\mathbf{Y}=\mathbf{y})$, and then the likelihood is

$$\prod_{i=1}^{n_{nv}}\prod_{\mathbf{y}=0}^{1}\prod_{\mathbf{w}=0}^{1}\mathrm{pr}(\mathbf{W}_i=\mathbf{w}|\mathbf{Y}_i=\mathbf{y})^{I(\mathbf{W}_i=\mathbf{w},\mathbf{Y}_i=\mathbf{y})}, \quad (8.12)$$

where $n_{nv} = n - n_v$ is the size of the validation data set. The likelihood for the data set itself is the product of (8.11) and (8.12), with all probabilities expressed in terms of the α's and γ's.

A maximum likelihood analysis yielded $\widehat{\beta} = 0.609$ (std. error = 0.350). For nondifferential misclassification, the analysis simplifies in that $\alpha_{x0} = \alpha_{x1} = \alpha_x$, and then $\widehat{\beta} = 0.958$ (std. error = 0.237).

The noticeable difference in $\widehat{\beta}$ between assuming differential and nondifferential misclassification suggests that, in this example, misclassification is differential. In Section 9.9, this issue is further explored by comparing estimates of $\alpha_{0d} = \Pr(\mathbf{W} = 1|\mathbf{X} = x, \mathbf{Y} = d)$ for $d = 1$ and $d = 0$.

8.5 Framingham Data

In this section we describe an example that has classical error structure.

The Framingham heart study was described in Section 1.6.6. Here $\mathbf{X}$, the transformed long-term systolic blood pressure, is not observable, and the likelihoods of Section 8.2.2 are appropriate. The sample size is $n =$ 1,615. As before, $\mathbf{Z}$ includes age, smoking status, and serum cholesterol. Transformed systolic blood pressure (SBP) is log(SBP−50).

At Exam #2, the mean and standard deviation of transformed systolic blood pressure are 4.374 and .226, respectively, while the corresponding figures at Exam #3 are 4.355 and .229. The difference between measurements at Exam #2 and Exam #3 has mean 0.019 and standard deviation .159, indicating a statistically significant difference in means due largely to the sample size ($n = 1,615$). However, the following analysis will allow for differences in the means. The standard deviations are sufficiently similar that we will assume that the two exams have the same variability.

We write $\mathbf{W}$ and $\mathbf{T}$ for the transformed SBP at Exams 3 and 2, respectively. Since Exam #2 is not a true replicate, we are treating it as a second measure, differing from Exam #3 only in the mean. Thus, $\mathbf{W} = \mathbf{X} + \mathbf{U}$ and $\mathbf{T} = \alpha_{11} + \mathbf{X} + \mathbf{V}$, where $\mathbf{U}$ and $\mathbf{V}$ are independent with common measurement error variance σ_u^2, and α_{11} represents the (small) difference between the two exams.

There is justification for the assumption that transformed systolic blood pressure can be modeled reasonably by an additive model with normally distributed, homoscedastic measurement error. We use the techniques of Section 1.7. The q-q plot of the differences of transformed systolic blood pressure in the two exams is reasonable, although not perfect, indicating approximate normality of the measurement errors. The regression fits of the intraindividual standard deviation versus the mean are plotted in the original and transformed scale in Figure 8.2. The trend in the former suggests heteroscedastic measurement errors, while the lack of pattern in the latter suggests the transformation is a reasonable one.

Since the transformed systolic blood pressures are themselves approximately normally distributed, we will also assume that $\mathbf{X}$ given $\mathbf{Z}$ is normally distributed with mean $\alpha_{21}^t \mathbf{Z}$ and variance σ_x^2.

Using the probit approximation to the logistic (Section 4.8.2), it is possible to compute the likelihood (8.7) analytically; see section B.7.2. We used this analytical calculation, rather than numerical integration. When using all the data, the likelihood estimate for systolic blood pressure had a logistic coefficient of 2.013 with an (information) estimated standard error of 0.496, which is essentially the same as the regression calibration analysis; compare with Table 5.1.

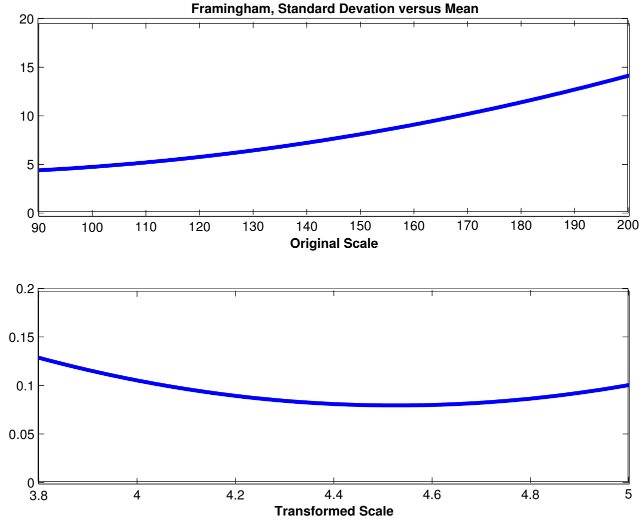


Figure 8.2 *Framingham systolic blood pressure data. Plot of cubic regression fits to the regression of intraindividual standard deviation against the mean. Top figure is in the original scale; bottom is for the transformation log(SBP−50). The noticeable trend in the top figure suggests that the measurement errors are heteroscedastic in that scale.*

8.6 Nevada Test Site Reanalysis

This section describes a problem that has a combination of Berkson and classical measurement errors, and it is one in which the errors are multiplicative rather than additive.

In Section 1.8.2, we described a simulation study in a model where part of the measurement error was Berkson and part was classical. The excess relative risk model (1.4) was used, and here for convenience we redisplay the model:

$$\mathrm{pr}(\mathbf{Y} = 1|\mathbf{X}, \mathbf{Z}) = H\left\{\beta_0 + \beta_z \mathbf{Z} + \log(1 + \beta_x \mathbf{X})\right\}. \tag{8.13}$$

The parameter β_x is the excess relative risk parameter. The mixture of classical and Berkson error models is given in equations (1.5) and (1.6). Again, for convenience, we redisplay this multiplicative measurement error model:

$$\log(\mathbf{X}) = \log(\mathcal{L}) + \mathbf{U}_b, \tag{8.14}$$

$$\log(\mathbf{W}) = \log(\mathcal{L}) + \mathbf{U}_c, \tag{8.15}$$

where $\mathbf{U}_b$ denotes Berkson-type error, and $\mathbf{U}_c$ denotes classical-type error. The standard classical measurement error model (1.1) is obtained by setting $\mathbf{U}_b = 0$. The Berkson model (1.2) is obtained by setting $\mathbf{U}_c = 0$.

In this section, we analyze the Nevada Test Site data, with outcome variable thyroiditis. The original data and their analyses were described by Stevens, Till, Thomas, et al. (1992); Kerber et al. (1993); and Simon, Till, Lloyd, et al. (1995). We use instead a revised version of these data that have corrected dosimetry as well as corrected health evaluations (Lyon, Alder, Stone, et al., 2006). In the risk model (8.13), the predictors $\mathbf{Z}$ consisted of gender and age at exposure, while $\mathbf{X}$ is the true dose. The data file gives an estimate for each individual of the total error variance in the log scale, but does not separate out the Berkson and classical uncertainties. In this illustration, we assumed that 40% of the uncertainty was classical, reflecting the important components due to dietary measurement error.

In these data, Owen Hoffman suggested the use of strata, because it is known that the doses received by individuals vary greatly, depending on where they were located. Thus, we fit models (8.14) and (8.15) in five different strata, namely (a) Washington County, (b) Lincoln County, (c) Graham County, (d) all others in Utah, and (e) all others. In these models, $\log(\mathcal{L})$ was normally distributed in a regression on gender and age, with the regression coefficients and the variance about the mean depending on the strata.

We performed three analyses of these data:

- The first model assumed that all uncertainty was Berkson and employed regression calibration. Specifically, since $\log(\mathbf{X}) = \log(\mathbf{W}) + \mathbf{U}_b$, with $\mathbf{U}_b$ normally distributed with mean zero and known variance σ^2_{ub} depending on the individual, $E(\mathbf{X}|\mathbf{W}) = \mathbf{W}\exp(\sigma^2_{ub}/2)$: This latter value was used in place of true dose.
- The second model assumes that 40% of the uncertainty is classical. We again implemented regression calibration; see below for the details.
- The third analysis was a maximum likelihood analysis; see below for details.

The results are described in Figure 8.3. The Berkson analysis yields an excess relative risk estimate $\widehat{\beta}_x = 5.3$, regression calibration $\widehat{\beta}_x = 8.7$, and maximum likelihood $\widehat{\beta}_x = 9.9$. The p-value using a likelihood ratio test for the hypothesis of no effect due to radiation is $< 10^{-7}$. Note how acknowledging the classical measurement error greatly increases the estimated excess relative risk, by a factor nearly of two. Of potential scientific interest is that the upper ends of the confidence intervals shown in Figure 8.3 change from 11.1 for the pure Berkson analysis to 18.8 for the mixture of Berkson and classical analysis, indicating the potential for a much stronger dose effect.

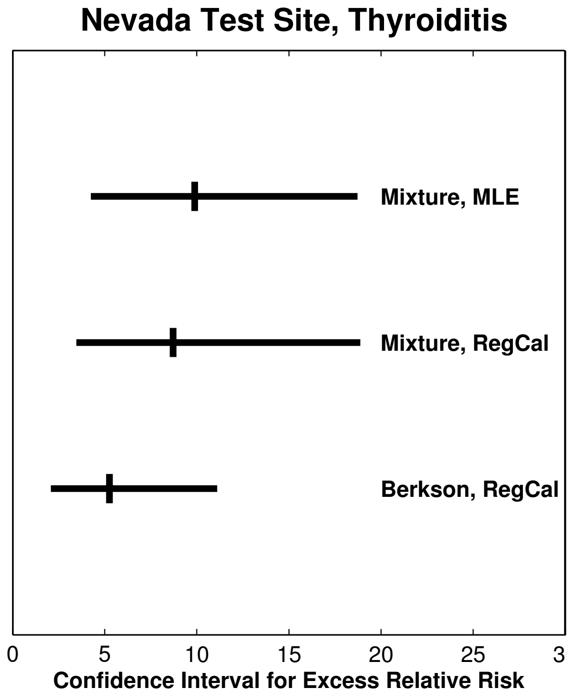


Figure 8.3 *Nevada Test Site analysis of the excess relative risk for thyroiditis. Estimates are the vertical lines, while the horizontal lines are 95% confidence intervals. Included are a pure Berkson analysis and two analyses that assume a mixture of classical and independent Berkson measurement errors, with 40% of the measurement error variance being classical. Note how the classical component greatly increases the excess relative risk estimate.*

8.6.1 Regression Calibration Implementation

Here is how we implemented regression calibration for the Nevada Test Site thyroiditis example. Let $\sigma^2_{i,\mathrm{tot}}$ be the variance of the uncertainty in true dose for an individual. Then the Berkson error variance for that individual is $\sigma^2_{i,\text{ub}} = 0.6 \times \sigma^2_{i,\text{tot}}$, while the classical error variance for that individual is $\sigma^2_{i,\text{uc}} = 0.4 \times \sigma^2_{i,\text{tot}}$. We assumed that, for an individual i who falls into stratum s, $\log(\mathcal{L}_i)$ was normally distributed with mean $\alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s}$ and variance σ^2_{Ls}. This means that $\log(\mathbf{X}_i)$ and $\log(\mathbf{W}_i)$ are jointly normally distributed with common mean $\alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s}$, variances $\sigma^2_{xi} = \sigma^2_{Ls} + \sigma^2_{i,\text{ub}}$ and $\sigma^2_{wi} = \sigma^2_{Ls} + \sigma^2_{i,\text{uc}}$, respectively, and covariance σ^2_{Ls}. Define $\rho_i = \sigma^2_{Ls}/(\sigma_{xi}\sigma_{wi})$. By simple algebraic calculations, this means that $\log(X_i)$ given $(\mathbf{W}_i, \mathbf{Z}_i)$ is normally distributed with mean $(\alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s})(1 - \sigma^2_{Ls}/\sigma^2_{wi}) + (\sigma^2_{Ls}/\sigma^2_{wi})\log(\mathbf{W}_i)$ and variance $\sigma^2_{xi}(1 - \rho_i^2)$. As in Section 4.5, this is a multiplicative measurement error with a lognormal structure, and hence it follows that

$$E(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i) = \exp\{(\alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s})(1 - \sigma^2_{Ls}/\sigma^2_{wi}) + \sigma^2_{xi}(1 - \rho^2)/2\}.$$

It remains to estimate α_{0s}, α_{1s}, and σ^2_{Ls}, and here we use the method of moments. First note that the regression of $\log(\mathbf{W}_i)$ on $\log(\mathcal{L}_i)$ for a person in stratum s is just $\alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s}$, so that α_{0s} and α_{1s} can be estimated by this regression. Since the residual variance for an individual is $\sigma^2_{wi} = \sigma^2_{Ls} + \sigma^2_{i,\text{uc}}$, if the (known) mean of the classical uncertainties $\sigma^2_{i,\text{uc}}$ within stratum s is σ^2_s, then the mean squared error of the regression has mean $\sigma^2_{Ls} + \sigma^2_s$: Subtracting σ^2_s from the observed regression mean squared error yields a method of moments estimate of σ^2_{Ls}.

8.6.2 Maximum Likelihood Implementation

The implementation of maximum likelihood is fairly straightforward. The four steps described in Figure 8.1 work as follows. Basically, we are going to compute the likelihood function for $\mathbf{Y}$ and $\log(\mathbf{W})$ given $\mathbf{Z}$, and we will work with the log scale.

The first step, of course, is the regression model (8.13). Write

$$\mathcal{H}\{\log(\mathbf{X}), \mathbf{Z}\} = H\left[\beta_0 + \beta_z \mathbf{Z} + \log\{1 + \beta_x \exp\{\log(\mathbf{X})\}\right].$$

The the likelihood function if $\log(\mathbf{X})$ could be observed is just a typical logistic likelihood:

$$\left[\mathcal{H}\{\log(\mathbf{X}), \mathbf{Z}\}\right]^{\mathbf{Y}} \left[1 - \mathcal{H}\{\log(\mathbf{X}), \mathbf{Z}\}\right]^{1-\mathbf{Y}}.$$

The second step is the error model, which is really of a form described in (2.1) of Section 2.2. Let $\rho_{i*} = \sigma^2_{Ls}/\sigma^2_{xi}$, so that given $\{\log(\mathbf{X}_i), \mathbf{Z}_i\}$, $\log(\mathbf{W}_i)$ is normally distributed with mean

$$\mu_{iw|x}\{\mathbf{Z}_i, \log(\mathbf{X}_i)\} = (\alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s})(1 - \sigma^2_{Ls}/\sigma^2_{xi}) + (\sigma^2_{Ls}/\sigma^2_{xi})\log(\mathbf{X}_i)$$

and variance $\sigma^2_{iw|x} = \sigma^2_{wi}(1 - \rho^2_{i*})$.

The third step is the distribution of $\log(\mathbf{X})$ given $\mathbf{Z}$, which we have

already done. Remember that $\log(\mathbf{X}_i)$ is normally distributed with mean $\mu_{xi} = \alpha_{0s} + \mathbf{Z}_i^t \alpha_{1s}$ and variance σ^2_{xi}.

We now just apply (8.7). Let $\phi(x, \mu, \sigma^2)$ be the normal density function, with mean μ and variance σ^2 evaluated at x. Then, the likelihood function for $\mathbf{Y}_i$ and $\log(\mathbf{W}_i)$ given $\mathbf{Z}_i$ is

$$\int \left[\mathcal{H}\{s, \mathbf{Z}_i\}\right]^{\mathbf{Y}_i} \left[1 - \mathcal{H}\{s, \mathbf{Z}_i\}\right]^{1-\mathbf{Y}_i} \times \phi\{\log(\mathbf{W}_i), \mu_{iw|x}(\mathbf{Z}_i, s), \sigma^2_{iw|x}\}\phi(s, \mu_{xi}, \sigma^2_{xi})ds.$$

Unfortunately, this likelihood function does not have a closed form. Rather than computing the integral using Monte Carlo methods (Section B.7.1), we used numerical quadrature. Specifically, Gaussian quadrature (Thisted, 1988) is a way of approximating integrals of the form $\int g(t)\exp(-t^2)ds$ as a sum $\sum_j w_j g(t_j)$. To apply this, we have to do a change of variables of the likelihood, namely, to replace s by $t = (s - \mu_{xi})/\sqrt{2\sigma^2_{xi}}$, so that the likelihood becomes

$$\int \left[\mathcal{H}\{\mu_{xi} + t\sqrt{2\sigma^2_{xi}}, \mathbf{Z}_i\}\right]^{\mathbf{Y}_i} \left[1 - \mathcal{H}\{\mu_{xi} + t\sqrt{2\sigma^2_{xi}}, \mathbf{Z}_i\}\right]^{1-\mathbf{Y}_i} \times \phi\{\log(\mathbf{W}_i), \mu_{iw|x}(\mathbf{Z}_i, \mu_{xi} + t\sqrt{2\sigma^2_{xi}}), \sigma^2_{iw|x}\}\exp(-t^2)dt.$$

In our implementation, we started from the regression calibration estimates and used the function optimizer “fmincon” in MATLAB.

8.7 Bronchitis Example

This section describes a cautionary tale about identifying Berkson error models, which we believe are often better analyzed via Bayesian methods.

In occupational medicine, an important problem is the assessment of the health hazard of specific harmful substances in a working area. One approach to modeling assumes that there is a threshold concentration, called the *threshold limiting value* (TLV), under which there is no risk due to the substance. Estimating the TLV is of particular interest in the industrial workplace. We consider here the specific problem of estimating the TLV in a dust-laden mechanical engineering plant in Munich.

The regressor variable $\mathbf{X}$ is the logarithm of 1.0 plus the average dust concentration in the working area over the period of time in question, and $\mathbf{Y}$ is the indicator that the worker has bronchitis. In addition, the duration of exposure $\mathbf{Z}_1$ and smoking status $\mathbf{Z}_2$ are also measured. Following Ulm (1991), we based our analysis upon the segmented logistic model

$$\text{pr}(\mathbf{Y} = 1|\mathbf{X}, \mathbf{Z})$$

$$= H\left\{\beta_0 + \beta_{x,1}(\mathbf{X} - \beta_{x,2})_+ + \beta_{z,1}\mathbf{Z}_1 + \beta_{z,2}\mathbf{Z}_2\right\}, \qquad (8.16)$$

where $(a)_+ = a$ if $a > 0$ and $= 0$ if $a \leq 0$. The parameter of primary interest is $\beta_{x,2}$, the TLV.

It is impossible to measure $\mathbf{X}$ exactly, and instead sample dust concentrations were obtained several times between 1960 and 1977. The resulting measurements are $\mathbf{W}$. There were 1,246 observations: 23% of the workers reported chronic bronchitis, and 74% were smokers. Measured dust concentration had a mean of 1.07 and a standard deviation of 0.72. The durations $\mathbf{Z}_1$ were effectively independent of concentrations, with correlation 0.093, compare with Ulm's (1991) Figure 3. Smoking status is also effectively independent of dust concentration, with the smokers having mean concentration 1.068, and the nonsmokers having mean 1.083. Thus, in this example, for likelihood calculations we will treat the $\mathbf{Z}$'s as if they were independent of $\mathbf{X}$.

A preliminary segmented regression analysis ignoring measurement error suggested an estimated TLV $\widehat{\beta}_{x,2} = 1.27$. We will call this the *naive TLV.*

As in Section 8.6, the data really consist of a complex mixture of Berkson and classical errors. The classical errors come from the measures of dust concentration in factories, while the Berkson errors come from the usual occupational epidemiology construct, wherein no direct measures of dust exposure are taken on individuals, but instead plant records of where they worked and for how long are used to impute some version of dust exposure. In this section, for illustrative purposes, we will assume a pure Berkson error structure. In the first edition of this book, we reported a much different classical error analysis with a flexible distribution for $\mathbf{X}$; see also Küchenhoff and Carroll (1995). A Bayesian treatment of segmented regression can be found in Gössi and Küchenhoff (2001); in Carroll, Roeder, and Wasserman (1999), who analyzed the Bronchitis example assuming Berkson errors and a semiparametric error distribution; in Section 9.5.4, where a classical error model is assumed and either validation or replication data are available; and in Section 9.7.3, where the Bronchitis data are analyzed assuming normally distributed Berkson errors. The likelihood analysis of segmented regression when validation data are available is discussed by Staudenmayer and Spiegelman (2002), who assumed a Berkson error model.

8.7.1 Calculating the Likelihood

We have already identified the model if $\mathbf{X}$ were observed (Step 1), and we have decided upon a Berkson error model with measurement error variance (Steps 2 and 3), so it remains to compute the likelihood function

(Step 4). For simplicity, write

$$\mathcal{H}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathcal{B}) = H\left\{\beta_0 + \beta_{x,1}(\mathbf{X} - \beta_{x,2})_+ + \beta_{z,1}\mathbf{Z}_1 + \beta_{z,2}\mathbf{Z}_2\right\}.$$

Let $\phi(x, \mu, \sigma^2)$ be the normal density function with mean μ and variance σ^2 evaluated at x. Then, as in the Nevada Test Site example in Section 8.6, from (8.8) the likelihood function is

$$\int \{\mathcal{H}(\mathbf{Y}, x, \mathbf{Z}, \mathcal{B})\}^{\mathbf{Y}} \{1 - \mathcal{H}(\mathbf{Y}, x, \mathbf{Z}, \mathcal{B})\}^{1-\mathbf{Y}} \phi(x, 0, \sigma_u^2) dx$$
$$= \int \{\mathcal{H}(\mathbf{Y}, \mathbf{W} + s(2\sigma_u^2)^{1/2}, \mathbf{Z}, \mathcal{B})\}^{\mathbf{Y}} \{1 - \mathcal{H}(\cdot)\}^{1-\mathbf{Y}} \exp(-s^2) ds,$$

which can be computed by Gaussian quadrature. Note that the maximization is supposed to be over $\mathcal{B}$ and σ_u^2.

8.7.1.1 Theoretical Identifiability

All the parameters, including the Berkson error variance, are identified, in the sense that if the sample size were infinite, then all parameters would be known. Küchenhoff and Carroll (1997) showed this fact in probit regression, and it is generally true in nonlinear models.

8.7.2 Effects of Measurement Error on Threshold Models

It is first of all interesting to understand what the effects of measurement error are on segmented regression models or threshold models. We made the point in Section 1.1 that measurement error causes loss of features. Here, that loss is quite profound. In Figure 8.4, we graph (solid line) the true probabilities as a function of $\mathbf{X}$ in a segmented logistic regression with intercept $\beta_0 = 0$, slope $\beta_x = 3$ and threshold $= 0$. Note the abrupt change in the probability surface at the threshold. We also plot (dashed line) the actual probabilities of the observed data as a function of $\mathbf{W}$ when there is Berkson measurement error with variance $\sigma_u^2 = 1$. Note how the observed data have smooth probabilities: Indeed, the true threshold nature of the data have been obliterated by the measurement error. One can easily imagine, then, that trying to identify the threshold or the error variance σ_u^2 is likely to be challenging.

8.7.3 Simulation Study and Maximum Likelihood

We performed a small simulation study to show how difficult it might be to estimate a threshold model in a Berkson error case. We fit the threshold model (8.16) to the observed data and used this fit to get estimates of the parameters. We kept the $\mathbf{W}$ and $\mathbf{Z}$ data fixed, as in the actual data, used the naive parameter estimates as the true

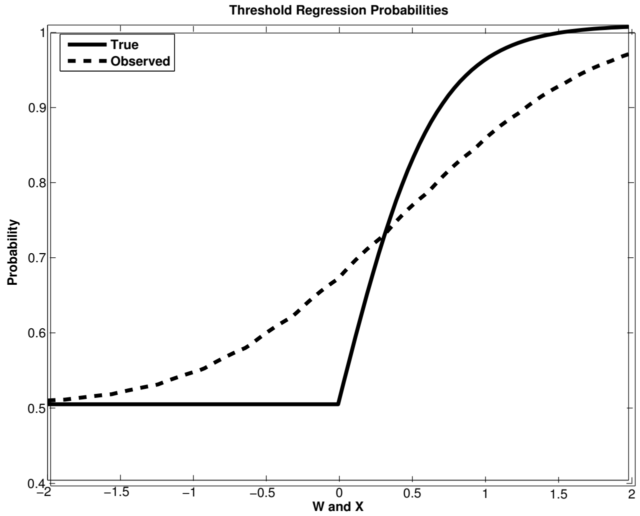


Figure 8.4 *The true probabilities (solid line) as a function of* $\mathbf{X}$ *and the observed probabilities (dashed line) as a function of* $\mathbf{W}$ *in segmented Berkson logistic regression with intercept* $\beta_0 = 0$*, slope* $\beta_x = 3$*, threshold* $= 0$*, and Berkson measurement error with variance* $= 1$*. Note how the observed data have smooth probabilities, while the true but unobserved data have the abrupt change at the threshold.*

parameters, and generated large Berkson errors $\mathbf{X} = \mathbf{W} + \mathbf{U}$, where $\text{var}(\mathbf{U}) = \text{var}(\mathbf{W}) = 0.72.^2 = 0.52$. We then generated simulation observations $\mathbf{Y}$ from model (8.16) and ran 200 simulated data sets. We then fit a maximum likelihood analysis to each simulated data set.

The true TLV in this simulation was $\beta_{x,2} = 1.27$, and the mean of the estimates across the simulations was 1.21, very nearly unbiased. The true Berkson error variance was 0.52, while the mean estimate over the simulations was 0.43, only slightly biased. So, one might ask, "What's the problem?" The problem is that in 35% of the simulated data sets, the MLE for $\sigma_u^2 = 0$! It is, to put it mildly, not very helpful when one knows that there is Berkson error but an algorithm announces that the Berkson error does not exist. This is one of those cases where there is technical identifiability of a parameter, but the free lunch of identifiability is rather skimpy.

This example also illustrates a problem with maximum likelihood when the likelihood is maximized at the boundary of the parameter space. Then the MLE takes a value which is the most extreme case of plausible values. In contrast, the usual Bayesian estimator, the mean of the posterior distribution, will not be this extreme, for example, would not be equal to zero when estimating a variance.

8.7.4 Berkson Analysis of the Data

If the reader has been paying attention, the previous discussion is obviously leading up to a problem with the analysis. We applied Berkson measurement error maximum likelihood to the bronchitis data, and the estimated measurement error variance was $\sigma_u^2 = 0.0$! Of course, the simulation study showed that this can happen in as many as one third of data sets, so it is an unfortunate finding but certainly no surprise. In some sense, this analysis is a cautionary tale that technical identifiability does not always lead to practical identifiability. The bioassay data of Section 4.7.3 are, of course, the counterpoint to this: There are indeed problems where technical and real identifiability coincide.

8.8 Quasilikelihood and Variance Function Models

Quasilikelihood and variance function (QVF) models are defined in Section A.7. In this approach, we model only the mean and variance functions of the response, and not its entire distribution, writing the mean function as $E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = m_{\mathbf{Y}}(\mathbf{Z},\mathbf{X},\mathcal{B})$ and the variance function as $\text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = \sigma^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)$.

Quasilikelihood and variance function techniques require that we compute the mean and variance functions of the *observed* data (and not the unobservable data). These are given by

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = E\left\{m_{\mathbf{Y}}(\cdot)|\mathbf{Z},\mathbf{W}\right\}, \tag{8.17}$$

$$\text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = \sigma^2 E\left\{g^2(\cdot)|\mathbf{Z},\mathbf{W}\right\} + \text{var}\left\{m_{\mathbf{Y}}(\cdot)|\mathbf{Z},\mathbf{W}\right\}. \tag{8.18}$$

Equations (8.17) and (8.18) define a variance function model. If we knew the functional forms of the mean and variance functions, then we could apply the fitting and model criticism techniques discussed in Section A.7. Note how both (8.17) and (8.18) require an estimate of a model for the distribution of the unobserved covariate given the observed covariates and the surrogate.

A QVF model analysis follows the same pattern of a likelihood analysis. As described in Figure 8.5, the steps required are as follows.

- **Step 1**: Specify the mean and variance of $\mathbf{Y}$ if $\mathbf{X}$ were observed.
- **Step 2**: Specify a model relating $\mathbf{W}$ to $\mathbf{X}$ that allows identification of all model parameters. This will have a classical component, so it is the same as Step 2 for a classical analysis, that is, a model for $\mathbf{W}$

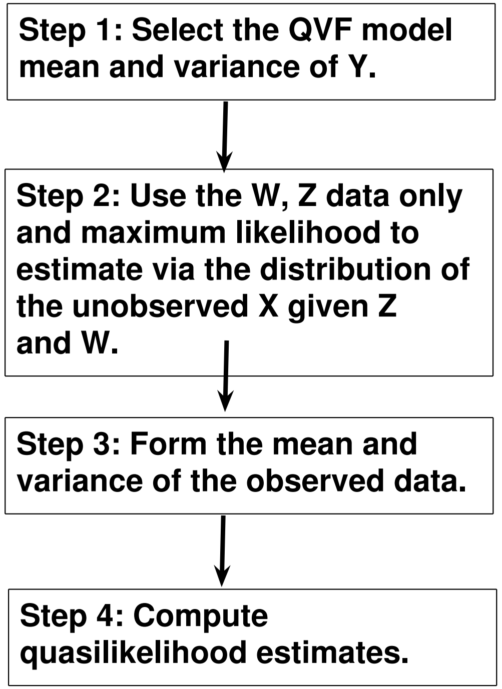

Step 1: Select the QVF model mean and variance of Y.
Step 2: Use the W, Z data only and maximum likelihood to estimate via the distribution of the unobserved X given Z and W.
Step 3: Form the mean and variance of the observed data.
Step 4: Compute quasilikelihood estimates.

Figure 8.5 *The steps in a quasilikelihood analysis with measurement error.*

given $(\mathbf{X}, \mathbf{Z})$ and a model for $\mathbf{X}$ given $\mathbf{Z}$. We write the densities given by these models as $f_{W|Z,X}(w|z, x, \widetilde{\alpha}_1)$ and $f_{X|Z}(x|z, \widetilde{\alpha}_2)$, respectively.

- **Step 3**: Do a maximum likelihood analysis of the $(\mathbf{W}, \mathbf{Z})$ data only to estimate the parameters $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$; see below for details.
- **Step 4**: Form (8.17)–(8.18), the observed data mean and variance functions, and then apply the fitting methods in Section A.7.

8.8.1 Details of Step 3 for QVF Models

The (reduced) likelihood for a single observation based upon only the observed covariates is

$$\int f_{W|Z,X}(\mathbf{W}|\mathbf{Z}, x, \widetilde{\alpha}_1) f_{X|Z}(x|\mathbf{Z}, \widetilde{\alpha}_2) d\mu(x),$$

where again the integral is replaced by a sum if $\mathbf{X}$ is discrete. The $(\mathbf{W}, \mathbf{Z})$ data are used to estimate $(\widetilde{\alpha}_1, \widetilde{\alpha}_2)$ by multiplying this reduced likelihood over the observations, taking logarithms, and then maximizing.

8.8.2 Details of Step 4 for QVF Models

The density or mass function of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is then given by

$$f_{X|Z,W}(x|z, w) = \frac{f_{W|Z,X}(w|z, x, \widetilde{\alpha}_1) f_{X|Z}(x|z, \widetilde{\alpha}_2)}{\int f_{W|Z,X}(w|z, v, \widetilde{\alpha}_1) f_{X|Z}(v|z, \widetilde{\alpha}_2) d\mu(v)}.$$

From this, one can obtain (8.17) and (8.18) by integration and either analytically or numerically. Thus, for example, (8.17) becomes

$$E(\mathbf{Y}|\mathbf{W}, \mathbf{Z}) = \int m_{\mathbf{Y}}(\mathbf{Z}, x, \mathcal{B}) f_{X|Z,W}(x|\mathbf{Z}, \mathbf{W}) dx.$$

The sandwich method or the bootstrap can be used for inference, although of course one must take into account the estimation of $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$, something the bootstrap does naturally.

Bibliographic Notes

Earlier references prior to the first edition of this text include Carroll, Spiegelman, Lan, et al. (1984) and Schafer (1987, 1993) for probit regression; Whittemore and Gong (1991) in a Poisson model; Crouch and Spiegelman (1990) and Wang, Carroll and Liang (1997) in logistic regression; and Küchenhoff and Carroll (1997) in a change point problem. Some recent references include Turnbull, Jiang, and Clark (1997); Gould, Stefanski, and Pollock (1997); Spiegelman and Casella (1997); Lyles, Munoz, Xu, et al. (1999); Buonaccorsi, Demidenko, and Toteson (2000); Nummi (2000); Higdon and Schafer (2001); Schafer, Lubin, Ron, et al. (2001); Schafer (2002); Aitkin and Rocci (2002); Mallick, Hoffman, and Carroll (2002); Augustin (2004); and Wannemuehler and Lyles (2005).

For simple linear regression, Schafer and Purdy (1996) compare maximum likelihood estimators with profile-likelihood confidence intervals to method-of-moments estimators with confidence intervals based on asymptotic normality. They note that in some situations the estimators can have skewed distributions and then likelihood-based intervals have more accurate coverage probabilities.

The cervical cancer data in Section 8.4 has validation data, so that the misclassification probabilities can be easily estimated. In other examples, validation data are absent but there is more than one surrogate. Gustafson (2005) discusses identifiability in such contexts, as well as Bayesian strategies for handling nonidentifiability.

CHAPTER 9

BAYESIAN METHODS

9.1 Overview

Over the last two decades, there has been an "MCMC revolution" in which Bayesian methods have become a highly popular and effective tool for the applied statistician. This chapter is a brief introduction to Bayesian methods and their applications in measurement error problems. The reader new to Bayesian statistics is referred to the bibliographic notes at the end of this chapter for further reading.

We will not go into the philosophy of the Bayesian approach, whether one should be an objective or a subjective Bayesian, and so forth. We recommend reading Efron (2005), who has a number of amusing comments on the differences between Bayesians and Frequentists, and also on the differences among Bayesians. Our focus here will be how to formulate measurement error models from the Bayesian perspective, and how to compute them. For those familiar with Bayesian software such as WinBUGS, a Bayesian analysis is sometimes relatively straightforward. Bayesian methods also allow one to use other sources of information, for example, from similar studies, to help estimate parameters that are poorly identified by the data alone. A disadvantage of Bayesian methods, which is shared by maximum likelihood, is that, compared to regression calibration, computation of Bayes estimators is intensive. Another disadvantage shared by maximum likelihood is that one must specify a full likelihood, and therefore one should investigate whether the estimator is robust to possible model misspecification.

9.1.1 Problem Formulation

Luckily, Bayesian methods start from a likelihood function, a topic we have already addressed in Chapter 8 and illustrated with a four-step approach in Figure 8.1.

In the Bayesian approach, there are five essential steps:

- **Step 1**: This is the same as the first step in a likelihood approach. Specifically, one must specify a parametric model for every component of the data. Any likelihood analysis begins with the model one would use if **X** were observable.

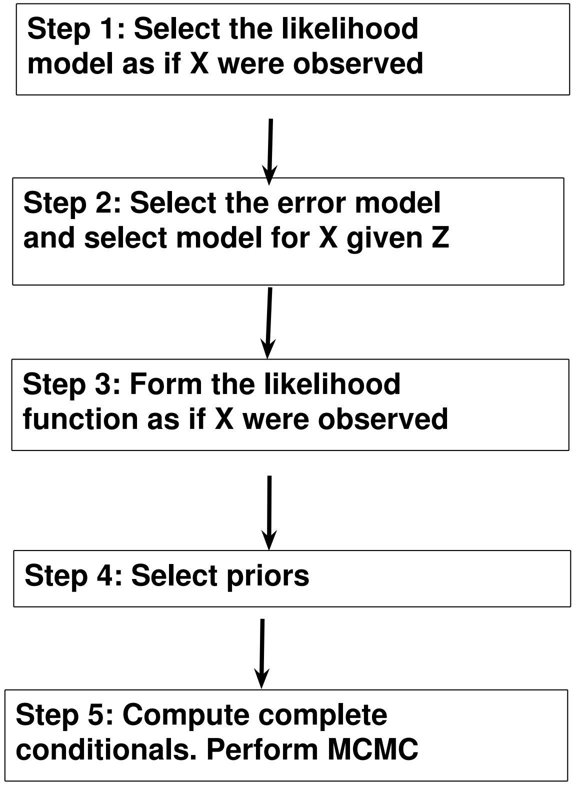

Step 1: Select the likelihood model as if X were observed
Step 2: Select the error model and select model for X given Z
Step 3: Form the likelihood function as if X were observed
Step 4: Select priors
Step 5: Compute complete conditionals. Perform MCMC

Figure 9.1 *Five basic steps in performing a Bayesian analysis of a measurement error problem. If automatic software such as WinBUGS is used, the complete conditionals, which often require detailed algebra, need not be computed.*

- **Step 2**: This step too agrees with the likelihood approach. The next crucial decision is the error model that is to be chosen. This could be a classical error model, a Berkson model, or a combination of the two. If one has classical components in the measurement error model, then typically one also needs to specify a distribution for the unobserved $\mathbf{X}$ given the observable covariates $\mathbf{Z}$.
- **Step 3**: The typical Bayesian approach treats $\mathbf{X}$ as missing data, and, in effect, imputes it multiple times by drawing from the conditional distribution of $\mathbf{X}$ given all other variables. Thus, at this step, the likelihood of all the data, including $\mathbf{W}$, is formed as if $\mathbf{X}$ were available.

- **Step 4**: In the Bayesian approach, parameters are treated as if they were random, one of the essential differences with likelihood methods. If one is going to treat parameters as random, then they need to be given distributions, called *prior distributions*. Much of the controversy among statisticians regarding Bayesian methods revolves around these prior distributions.
- **Step 5**: The final step is to compute Bayesian quantities, in particular the *posterior distribution* of parameters given all the *observed* data. There are various approaches to doing this, most of them revolving around Markov Chain Monte Carlo (MCMC) methods, often based on the Gibbs sampler. In some problems, such as with WinBUGS, users do not actually have to do anything but run a program, and the appropriate posterior quantities become available. In other cases, though, either the standard program is not suitable to the problem, or the program does not work well, in which case one has to tailor the approach carefully. This usually involves detailed algebraic calculation of what are called the *complete conditionals*, the distribution of the parameters, and the $\mathbf{X}$ values, given everything else in the model. We give a detailed example of this process in Section 9.4.

9.1.2 Posterior Inference

Bayesian inference is based upon the posterior density, which is the conditional density of unobserved quantities (the parameters and unobserved covariates) given the observed data, and summarizes all of the information about the unobservables. For example, the mean, median, or mode of the posterior density are all suitable point estimators. A region with probability $(1-\alpha)$ under the posterior is called a *credible set*, and is a Bayesian analog to a confidence region. To calculate the posterior, one can take the joint density of the data and parameters and, at least in principle, integrate out the parameters to get the marginal density of the data. One can then divide the joint density by this marginal density to get the posterior density.

There are many "textbook examples" where the posterior can be computed analytically, but in practical applications this is often a non trivial problem requiring high-dimensional numerical integration. The computational problem has been the subject of much recent research. The method currently receiving the most attention in the literature is the Gibbs sampler and related methods such as the Metropolis–Hastings algorithm (Hastings, 1970; Geman & Geman, 1984; Gelfand & Smith, 1990).

The Gibbs sampler, which is often called Markov Chain Monte Carlo (MCMC), generates a Markov chain whose stationary distribution is the posterior distribution. The key feature of the Gibbs sampler is that this chain can be simulated using only the joint density of the parameters, the unobserved **X**-values and the observed data, for example, the product of the likelihood and the prior, and not the unknown posterior density which would require an often intractable integral. If the chain is run long enough, then the observations in a sample from the chain are approximately identically distributed, with common distribution equal to the posterior. Thus posterior moments, the posterior density, and other posterior quantities can be estimated from a sample from the chain.

The Gibbs sampler "fills in" or imputes the values of the unobserved covariates **X** by sampling from their conditional distribution given the observed data and the other parameters. This type of imputation differs from the imputation of regression calibration in two important ways. First, the Gibbs sampler makes a large number of imputations from the conditional distribution of **X**, whereas regression calibration uses a single imputation, namely the conditional expectation of **X** given **W** and **Z**. Second, the Gibbs sampler conditions on **Y** as well as **W** and **Z** when imputing values of **X**, but regression calibration does not use information about **Y** when imputing **X**.

9.1.3 Bayesian Functional and Structural Models

We made the point in Section 2.1 that our view of functional and structural modeling is that in the former, we make no or at most few assumptions about the distribution of the unobserved **X**-values. Chapters 5 and 7 describe methods that are explicitly functional, while regression calibration is approximately functional.

In contrast, likelihood methods (Chapter 8) and Bayesian methods necessarily must specify a distribution for **X** in one way or another, and here the distinction between functional and structural is blurred. Effectively, structural Bayesian likelihood modeling imposes a simple model on **X**, such as the normal model, while functional methods specify flexible distributions for **X**. We use structural models in this chapter. Examples of this approach are given by Schmid and Rosner (1993), Richardson and Gilks (1993), and Stephens and Dellaportas (1992).

There are at least several ways to formulate a Bayesian functional model. One way would allow the distribution of **X** to depend on the observation number, i. Müller and Roeder (1997) used this idea for the case when **X** is partially observed. They assume that the $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ are jointly normally distributed with mean μ_i and covariance matrix Σ_i, where $\theta_i = (\mu_i, \Sigma_i)$ is modeled by a Dirichlet process distribution which itself has unknown hyperparameters. Lindley and El Sayyad (1968) wrote the first Bayesian paper on functional models, covering the linear regression case. Because of their complexity, we do not consider Bayesian functional models here.

A second possibility intermediate between functional and hard-core structural approaches is to specify flexible distributions, much as we suggested in Section 8.2.2. Carroll, Roeder, and Wasserman (1999) and Richardson, Leblond, Jaussent, and Green (2002) used mixtures of normal distributions. Gustafson, Le, and Vallee (2002) used an approach based on approximating the distribution of **X** by a discrete distribution.

In this chapter, the $\mathbf{Z}_i$'s are treated as fixed constants, as we have done before in non-Bayesian treatments. This makes perfect sense, since Bayesians only need to treat unknown quantities as random variables. Thus, the likelihood is the conditional density of the $\mathbf{Y}_i$'s, $\mathbf{W}_i$'s, and any $\mathbf{X}_i$'s that are observed, given the parameters and the $\mathbf{Z}_i$'s. The posterior is the conditional density of the parameters given all data, that is, the $\mathbf{Z}_i$'s, $\mathbf{Y}_i$'s, $\mathbf{W}_i$'s, and any observed $\mathbf{X}_i$'s.

9.1.4 Modularity of Bayesian MCMC

The beauty of the Bayesian paradigm combined with modern MCMC computing is its tremendous flexibility. The technology is "modular" in that the methods of handling, for example, multiplicative error, segmented regression and the logistic regression risk model can be combined easily. In effect, if one knows how to handle these problems separately, it is often rather easy to combine them into a single analysis and program.

9.2 The Gibbs Sampler

As in Chapter 8, especially equation (8.7), the first three steps of our Bayesian paradigm result in the likelihood computed as if **X** were observable. Dropping the second measure **T**, this likelihood for an individual observation becomes

$$\begin{aligned} f(\mathbf{Y}, \mathbf{W}, \mathbf{X}|\mathbf{Z}, \Omega) \quad = \quad & f_{Y|Z,X}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \mathcal{B}) \\ & \times f_{W|Z,X}(\mathbf{W}|\mathbf{Z}, \mathbf{X}, \widetilde{\alpha}_1) f_{X|Z}(\mathbf{X}|\mathbf{Z}, \widetilde{\alpha}_2), \end{aligned}$$

where Ω is the collection of all unknown parameters. As in the fourth step of the Bayesian paradigm, we let Ω have a prior distribution $\pi(\Omega)$. The likelihood of all the "data" then becomes

$$\pi(\Omega) \prod_{i=1}^{n} f(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i|\mathbf{Z}_i, \Omega).$$

To keep this section simple, we have not included the possibility of validation data here, but that could be done with only some additional effort, mostly notational. To keep notation compact, we will write the ensemble of $\mathbf{Y}$, $\mathbf{X}$, etc., as $\widetilde{\mathbf{Y}}$, $\widetilde{\mathbf{X}}$, etc. This means that the likelihood can be expressed as

$$\pi(\Omega) f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}}, \Omega).$$

The posterior distribution of Ω is then

$$f(\Omega \Big| \widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}) = \frac{\pi(\Omega) \int f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{x} | \widetilde{\mathbf{Z}}, \Omega) d\widetilde{x}}{\int \pi(\omega) f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{x} | \widetilde{\mathbf{Z}}, \omega) d\widetilde{x} d\omega}. \tag{9.1}$$

The practical problem is that, even if the integration in $\widetilde{x}$ can be accomplished or approximated as in Chapter 8, the denominator of (9.1) may be very difficult to compute. Numerical integration typically fails to provide an adequate approximation even when there are as few as three or four components to Ω.

The Gibbs sampler is one solution to the dilemma. The Gibbs sampler is an iterative, Monte Carlo method consisting of the following main steps, starting with initial values of Ω:

- Generate a sample of the unobserved $\mathbf{X}$-values by sampling from their posterior distributions given the current value of Ω, the posterior distribution of $\mathbf{X}_i$ being

  $$f(\mathbf{X}_i | \mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i, \Omega) = \frac{f(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i | \mathbf{Z}_i, \Omega)}{\int f(\mathbf{Y}_i, \mathbf{W}_i, x | \mathbf{Z}_i, \Omega) dx}. \tag{9.2}$$

  As we indicate below, this can be done without having to evaluate the integral in (9.2).
- Generate a new value of Ω from its posterior distribution given the observed data and the current generated $\mathbf{X}$-values, namely,

  $$f(\Omega \Big| \widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}, \widetilde{\mathbf{X}}) = \frac{\pi(\Omega) f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}}, \Omega)}{\int \pi(\omega) f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}}, \omega) d\omega}. \tag{9.3}$$

  Often, this is done one element of Ω at a time, holding the others fixed (as described below, here too we do not need to compute the integral). Thus, for example, if the j^{th} value of Ω is ω_j, and the other components of Ω are $\Omega_{(-j)}$, then the posterior in question is simply

  $$\begin{aligned} & f(\omega_j | \widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}, \widetilde{\mathbf{X}}, \Omega_{(-j)}) \qquad (9.4) \\ & \quad = \frac{\pi(\omega_j, \Omega_{(-j)}) f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}}, \omega_j, \Omega_{(-j)})}{\int \pi(\omega_j^*, \Omega_{(-j)}) f(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{X}} | \widetilde{\mathbf{Z}}, \omega_j^*, \Omega_{(-j)}) d\omega_j^*}. \end{aligned}$$
- Repeat this many times. Discard the first few of the generated samples, the so-called burn-in period.

- Quantities such as the posterior mean and posterior quantiles are estimated by the sample mean and quantiles of $\Omega_1, \Omega_2, \ldots$, while kernel density estimates are used to approximate the entire posterior density or the marginal posterior density of a single parameter or subset of parameters.

An important point is that the first two steps do *not* require that one evaluates the integral in the denominator on the right-hand sides of (9.2), (9.3), and (9.4).

Generating pseudorandom observations from (9.4) is the heart of the Gibbs sampler. Often the prior on ω_j is conditionality conjugate so that the full conditional for ω_j is in the same parametric family as the prior, for example, both are normal or both are inverse-gamma; see Section A.3 for a discussion of the inverse-gamma distribution. In such cases, the denominator of (9.4) can be determined from the form of the posterior and the integral need not be explicitly calculated.

If we do not have conditional conjugacy, then drawing from the full conditional of ω_j is more difficult. In this situation, we will use a Metropolis–Hastings, step which will be described soon. The Metropolis–Hastings algorithm does not require that the integral in (9.4) be evaluated.

9.3 Metropolis–Hastings Algorithm

The Metropolis–Hastings algorithm (MH algorithm) is a very versatile and flexible tool, and even includes the Gibbs sampler as a special case. Suppose we want to sample from a certain density, which in applications to Bayesian statistics is the posterior, and that the density is $Cf(\cdot)$, where f is known but the normalizing constant $C > 0$ is difficult to evaluate; see, for example, (9.3). The MH algorithm uses f without knowledge of C to generate a Markov chain whose stationary distribution is $Cf(\cdot)$.

To simplify the notation, we will subsume the unobserved $\mathbf{X}$ into Ω; this involves no loss of generality, since a Bayesian treats all unknown quantities in the same way. Suppose that the current value of Ω is Ω_{curr}. The idea is to generate (see below) a ”candidate” value Ω_{cand} and either accept it as the new value or reject it and stay with the current value. Over repeated application, this process results in random variables with the desired distribution.

Mechanically, one has to have a candidate distribution, which may depend upon the current value. We write this candidate density as $q(\Omega_{\text{cand}} | \Omega_{\text{curr}})$. Gelman, Stern, Carlin, and Rubin (2004) call $q(\cdot|\cdot)$ a “jumping rule,” since it may generate the jump from Ω_{curr} to Ω_{cand}. Thus, a candidate Ω_{cand} is generated from $q(\cdot|\Omega_{\text{curr}})$. This candidate is accepted and becomes Ω_{curr} with probability

$$r = \min\left\{1, \frac{f(\Omega_{\text{cand}})q(\Omega_{\text{curr}}|\Omega_{\text{cand}})}{f(\Omega_{\text{curr}})q(\Omega_{\text{cand}}|\Omega_{\text{curr}})}\right\}. \tag{9.5}$$

More precisely, a uniform(0,1) random variable V is drawn, and then we set $\Omega_{\text{curr}} = \Omega_{\text{cand}}$ if $V \leq r$.

The popular "random-walk" MH algorithm uses $q(\Omega_{\text{cand}}|\Omega_{\text{curr}}) = h(\Omega_{\text{cand}} - \Omega_{\text{curr}})$ for some probability density h. Often, as in our examples, $h(\cdot)$ is symmetric so that

$$r = \min\left\{1, \frac{f(\Omega_{\text{cand}})}{f(\Omega_{\text{curr}})}\right\}. \tag{9.6}$$

The "Metropolis–Hastings within Gibbs algorithm" uses the MH algorithm at those steps in a Gibbs sampler where the full conditional is difficult to sample. Suppose sampling ω_j is one such step. If we generate the candidate $\omega_{j,\text{cand}}$ from $h(\cdot - \omega_{j,\text{curr}})$ where h is symmetric and $\omega_{j,\text{curr}}$ is the current value of ω_j, then r in (9.6) is

$$r = \min\left\{1, \frac{f(\omega_{j,\text{cand}}|\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}, \omega_{\ell,\text{curr}} \text{ for } \ell \neq j)}{f(\omega_{j,\text{curr}}|\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}, \omega_{\ell,\text{curr}} \text{ for } \ell \neq j)}\right\}.$$

Often, h is a normal density, a heavy-tailed normal mixture, or a t-density. The scale parameter of this density should be chosen so that typical values of $\omega_{j,\text{cand}}$ are neither too close to nor too far from $\omega_{j,\text{curr}}$. If $\omega_{j,\text{cand}}$ is too close to $\omega_{j,\text{curr}}$ with high probability, then the MH algorithm takes mostly very small steps and does not move quickly enough. If $\omega_{j,\text{cand}}$ is generally too far from $\omega_{j,\text{curr}}$, then the probability of acceptance is small. To get good performance of the Metropolis within Gibbs algorithm, we might use a Normal$(0, \sigma^2)$ proposal density where σ^2 is tuned to the algorithm so that the acceptance probability is between 25% and 50%. Gelman, Carlin, Stern, and Rubin (2004, p. 306) state that the optimal jumping rule has 44% acceptance in one dimension and about 23% acceptance probability in high dimensions when the jumping and target densities have the same shape. To allow for occasional large jumps, one might instead use a heavy-tailed normal mixture of 90% Normal$(0, \sigma^2)$ and 10% Normal$(0, L\sigma^2)$, where L might be 2, 3, 5, or even 10. This density is very easy to sample from, since we need only generate independent $Z \sim$ Normal(0, 1) and $U \sim [0, 1]$. Then, we multiply Z by σ or $\sqrt{L}\,\sigma$ according as $U \leq 0.9$ or $U > 0.9$. The Normal$(0, L\sigma^2)$ component gives the mixture heavy tails and allows the sampler to take large steps occasionally. One can experiment with the value of L to see which gives the best mixing, that is, the least autocorrelation in the sample.

More information on the Gibbs sampler and the MH algorithm can be found in Roberts, Gelman, and Gilks (1997), Chib and Greenberg (1995), Gelman et al. (2004), and in many other books and papers. See Roberts and Rosenthal (2001) for more discussion about scaling of MH jumping rules.

9.4 Linear Regression

In this section, an example is presented where the full conditionals are all conjugate. For those new to Bayesian computations, we will show in some detail how the full conditionals can be found. In the following sections, this example will be modified to models where some, but not all, full conditionals are conjugate.

Suppose we have a linear regression with a scalar covariate $\mathbf{X}$ measured with error and a vector $\mathbf{Z}$ of covariates known exactly. Then the first three steps in Figure 9.1 are as follows. The so-called "outcome model" for the outcome $\mathbf{Y}$ given all of the covariates (observed or not) is

$$\mathbf{Y}_i = \text{Normal}(\mathbf{Z}_i^t\beta_z + \mathbf{X}_i\beta_x, \sigma_\epsilon^2). \tag{9.7}$$

Suppose that we have replicates of the surrogate $\mathbf{W}$ for $\mathbf{X}$. Then the so-called "measurement model" is

$$\mathbf{W}_{i,j} = \text{Normal}(\mathbf{X}_i, \sigma_u^2),\ \ j = 1, \ldots, k_i. \tag{9.8}$$

Finally, suppose that the "exposure model" for the covariate measured with error, $\mathbf{X}$, given $\mathbf{Z}$ is

$$\mathbf{X}_i = \text{Normal}(\alpha_0 + \mathbf{Z}_i^t\alpha_z, \sigma_x^2). \tag{9.9}$$

The term *exposure model* comes from epidemiology, where $\mathbf{X}$ is often exposure to a toxicant.

For this model, it is possible to have conjugate priors for all of the full conditionals. The prior we will use is that independently

$$\beta_x = \text{Normal}(0, \sigma_\beta^2),\ \ \beta_z = \text{Normal}(0, \sigma_\beta^2\boldsymbol{I})$$

$$\alpha_0 = \text{Normal}(0, \sigma_\alpha^2),\ \ \alpha_z = \text{Normal}(0, \sigma_\alpha^2\boldsymbol{I}),$$

$$\sigma_\epsilon^2 = \text{IG}(\delta_{\epsilon,1}, \delta_{\epsilon,2}),\ \ \sigma_u^2 = \text{IG}(\delta_{u,1}, \delta_{u,2}),\ \ \sigma_x^2 = \text{IG}(\delta_{x,1}, \delta_{x,2}).$$

As discussed in Section A.3, this prior is conjugate for the full conditionals. Here IG$(\cdot, \cdot)$ is the inverse gamma density, and the hyperparameters σ_β and σ_μ are chosen to be "large" and the δ hyperparameters to be "small" so that the priors are relatively noninformative. In particular, because σ_β and σ_μ are large, using a mean of zero for the normal priors should not have much influence on the posterior. See Section A.3 for the definition of the inverse gamma distribution and discussion about choosing the hyperparameters of an inverse gamma prior. The unknowns in this model are $(\beta_x, \beta_z, \sigma_\epsilon, \sigma_x, \sigma_u)$, $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, and $(\alpha_0, \alpha_z, \sigma_x)$.

Define

$$\boldsymbol{C}_i = \begin{pmatrix} \mathbf{Z}_i \\ \mathbf{X}_i \end{pmatrix}, \ \ \mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)^t, \ \ \text{and} \ \ \beta = \begin{pmatrix} \beta_z \\ \beta_x \end{pmatrix}.$$

The likelihood for a single observation is

$$\begin{aligned} &f(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i | \mathbf{Z}_i, \Omega) = (2\pi)^{-3/2} \frac{1}{\sigma_x \sigma_\epsilon \sigma_u^{k_i}} \\ &\times \exp\{-\left(\mathbf{Y}_i - \boldsymbol{C}_i^t \beta\right)^2 / (2\sigma_\epsilon^2)\} \qquad (9.10) \\ &\times \exp\left\{-\textstyle\sum_{j=1}^{k_i} (\mathbf{W}_{i,j} - \mathbf{X}_i)^2 / (2\sigma_u^2) - (\mathbf{X}_i - \alpha_0 - \mathbf{Z}_i^t \alpha_z)^2 / (2\sigma_x^2)\right\}. \end{aligned}$$

The joint likelihood is, of course, the product over index i of the terms (9.10). The joint density of all observed data and all unknown quantities (parameters and true $\mathbf{X}$'s for nonvalidation data) is the product of the joint likelihood and the joint prior.

In our calculations, we will use the following:

Rule: If for some p-dimensional parameter θ we have

$$f(\theta | \text{others}) \propto \exp\left\{-\left(\theta^t \boldsymbol{A} \theta - 2\boldsymbol{b}\theta\right)/2\right\}$$

where the constant of proportionality is independent of θ, then $f(\theta|\text{others})$ is Normal$(\boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1})$.

To find the full conditional for β, we isolate the terms depending on β in this joint density. We write the full conditional of β given the others as $f(\beta|\text{others})$. This gives us

$$f(\beta | \text{others}) \propto \exp\left\{-\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{C}_i^t \beta)^2 - \frac{1}{2\sigma_\beta^2} \beta^t \beta\right\}, \qquad (9.11)$$

where the first term in the exponent comes from the likelihood and the second comes from the prior. Let $\mathcal{C}$ have i^{th} row $\boldsymbol{C}_i^t$ and let $\Delta = \sigma_\epsilon^2 / \sigma_\beta^2$. Then (9.11) can be rearranged to

$$f(\beta | \text{others}) \propto \exp\left[-\frac{1}{2\sigma_\epsilon^2} \left\{\beta^t \left(\mathcal{C}^t \mathcal{C} + \Delta \boldsymbol{I}\right) \beta + 2\mathcal{C}^t \mathbf{Y} \beta\right\}\right], \qquad (9.12)$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^t$. Using the Rule,

$$f(\beta | \text{others}) = N\left(\left\{\mathcal{C}^t \mathcal{C} + \Delta \boldsymbol{I}\right\}^{-1} \mathcal{C}^t \mathbf{Y}, \, \sigma_\epsilon^2 \left(\mathcal{C}^t \mathcal{C} + \Delta \boldsymbol{I}\right)^{-1}\right). \qquad (9.13)$$

Here we see how the Gibbs sampler can avoid the need to calculate integrals. The normalizing constant in (9.12) can be found from (9.13) simply by knowing the form of the normal distribution.

Result (9.13) is exactly what we would get without measurement error, except that for the nonvalidation data the $\mathbf{X}$'s in $\mathcal{C}$ are "filled-in" rather than known. Therefore, $\mathcal{C}$ will vary on each iteration of the Gibbs sampler. The parameters Δ and σ_ϵ will also vary, even if there is no measurement error.

The full conditional for $\alpha = (\alpha_0, \alpha_z^t)^t$ can be found in the same way as for β. First, analogous to (9.11),

$$f(\alpha | \text{others}) \propto \exp\left\{-\frac{\sum_{i=1}^n \{\mathbf{X}_i - (\alpha_0 + \mathbf{Z}_i^t \alpha_z)\}^2}{2\sigma_x^2} - \frac{\alpha^t \alpha}{2\sigma_\alpha^2}\right\}.$$

Let $D_i = (1 \ \mathbf{Z}_i^t)^t$ and let $\mathcal{D}$ be the matrix with i^{th} row equal to D_i^t. Also, let $\eta = \sigma_x^2 / \sigma_\alpha^2$. Then, analogous to (9.13),

$$f(\alpha | \text{others}) = N\left\{\left(\mathcal{D}^t \mathcal{D} + \eta \boldsymbol{I}\right)^{-1} \mathcal{D}^t \mathbf{X}, \, \sigma_x^2 \left(\mathcal{D}^t \mathcal{D} + \eta \boldsymbol{I}\right)^{-1}\right\}, \qquad (9.14)$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^t$.

To find the full conditional for $\mathbf{X}_i$, define $\overline{\mathbf{W}}_i = \sum_{J=1}^{k_i} W_{i,j}/k_i$. Then

$$\begin{aligned} f(\mathbf{X}_i | \text{others}) &\propto \exp\left[-(\mathbf{Y}_i - \mathbf{X}_i \beta_x - \mathbf{Z}_i^t \beta_z)^2 / (2\sigma_\epsilon^2)\right] \qquad (9.15) \\ &\times \exp\left\{-(\mathbf{X}_i - \alpha_0 - \mathbf{Z}_i^t \alpha_z)^2 / (2\sigma_x^2) - k_i (\overline{\mathbf{W}}_i - \mathbf{X}_i)^2 / (2\sigma_u^2)\right\}. \end{aligned}$$

After some algebra and applying the Rule again, $f(\mathbf{X}_i|\text{others})$ is seen to be normal with mean

$$\frac{(\mathbf{Y}_i - \mathbf{Z}_i^t \beta_z)(\beta_x / \sigma_\epsilon^2) + (\alpha_0 + \mathbf{Z}_i^t \alpha_z)/\sigma_x^2 + \overline{\mathbf{W}}_i / \sigma_{\overline{\mathbf{W}}}^2}{(\beta_x^2 / \sigma_\epsilon^2) + (1/\sigma_x^2) + 1/\sigma_{\overline{\mathbf{W}}}^2}$$

and variance

$$\left\{(\beta_x^2 / \sigma_\epsilon^2) + (1/\sigma_x^2) + (1/\sigma_{\overline{\mathbf{W}}}^2)\right\}^{-1}.$$

Notice that the mean of this full conditional distribution for $\mathbf{X}_i$ given everything else depends on $\mathbf{Y}_i$, so that, unlike in regression calibration, $\mathbf{Y}_i$ is used for imputation of $\mathbf{X}_i$.

Now we will find the full conditional for σ_ϵ^2. Recall that the prior is IG$(\delta_{\epsilon,1}, \delta_{\epsilon,2})$, where from Appendix A.3 we know that the $IG(\alpha, \beta)$ distribution has mean $\beta/(\alpha - 1)$ if $\alpha > 1$ and density proportional to $x^{-(\alpha+1)}\exp(-\beta/x)$. Isolating the terms depending on σ_ϵ^2 in the joint density of the observed data and the unknowns, we have

$$\begin{aligned} &f(\sigma_\epsilon^2 | \text{others}) \\ &\qquad \propto (\sigma_\epsilon^2)^{-(\delta_{\epsilon,1} + n/2 + 1)} \exp\left\{\frac{-\delta_{\epsilon,2} + -\frac{1}{2} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i \beta_x - \mathbf{Z}_i^t \beta_z)^2}{\sigma_\epsilon^2}\right\}, \end{aligned}$$

which implies that

$$f(\sigma_\epsilon^2 | \text{others}) = \text{IG}\left[(\delta_{\epsilon,1} + n/2), \left\{\delta_{\epsilon,2} + (1/2) \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i \beta_x - \mathbf{Z}_i^t \beta_z)^2\right\}\right].$$

By similar calculations,

$$f(\sigma_x^2|\text{others}) \propto (\sigma_x^2)^{-(\delta_{x,1}+n/2+1)} \exp\left\{\frac{-\delta_{x,2} - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \mu_x)^2}{\sigma_x^2}\right\},$$

so that

$$f(\sigma_x^2|\text{others}) = \text{IG}\left[(\delta_{x,1} + (n/2)), \left\{\delta_{x,2} + (1/2)\sum_{i=1}^{n}(\mathbf{X}_i - \mu_x)^2\right\}\right].$$

Let $M_J = \sum_{i=1}^{n} k_i/2$. Then we have in addition that

$$f(\sigma_u^2|\text{others})$$
$$\propto (\sigma_u^2)^{-(\delta_{u,1}+M_J+1)} \exp\left\{\frac{-\delta_{u,2} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(\mathbf{W}_{i,j} - \mathbf{X}_i)^2}{\sigma_u^2}\right\},$$

whence

$$f(\sigma_u^2|\text{others}) = \text{IG}\left[(\delta_{u,1} + M_J), \left\{\delta_{u,2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(\mathbf{W}_{i,j} - \mathbf{X}_i)^2\right\}\right].$$

The Gibbs sampler requires a starting value for Ω. For β_x, β_z, and σ_ϵ, one can use estimates from the regression of $\mathbf{Y}_i$ on $\mathbf{Z}_i$ and $\mathbf{X}_i$ (validation data) or $\overline{\mathbf{W}}$ (nonvalidation data). Although there will be some bias, these naive estimators should be in a region of reasonably high posterior probability, and bias should not be a problem since they are being used only as starting values. We start $\mathbf{X}_i$ at $\overline{\mathbf{W}}_i$. Also, μ_x and σ_x can be started at the sample mean and standard deviation of the starting values of the $\mathbf{X}_i$'s. The replication data can be used to find an analysis of variance estimate of σ_u^2 for use as a starting value; see equation (4.3).

9.4.1 Example

We simulated data with the following parameters: $n = 200$, $\beta^t = (\beta_0, \beta_x, \beta_z) = (1, 0.5, 0.3)$, $\alpha^t = (\alpha_0, \alpha_z) = (1, 0.2)$, $\mathbf{X}_i = \alpha_0 + \alpha_z\mathbf{Z}_i + \mathbf{V}_i$, where $\mathbf{V}_i \sim \text{Normal}(0, \sigma_x^2)$ with $\sigma_x = 1$. The $\mathbf{Z}_i$ were independent Normal(1, 1), and since the analysis is conditioned on their values, their mean and variance are not treated as parameters. Also,

$$\mathbf{Y}_i = \beta_0 + \beta_x\mathbf{X}_i + \beta_z\mathbf{Z}_i + \epsilon_i, \tag{9.16}$$

where $\epsilon_i = \text{Normal}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.3$, and $\mathbf{W}_{i,j} = \text{Normal}(\mathbf{X}_i, \sigma_u^2)$, with $\sigma_u^2 = 1$. The observed data are $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_{i,1}, \mathbf{W}_{i,2})$.

We used Gibbs sampling with 10,000 iterations after a burn-in period of 2,000 iterations. The prior parameters were $\sigma_\beta = \sigma_\alpha = 1000$, $\delta_{\epsilon,1} = 3, \delta_{\epsilon,2} = 1$, $\delta_{x,1} = 3, \delta_{x,2} = 1$, and $\delta_{u,1} = 3, \delta_{u,2} = 1$. As discussed in Section A.3, the choice of $\delta_{\epsilon,1} = 3$ and $\delta_{\epsilon,2} = 1$ suggests a prior

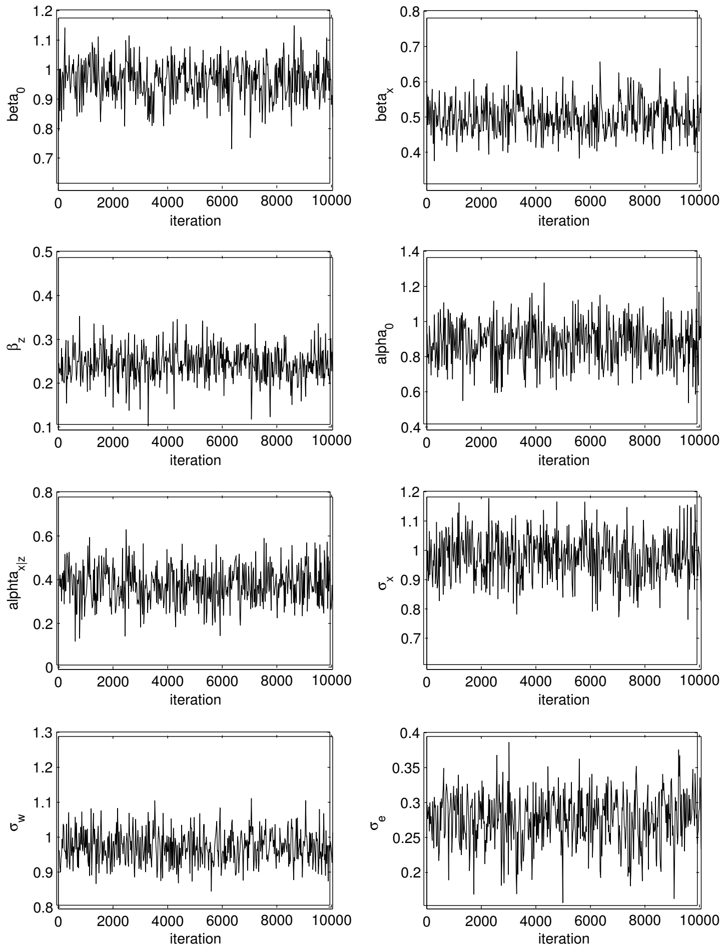


Figure 9.2 *Every 20th iteration of the Gibbs sampler for the linear regression example.*

guess at σ_ϵ^2 of $\delta_{\epsilon,2}/\delta_{\epsilon,1} = 1/3$, and that the prior has the amount of information that would be obtained from $2\delta_{\epsilon,1} = 6$ observations. The same is true of the other δ's. We experimented with other choices of these prior parameters, in particular, smaller values of the effective prior sample size, and found that the posterior was relatively insensitive to the priors, provided that $\delta_{\epsilon,2}$ is not too large.

Starting values for the unobserved covariates were $\mathbf{X}_i = \overline{\mathbf{W}}_i = (\mathbf{W}_{i,1}+$

$\mathbf{W}_{i,2})/2$. The starting values of the parameters were chosen independently: $\sigma_x, \sigma_u, \sigma_\epsilon \sim \text{Uniform}(0.05, 3)$. The starting value for β and α were generated from (9.13) and (9.14).

Figure 9.2 shows every 20th iteration of the Gibbs sampler. These are the so-called trace plots that are used to monitor convergence of the Gibbs sampler, that is, at convergence, they should have no discernible pattern. No patterns are observed, and thus the sampler appears to have mixed well. This subset of the iterations was used to make the plots clearer; for estimation of posterior means and variance, all iterates were used. Using all iterates, the sample autocorrelation for β_x looks like an AR(1) process with a first-order autocorrelation of about 0.7. We used a large number (10,000) of iterations to reduce the potentially high Monte Carlo variability due to autocorrelation.

To study the amount of Monte Carlo error from Gibbs sampling and to see if 10,000 iterations is adequate, the Gibbs sampler was repeated four more times on the same simulated data set but with new random starting values for σ_x, σ_u, and σ_ϵ. The averages of the five posterior means and standard deviations for β_x were 0.4836 and 0.0407. The standard deviation of the five posterior means, which estimates Monte Carlo error, was only 0.00093. Thus, the Monte Carlo error of the estimated posterior means was small relative to the posterior variances, and of course this error was reduced further by averaging the five estimates. The results for the other parameters were similar.

It is useful to compare this Bayesian analysis to a naive estimate that ignores measurement error. The naive estimate from regressing $\mathbf{Y}_i$ on $\overline{\mathbf{W}}_i$ and $\mathbf{Z}_i$ was $\widehat{\beta}_x = 0.346$ with a standard error of 0.0233, so the naive estimator is only about half as variable as the Bayes estimator, but the mean square error of the naive estimator will be much larger and due almost entirely to bias. The estimated attenuation was 0.701, so the bias-corrected estimate was $0.346/0.701 = 0.494$. Ignoring the uncertainty in the attenuation, the standard error of the bias-corrected estimate is $0.0233/0.701 = 0.0322$. This standard error is smaller than the posterior standard deviation but is certainly an underestimate of variability, and if we wanted to use the bias-corrected estimator we would want to use the bootstrap or the sandwich formula to get a better standard error.

In summary, in this example the Bayes estimate of β_x is similar to the naive estimate corrected for attenuation, which coincides with the regression calibration estimate. The Bayes estimator takes more work to program but gives a posterior standard deviation that takes into account uncertainty due to estimating other parameters. The estimator corrected for attenuation would require bootstrapping or some type of asymptotic approximation, for example, the delta-method or the sandwich formula from estimating equations theory, to account for this uncertainty. However, for linear regression, Bayesian MCMC is a bit of overkill. The real strength of Bayesian MCMC is the ability to handle more difficult problems, for example, segmented regression with multiplicative errors, a problem that appears not to have been discussed in the literature but which can be tackled by MCMC in a straightforward manner; see Section 9.1.4.

9.5 Nonlinear Models

The ideas in Section 9.4 can be generalized to complex regression models in $\mathbf{X}$.

9.5.1 A General Model

The models we will study are all special cases of the following general outcome model

$$[\mathbf{Y}_i | \mathbf{X}_i, \mathbf{Z}_i, \beta, \theta, \sigma_\epsilon] = \text{Normal}\{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta), \sigma_\epsilon^2\}, \tag{9.17}$$

where

$$m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta) = \phi(\mathbf{X}_i, \mathbf{Z}_i)^t \beta_1 + \psi(\mathbf{X}_i, \mathbf{Z}_i, \theta)^t \beta_2 \tag{9.18}$$

is a linear function in β_1, β_2 and nonlinear in θ. The functions ϕ and ψ may include nonlinear terms in $\mathbf{X}$ and $\mathbf{Z}$, as well as interactions, and may be scalar or vector valued. When $\psi \equiv 0$, particular cases of model (9.17) include linear and polynomial regression, interaction models, and multiplicative error models. An example of nonlinear component is $\psi(\mathbf{X}_i, \mathbf{Z}_i, \theta) = |\mathbf{X}_i - \theta|_+$, which appears in segmented regression with an unknown break point location. We assume that the other components of the linear model in Section 9.4 remain unchanged and that $\mathbf{X}_i$ is scalar, though this assumption could easily be relaxed. The unknowns in this model are $(\beta, \theta, \sigma_\epsilon, \sigma_u)$, $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, $(\alpha_0, \alpha_z, \sigma_x)$.

In addition to the priors considered in Section 9.4, we consider a general prior $\pi(\theta)$ for θ and assume that all priors are mutually independent. It is easy to check that the full conditionals $f(\alpha|\text{others})$, $f(\sigma_x^2|\text{others})$, and $f(\sigma_u^2|\text{others})$ are unchanged, and that

$$f(\sigma_\epsilon^2|\text{others}) = \text{IG}\left[\delta_{\epsilon,1} + (n/2),\ \delta_{\epsilon,2} + (1/2)\sum_{i=1}^{n}\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2\right].$$

Denoting by $\mathcal{C}(\theta)$ the matrix with i^{th} row

$$\boldsymbol{C}_i^t(\theta) = [\phi(\mathbf{X}_i, \mathbf{Z}_i), \psi(\mathbf{X}_i, \mathbf{Z}_i, \theta)],$$

letting $\beta = (\beta_1^t, \beta_2^t)^t$, and letting $\Delta = \sigma_\epsilon^2/\sigma_\beta^2$, the full conditional for β becomes normal with mean $\{\mathcal{C}(\theta)^t\mathcal{C}(\theta) + \Delta\boldsymbol{I}\}^{-1}\mathcal{C}(\theta)^t\mathbf{Y}$ and covariance matrix $\{\mathcal{C}(\theta)^t\mathcal{C}(\theta) + \Delta\boldsymbol{I}\}^{-1}$.

By grouping together all terms that depend on θ one obtains

$$f(\theta|\text{others}) \propto \exp\left[-\sum_{i=1}^{n}\frac{\{\mathbf{Y}_i^{(1)}-\psi(\mathbf{X}_i,\mathbf{Z}_i,\theta)\beta_2\}^2}{2\sigma_\epsilon^2}\right]\pi(\theta), \tag{9.19}$$

where $\mathbf{Y}_i^{(1)} = \mathbf{Y}_i - \phi(\mathbf{X}_i,\mathbf{Z}_i)\beta_1$. Since ψ is a nonlinear function in θ, this full conditional is generally not in a known family of distributions regardless of how $\pi(\theta)$ is chosen. One can update θ using a random walk MH step using Normal$(\theta, B\sigma_\theta^2)$ as the proposal density, where B is tuned to get a moderate acceptance rate.

The full conditional for $\mathbf{X}_i$ is

$$\begin{aligned} f(\mathbf{X}_i|\text{others}) &\propto \exp\left[-\{\mathbf{Y}_i - m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)\}^2/(2\sigma_\epsilon^2)\right] \qquad (9.20)\\ &\times\exp\left\{(\mathbf{X}_i-\alpha_0-\alpha_z\mathbf{Z}_i)^2/(2\sigma_x^2)+k_i(\overline{\mathbf{W}}_i-\mathbf{X}_i)^2/(2\sigma_u^2)\right].\end{aligned}$$

To update $\mathbf{X}_i$, we use a random walk MH step with Normal$(\mathbf{X}_i, B\,\sigma_u^2/k_i)$ with the "dispersion" factor, B, chosen to provide a reasonable acceptance rate.

We now discuss the details of implementation for polynomial, multiplicative measurement error and segmented regression.

9.5.2 Polynomial Regression

A particular case of the outcome model (9.17) is the polynomial regression in $\mathbf{X}$

$$\mathbf{Y}_i = \mathbf{Z}_i^t\beta_z + \mathbf{X}_i\beta_{x,1} + \cdots + \mathbf{X}_i^p\beta_{x,p} + \epsilon_i, \tag{9.21}$$

for some $p > 1$, where ϵ_i are independent Normal$(0,\sigma_\epsilon^2)$, obtained by setting $\phi(\mathbf{X}_i,\mathbf{Z}_i) = (\mathbf{Z}_i^t,\mathbf{X}_i,\ldots,\mathbf{X}_i^p)$ and $\psi(\mathbf{X}_i,\mathbf{Z}_i,\theta) = 0$. The i^{th} row of $\mathcal{C} := \mathcal{C}(\theta)$ is $\boldsymbol{C}_i^t = \phi(\mathbf{X}_i,\mathbf{Z}_i)$ and $\beta = (\beta_z^t,\beta_{x,1},\ldots,\beta_{x,p})^t$. With this notation, all full conditionals are as described in Section 9.5.1. In particular, the full conditional of θ in (9.19) is not necessary because $\psi = 0$. In this example, the full conditional for $\mathbf{X}_i$ is the only nonstandard distribution and can be obtained as a particular case of (9.20) as

$$\begin{aligned} f(\mathbf{X}_i|\text{others}) &\propto \exp\left\{-(\mathbf{Y}_i - \boldsymbol{C}_i^t\beta)^2/(2\sigma_\epsilon^2)\right\} \qquad (9.22)\\ &\times\exp\left\{-(\mathbf{X}_i-\alpha_0-\mathbf{Z}_i^t\alpha_z)^2/(2\sigma_x^2)-k_i(\overline{\mathbf{W}}_i-\mathbf{X}_i)^2/(2\sigma_u^2)\right\}.\end{aligned}$$

The full conditional for $\mathbf{X}_i$ is nonstandard because $\boldsymbol{C}_i$ contains powers of $\mathbf{X}_i$.

To illustrate these ideas, consider the quadratic regression in $\mathbf{X}$

$$\mathbf{Y}_i = \beta_0 + \beta_{x,1}\mathbf{X}_i + \beta_{x,2}\mathbf{X}_i^2 + \beta_z\mathbf{Z}_i + \epsilon_i, \tag{9.23}$$

with $\beta_{x,2} = 0.2$ and the other parameters unchanged. To update $\mathbf{X}_i$ the proposal density was Normal$(\mathbf{X}_i, B\,\sigma_u^2/k_i)$. After some experimentation, the "dispersion" factor B was chosen to be 1.5 to get approximately 25% acceptance. We found that the performance of the Gibbs sampler was not particularly sensitive to the value of B, and B equal to 1 or 2.5 also worked well.

As in the linear example, we used five runs of the Gibbs sampler, each with 10,000 iterations, and with the same starting value distribution as before. The posterior means of β_0, $\beta_{x,1}$, $\beta_{x,2}$, and β_z were 1.015, 0.493 0.191, and 0.348, close to the true values of the parameters, which were 1.0, 0.5, 0.2, and 0.3. In contrast, the naive estimates obtained by fitting (9.23) with $\mathbf{X}_i$ replaced by $\overline{\mathbf{W}}_i$ were 1.18, 0.427, 0.104, and 0.394, so, in particular, the coefficient of $\mathbf{X}^2$ was biased downward by nearly 50%. The posterior standard deviations were 0.057, 0.056, 0.027, and 0.040, while the standard errors of the naive estimates were 0.079, 0.052, 0.021, and 0.049.

9.5.3 Multiplicative Error

We now show that a linear regression model (9.7) with multiplicative measurement error is a particular case of model (9.17). As discussed in Section 4.5, this model is relatively common in applications. Indeed, if $\mathbf{X}_i^* = \log(\mathbf{X}_i)$ and $\mathbf{W}_{i,j}^* = \log(\mathbf{W}_{i,j})$ then the outcome model becomes

$$Y_i = \mathbf{Z}_i^t\beta_z + e^{\mathbf{X}_i^*}\beta_x + \epsilon_i,$$

which can be obtained from (9.17) by setting $\phi(\mathbf{X}_i^*,\mathbf{Z}_i) = (\mathbf{Z}_i^t, e^{\mathbf{X}_i^*})$ and $\psi(\mathbf{X}_i^*,\mathbf{Z}_i,\theta) = 0$. The i^{th} row of $\mathcal{C} := \mathcal{C}(\theta)$ is $\boldsymbol{C}_i^t = \phi(\mathbf{X}_i^*,\mathbf{Z}_i)$ and $\beta = (\beta_z^t,\beta_x)^t$.

We replace the exposure model (9.9) by a lognormal exposure model where (9.24) holds with $\mathbf{X}_i$ replaced by

$$\mathbf{X}_i^* \sim \text{Normal}(\alpha_0 + \mathbf{Z}_i^t\alpha_z, \sigma_x^2). \tag{9.24}$$

The measurement model is

$$[\mathbf{W}_{i,j}^*|\mathbf{X}_i] \sim \text{Normal}(\mathbf{X}_i^*, \sigma_u^2),\ j = 1,\ldots,k_i,\ i = 1,\ldots,n. \tag{9.25}$$

With this notation, the full conditionals for this model are the same as in Section 9.5.1. One trivial change is that $\mathbf{X}_i$ is replaced everywhere by $\mathbf{X}_i^*$, and the full conditional of θ is not needed because $\psi = 0$.

To illustrate these ideas, we simulated 200 observations with $\beta_0 = 1$, $\beta_x = 0.3$, $\beta_z = 0.3$, $\alpha_0 = 0$, $\alpha_z = 0.2$, $\sigma_x = 1$, and $\sigma_u = 1$. The $\mathbf{Z}_i$ were Normal$(-1,1)$. We ran the Gibbs sampler with tuning parameter $B = 2.5$, which gave a 30% acceptance rate. Figure 9.3 shows the output from one of five runs of the Gibbs sampler. There were 10,500 iterations, of which the first 500 were discarded. One can see that β_0 and, especially, β_x mix more slowly than the other parameters, yet their mixing seems adequate. In particular, the standard deviation of the five

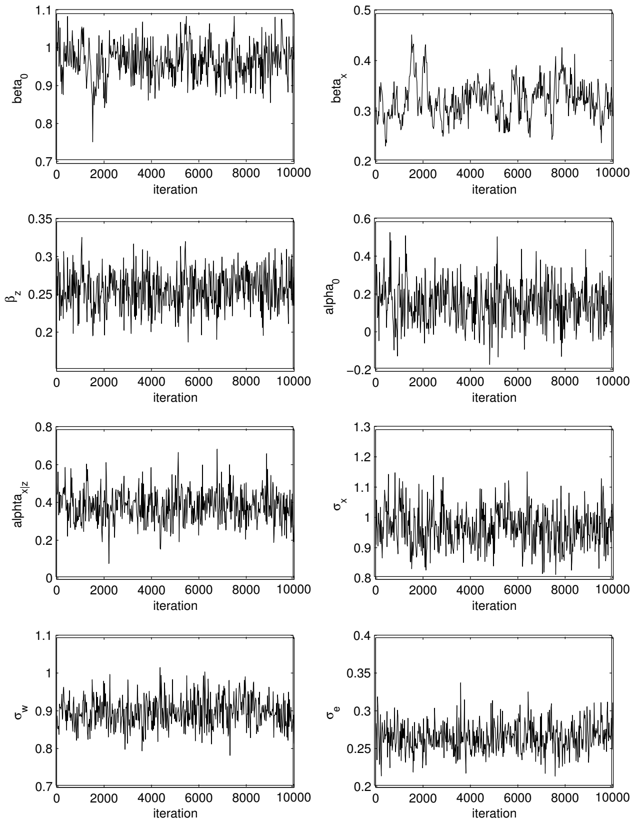


Figure 9.3 *Every 20th iteration of the Gibbs sampler for the linear regression example with multiplicative error.*

posterior means for β_x was 0.0076 giving a Monte Carlo standard error of $0.0078/\sqrt{5} = 0.0034$, while the posterior standard deviation of that parameter was 0.0377, about 10 times larger than the Monte Carlo standard error.

9.5.4 Segmented Regression

A commonly used regression model is a segmented line, that is, two lines joined together at a knot. This model can be written as

$$\mathbf{Y}_i = \mathbf{Z}_i^t\beta_z + \beta_{x,1}\mathbf{X}_i + \beta_{x,2}(\mathbf{X}_i - \theta)_+ + \epsilon_i, \tag{9.26}$$

where we use the notation $a_+ = \min(0,\ a)$, θ is the knot, $\beta_{x,1}$ is the slope of $\mathbf{Y}$ on $\mathbf{X}$ before the knot, and $\beta_{x,2}$ is the change in this slope at the knot. An intercept could be included in $\mathbf{Z}_i^t\beta_z$.

The outcome model (9.26) is a particular case of model (9.17) with $\phi(\mathbf{X}_i, \mathbf{Z}_i) = (\mathbf{Z}_i^t, \mathbf{X}_i)$ and $\psi(\mathbf{X}_i, \mathbf{Z}_i, \theta) = (\mathbf{X}_i - \theta)_+$. The i^{th} row of $\mathcal{C}(\theta)$ is $\boldsymbol{C}_i^t(\theta) = \{\mathbf{Z}_i^t, \mathbf{X}_i, (\mathbf{X}_i - \theta)_+\}^t$ and $\beta = (\beta_z^t, \beta_{x,1}, \beta_{x,2})^t$. With this notation, all full conditionals are as described in Section 9.5.1.

To illustrate segmented regression with measurement error and unknown knot location we simulated data with $n = 200$, $J = 2$, $\beta_0 = 1$, $\beta_x = 1$, $\beta_{x,2} = 0.8$, $\beta_z = 0.1$, $\theta = 1$, $\alpha_0 = 1$, $\alpha_z = 0$, $\sigma_\epsilon = 0.15$, $\sigma_x = 1$, and $\sigma_u = 1$. The $\mathbf{Z}_i$ were Normal$(1, 1)$. Since $\alpha_z = 1$, the $\mathbf{X}_i$ were Normal$(1, 1)$ independently of the $\mathbf{Z}_i$.

We ran the Gibbs sampler 5 times, each with 10,000 iterations. Starting values for θ were Uniform$(0.5,\ 1.5)$. In the prior for θ, we used the Normal$(\mu_\theta, \sigma_\theta^2)$ distribution with $\mu_\theta = \overline{\mathbf{W}}$ and $\sigma_\theta = 5\,s(\overline{\mathbf{W}})$, where $s(\overline{\mathbf{W}})$ was the sample standard deviation of $\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_n$. This prior was designed to have high prior probability over the entire range of observed values of $\mathbf{W}$. In the proposal density for θ, we used $B = 0.01$. This value was selected by trial and error and gave an acceptance rate of 36% and adequate mixing. The posterior mean and standard deviation of θ were 0.93 and 0.11, respectively. The Monte Carlo standard error of the posterior mean was only 0.005.

Figure 9.4 reveals how well the Bayesian modeling imputes the $\mathbf{X}_i$ and leads to good estimates of θ. The top left plot shows the true $\mathbf{X}_i$ plotted with the $\mathbf{Y}_i$. The bottom right plot is similar, except that instead of the unknown $\mathbf{X}_i$ we use the imputed $\mathbf{X}_i$ from the 10,000th iteration of the fifth run of the Gibbs sampler. Notice that the general pattern of $\mathbf{X}$ versus $\mathbf{Y}$ is the same for the true and the imputed $\mathbf{X}_i$. In contrast, a plot of $\mathbf{Y}_i$ and either $\overline{\mathbf{W}}_i$ or $\widehat{E}(\mathbf{X}_i|\overline{\mathbf{W}}_i) = (1-\widehat{\lambda})\overline{\mathbf{W}}_i + \widehat{\lambda}\overline{\mathbf{W}}_i$ shows much less similarity with the $(\mathbf{X}_i, \mathbf{Y}_i)$ plot. Here, $\widehat{\lambda}$ is the estimated attenuation and $\overline{\mathbf{W}}$ is the mean of $\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_n$.

The plot of the imputed $\mathbf{X}_i$ versus $\mathbf{Y}_i$ shows the existence and location of the knot quite clearly, and it is not surprising that θ can be estimated with reasonably accuracy. Of course, this "feedback" of information about the $\mathbf{X}_i$ to information about θ works both ways. Accurate knowledge of θ well helps impute the $\mathbf{X}_i$. One estimates both the $\mathbf{X}_i$ and θ well in this example because their joint posterior has highest probability near their true values.

9.6 Logistic Regression

In this section, we assume the same model with nonlinear measurement error as in Section 9.5 but with a binary outcome. We use the logistic

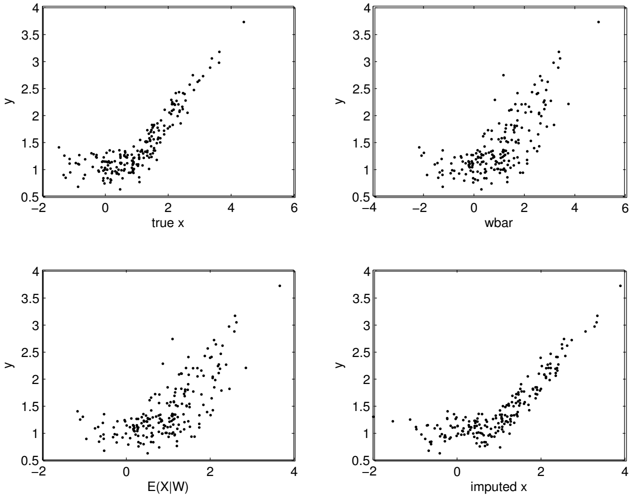


Figure 9.4 *Segmented regression. Plots of* $\mathbf{Y}_i$ *and* $\mathbf{X}_i$ *and three estimator of* $\mathbf{X}_i$. *Top left:* $\mathbf{Y}$ *plotted versus the true* $\mathbf{X}$. *Top right:* $\mathbf{Y}$ *plotted versus the mean of the replicated* $\mathbf{W}$*-values. Bottom left:* $\mathbf{Y}$ *plotted versus the regression calibration estimates of* $\mathbf{X}$. *Bottom right:* $\mathbf{Y}$ *plotted versus the imputed* $\mathbf{X}$ *in a single iteration of the Gibbs sampler. Note how the Gibbs sampler more faithfully reproduces the true* $\mathbf{X}$*-values.*

regression model

$$\log\left\{\frac{P(\mathbf{Y}_i=1|\mathbf{X}_i,\mathbf{Z}_i)}{P(\mathbf{Y}_i=0|\mathbf{X}_i,\mathbf{Z}_i)}\right\} = m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta),$$

so the outcome likelihood is proportional to

$$\exp\left[\sum_{i=1}^{n}\mathbf{Y}_i m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta) - \sum_{i=1}^{n}\log\left\{1+e^{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)}\right\}\right],$$

$$[\beta,\theta|\text{others}] \propto \exp\left[\sum_{i=1}^{n}\mathbf{Y}_i m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta) - \sum_{i=1}^{n}\log\left\{1+e^{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)}\right\} - \frac{\beta^t\beta}{2\sigma_\beta^2}\right]\pi(\theta), \qquad (9.27)$$

and

$$[\mathbf{X}_i|\text{others}] \propto \exp\left[\sum_{i=1}^{n}\mathbf{Y}_i m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta) - \sum_{i=1}^{n}\log\left\{1+e^{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)}\right\}\right.$$

$$\left. + \frac{(\mathbf{X}_i-\alpha_0-\alpha_z\mathbf{Z}_i)^2}{\sigma_x^2} + \frac{(\overline{\mathbf{W}}_i-\mathbf{X}_i)^2}{\sigma_{\overline{\mathbf{W}}}^2}\right]. \qquad (9.28)$$

To update $\mathbf{X}_i$ we use a random-walk MH step with the same Normal($\mathbf{X}_i$, $B\,\sigma^2_{\overline{\mathbf{W}}}$) proposal density as for polynomial regression. To update β we use a random-walk MH step with proposal density $N\{\beta,\, B'\text{var}(\widehat{\beta})\}$, where $\text{var}(\widehat{\beta})$ is the covariance matrix of the naive logistic regression estimator using $\overline{\mathbf{W}}$ in place of $\mathbf{X}$ and B' is another tuning constant. A similar strategy may be applied to update θ when ψ in (9.18) is not identically zero.

To illustrate the fitting algorithms for logistic regression with measurement error, we simulated data from a quadratic regression similar to the one in Section 9.5.2 but with a binary response following the logistic regression model. The intercept β_0 was changed to -1 so that there were roughly equal numbers of 0s and 1s among the $\mathbf{Y}_i$. Also, the sample size was increased to $n = 1,500$ to ensure reasonable estimation accuracy for β. Otherwise, the parameters were the same as the example in Section 9.5.2. The tuning parameters in the MH steps were $B = B' = 1.5$. This gave acceptance rates of about 52% for the $\mathbf{X}_i$ and about 28% for β.

Figure 9.5 show the output from one of the five runs of the Gibbs sampler. The samplers appear to have converged and to have mixed reasonably well. The posterior mean of β was $(-1.18, 0.55, 0.24, 0.30)$, which can be compared to $\beta = (-1, 0.5, 0.2, 0.3)$. The posterior standard deviations were $(0.13, 0.17, 0.09, 0.06)$. The Monte Carlo error, as measured by the between-run standard deviations of the posterior means, was less than one-tenth as large as the posterior standard deviations.

9.7 Berkson Errors

The Bayesian analysis of Berkson models is similar to, but somewhat simpler than, the Bayesian analysis of error models. The reason for the simplicity is that we need a Berkson error model only for $[\mathbf{X}|\mathbf{W}]$ or $[\mathbf{X}|\mathbf{W},\mathbf{Z}]$. If, instead, we had an error model $[\mathbf{W}|\mathbf{X},\mathbf{Z}]$ then, as we have seen, we would also need a structural model $[\mathbf{X}|\mathbf{Z}]$.

We will consider nonlinear regression with a continuously distributed $\mathbf{Y}$ first and then logistic regression.

9.7.1 Nonlinear Regression with Berkson Errors

Suppose that we have outcome model (9.17), which for the reader's convenience is

$$[\mathbf{Y}_i|\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta,\sigma_\epsilon] = \text{Normal}\{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta),\sigma_\epsilon^2\}, \qquad (9.29)$$

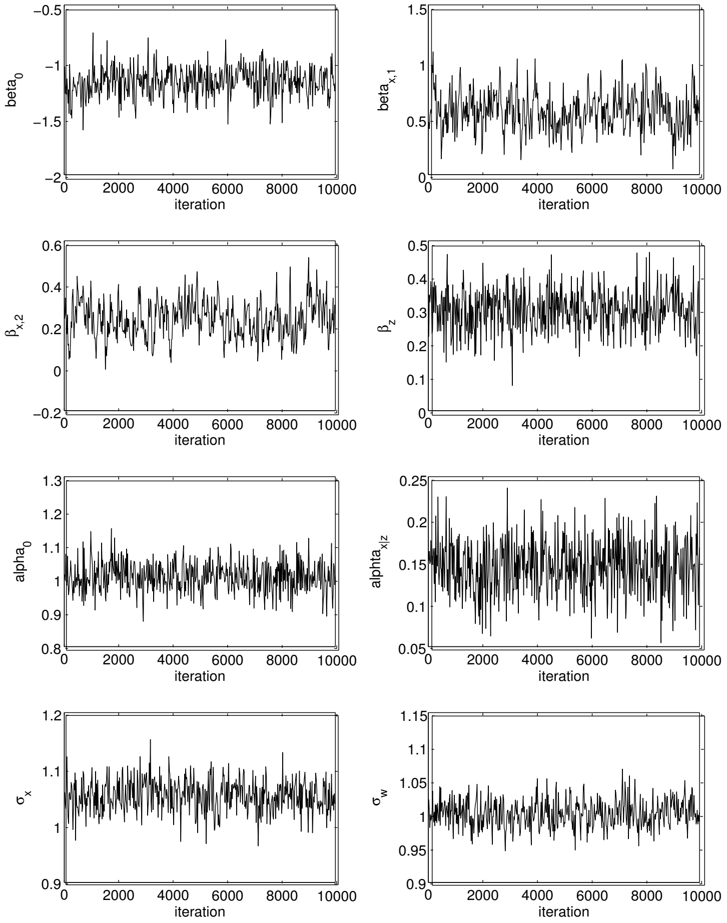


Figure 9.5 *Every 20th iteration of the Gibbs sampler for the quadratic logistic regression example.*

but now with Berkson error so that we observe $\mathbf{W}_i$ where

$$\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i, \;\; E(\mathbf{U}_i|\mathbf{Z}_i, \mathbf{W}_i) = 0.$$

Model (9.29) is nonlinear in general, but includes linear models as a special case. The analysis in Section 9.5.1, which was based upon replicated classical measurement error and a structural model that says that $\mathbf{X}|\mathbf{Z} \sim \text{Normal}(\alpha_0 + \alpha_z \mathbf{Z})$, must be changed slightly because of the Berkson errors. The only full conditionals that change are for the $\mathbf{X}_i$.

Specifically, equation (9.20), which is

$$f(\mathbf{X}_i|\text{others}) \propto \exp\left[-\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2/(2\sigma_\epsilon^2)\right]$$
$$\times \exp\left\{-(\mathbf{X}_i - \alpha_0 - \alpha_z \mathbf{Z}_i)^2/(2\sigma_x^2) - k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right\},$$

is modified to

$$f(\mathbf{X}_i|\text{others}) \propto \exp\left[-\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2/(2\sigma_\epsilon^2)\right] \quad (9.30)$$
$$\times \exp\left\{-(\mathbf{W}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right].$$

Thus, we see two modifications. The term $-(\mathbf{X}_i - \alpha_0 - \alpha_z \mathbf{Z}_i)^2/(2\sigma_x^2)$ in (9.20), which came from the structural assumption, is not needed and $k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2$ is replaced by $(\mathbf{W}_i - \mathbf{X}_i)^2$ since there are no replicates in the Berkson model. That's it for changes—everything else is the same!

This analysis illustrates a general principle, which may have been obvious to the reader, but should be emphasized. When we have a Berkson model that gives $[\mathbf{X}|\mathbf{Z}, \mathbf{W}]$, we do not need a model for marginal density $[\mathbf{W}]$ of $\mathbf{W}$. The $\mathbf{W}_i$ are observed so that we can condition upon them. In contrast, if we have an error model for $[\mathbf{W}|\mathbf{Z}, \mathbf{X}]$, we cannot do a conditional analysis given the $\mathbf{X}_i$ since these are unobserved, and therefore a structural model for $[\mathbf{X}]$ or, perhaps, $[\mathbf{X}|\mathbf{Z}]$ is also needed.

9.7.2 Logistic Regression with Berkson Errors

When errors are Berkson, the analysis of a logistic regression model described in Section 9.6 changes in a way very similar to the changes just seen for nonlinear regression. In particular, equation (9.28), which is

$$[\mathbf{X}_i|\text{others}] \propto \exp\left[\sum_{i=1}^n \mathbf{Y}_i m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta) - \sum_{i=1}^n \log\left\{1 + e^{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)}\right\}\right.$$
$$\left. + \frac{(\mathbf{X}_i - \alpha_0 - \alpha_z \mathbf{Z}_i)^2}{\sigma_x^2} + \frac{(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2}{\sigma_{\mathbf{W}}^2}\right],$$

becomes

$$[\mathbf{X}_i|\text{others}] \propto \exp\left[\sum_{i=1}^n \mathbf{Y}_i m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta) - \sum_{i=1}^n \log\left\{1 + e^{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)}\right\}\right.$$
$$\left. + \frac{(\mathbf{W}_i - \mathbf{X}_i)^2}{\sigma_u^2}\right]. \qquad (9.31)$$

As before, the term $(\mathbf{X}_i - \alpha_0 - \alpha_z \mathbf{Z}_i)^2/\sigma_x^2$ in (9.28) came from the structural model and is not needed for a Berkson analysis, and $\overline{\mathbf{W}}_i$ is replaced by $\mathbf{W}_i$ because there is no replication.

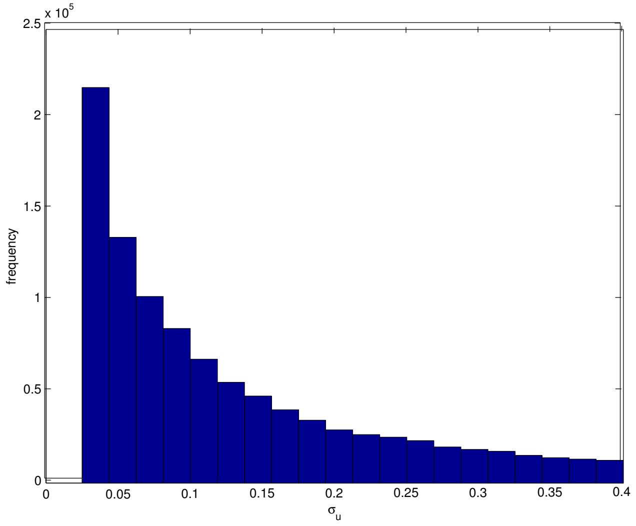


Figure 9.6 *Munich bronchitis data. Histogram of 1,250,000 samples from the posterior for σ_u.*

9.7.3 Bronchitis Data

We now continue the analysis of the bronchitis data described in Section 8.7. Recall that in that section we found that the MLE of the Berkson measurement error standard deviation, σ_u, was zero. Our Bayesian analysis will show that σ_u is poorly determined by the data. Although σ_u is theoretically identifiable, for practical purposes it is not identified. Gustafson (2005) has an extensive discussion of nonidentified models. He argues in favor of using informative priors on nonidentified nuisance parameters, such as σ_u here. The following analysis applies Gustafson's strategy to σ_u.

We will use a Uniform (0.025, 0.4) prior for σ_u. This prior seems reasonable, since σ_w is 0.72, so the lower limit of the prior implies very little measurement error. Also, the upper limit is over twice the value, 0.187, assumed in previous work by Gössi and Küchenhoff (2001). We will use a Uniform $\{1.05 \min(W_i),\, 0.95 \max(W_i)\}$ prior for $\beta_{x,2}$. This prior is reasonable since $\beta_{x,2}$ is a TLV (threshold limiting value) within the range of the observed data. The prior on β, the vector of all regression coefficient, is Normal$(0, 10^6 I)$.

There were five MCMC runs, each of 250,000 iterations excluding a burn-in of 1,000 iterations. Figure 9.6 is a histogram of the 1,250,000 values of σ_u^2 from the five runs combined. The posterior is roughly proportional to the likelihood, since there are uniform priors on σ_u and $\beta_{x,2}$ and a very diffuse prior on β. The histogram is monotonically decreasing, in agreement with the MLE of 0 for σ_u. However, the posterior is very diffuse and much larger values of σ_u are plausible under the posterior. In fact, the posterior mean, standard deviation, 0.025 quantile, and 0.975 quantile of σ_u were 0.13, 0.098, 0.027, and 0.37, respectively. The 95% credible interval of (0.027, 0.37) is not much different from (0.0344, 0.3906), the interval formed by the 2.5 and 97.5 percentiles of the prior. Thus, the data provide some, but not much, information about σ_u.

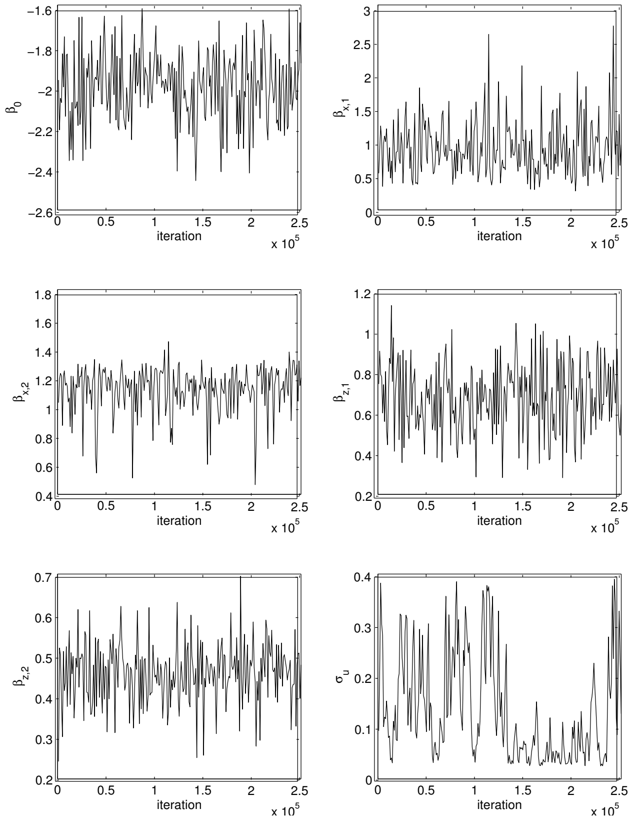


Figure 9.7 *Trace plots for the Munich bronchitis data.*

Figure 9.7 shows trace plots for the first of the five MCMC runs. Trace plots for the other runs are similar. The mixing for σ_u is poor, but the mixing for the other parameters is much better. The poor mixing of σ_u

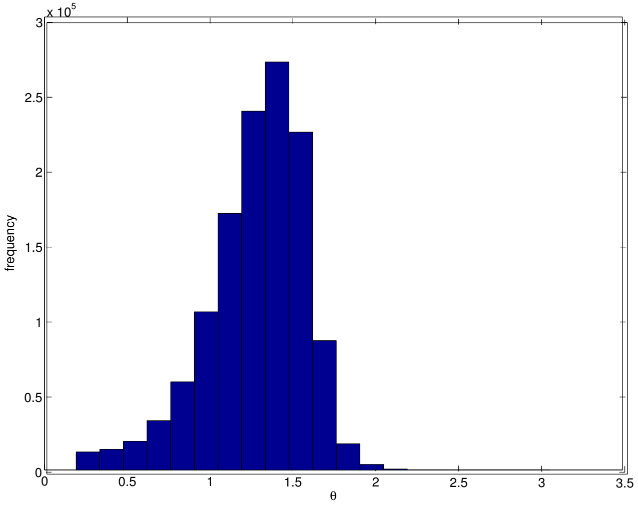


Figure 9.8 *Munich bronchitis data. Histogram of 1,250,000 samples from the posterior for TLV, $\beta_{x,2}$.*

was the reason we used 250,000 iterations per run rather than a smaller value, such as 10,000, which was used in previous examples.

We experimented with a Uniform$(0, 10)$ prior for σ_u and encountered difficulties. On some runs, the sampler would get stuck at $\sigma_u = 0$ and $\mathbf{X}_i = \mathbf{W}_i$ for all i. On runs where this problem did not occur the mixing was very poor for σ_u, and fair to poor for the other parameters. We conclude that a reasonably informative prior on σ_u is necessary. However, fixing σ_u at a single value, as Gössi and Küchenhoff (2001) have done, is not necessary.

Figure 9.8 is a histogram of the 1,250,000 value of $\beta_{x,2}$ from the combined runs with burn-ins excluded. The posterior mean of $\beta_{x,2}$ was 1.28, very close to the naive of 1.27 found in Section 8.7. This is not surprising, since the simulations in Section 8.7.3 showed that the naive estimator had only a slight negative bias. The 95% highest posterior density credible interval was (0.53, 1.73).

9.8 Automatic Implementation

Bayesian analysis for complex models with covariates measured with error needs to be based on carefully constructed prior, full conditional, and proposal distributions combined with critical examination of the convergence and mixing properties of the Markov chains. The MATLAB programs used in the previous sections are specially tailored and optimized to address these issues. However, standard software, such as WinBUGS, may prove to be a powerful additional tool in applications where many models are explored. We now show how to use WinBUGS for fitting models introduced in Sections 9.4 and 9.5.

9.8.1 Implementation and Simulations in WinBUGS

We describe in detail the implementation of the linear model in Section 9.4 and note only the necessary changes for the more complex models. The complete commented code presented in Appendix B.8.1 follows step by step the model description in Section 9.4.

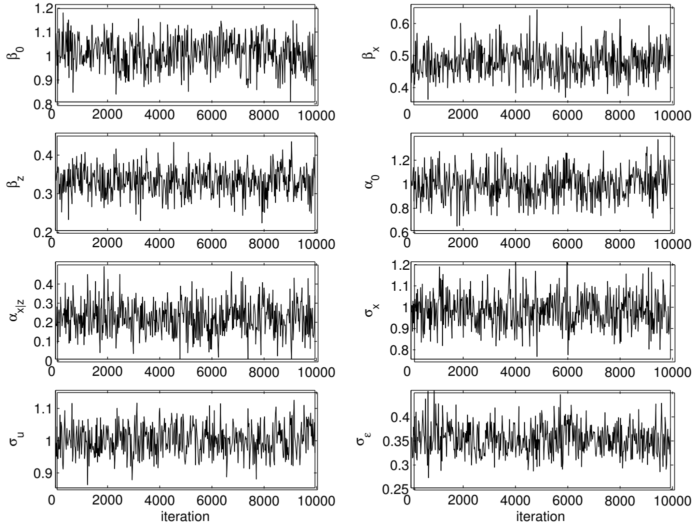


Figure 9.9 *Every 20th iteration for the WinBUGS Gibbs sampler for the linear regression example.*

The first `for` loop specifies the outcome, measurement, and exposure model (9.7), (9.8), and (9.9). Note that `Nobservations` is the sample size and that the # sign indicates a comment. The code is structured and intuitive. For example, the two lines in the outcome model

```
Y[i]~dnorm(meanY[i],taueps)
meanY[i]<-beta[1]+beta[2]*X[i]+beta[3]*Z[i]
```

specify that the outcome of the i^{th} subject, $\mathbf{Y}_i$, has a normal distribution with mean $m_Y(i) = \beta_1 + \beta_2\mathbf{X}_i + \beta_3\mathbf{Z}_i$ and precision parameter $\tau_\epsilon = 1/\sigma_\epsilon^2$. It is quite common in Bayesian analysis to specify the normal distribution in terms of its precision instead of its variance.

The nested `for` loop corresponding to the replication model

```
for (j in 1:Nreplications) {W[i,j]~dnorm(X[i],tauu)}
```

specifies that, conditional on the unobserved exposure, $\mathbf{X}_i$, of the i^{th} subject the proxies $\mathbf{W}_{i,j}$ are normally distributed with mean $\mathbf{X}_i$ and precision $\tau_u = 1/\sigma_u^2$. Here `Nreplications` is the number of replications and it happened to be the same for all subjects. A different number of replications could easily be accommodated by replacing the scalar `Nreplications` by a vector `Nreplications[]`.

The code corresponding to the measurement error model

```
X[i]~dnorm(meanX[i],taux)
meanX[i]<-alpha[1]+alpha[2]*Z[i]
```

specifies that the exposure of the i^{th} subject, $\mathbf{X}_i$, has a normal distribution with mean $\alpha_1 + \alpha_2\mathbf{Z}_i$ and precision parameter $\tau_x = 1/\sigma_x^2$.

The code for prior distributions

```
tauu~dgamma(3,1)
taueps~dgamma(3,1)
taux~dgamma(3,1)
```

specifies that the precision parameters $\tau_u, \tau_\epsilon, \tau_x$ have independent Gamma priors with parameters 3 and 1. The `dgamma(a,b)` notation in WinBUGS specifies a Gamma distribution with mean a/b and variance a/b^2. The code for prior distributions

```
for (i in 1:nalphas){alpha[i]~dnorm(0,1.0E-6)}
for (i in 1:nbetas){beta[i]~dnorm(0,1.0E-6)}
```

specifies that the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3$ have independent normal priors with mean zero and precision 10^{-6}. Here `nalphas` and `nbetas` denote the number of α and β parameters.

The last part of the code contains only definitions of explicit functions of the model parameters. For example,

```
sigmaeps<-1/sqrt(taueps)
sigmau<-1/sqrt(tauu)
sigmax<-1/sqrt(taux)
```

define the standard deviations $\sigma_\epsilon = 1/\sqrt{\tau_\epsilon}$, $\sigma_u = 1/\sqrt{\tau_u}$ and $\sigma_x = 1/\sqrt{\tau_x}$ for the outcome, replication and exposure models, respectively, and

```
lambda<-tauu/(tauu+taux)
```

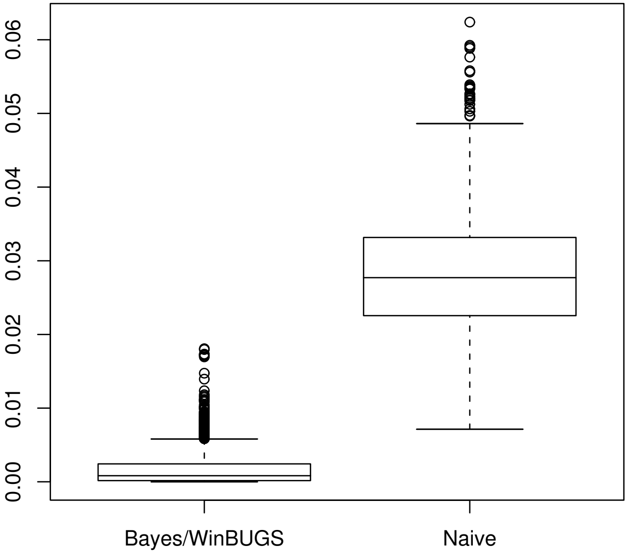


Figure 9.10 *Squared error for the Bayes and naive methods for estimating the exposure effect β_x in the linear model with measurement error in the model (9.16).*

defines the reliability ratio $\lambda = \tau_u/(\tau_u + \tau_x) = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$.

To assess the quality of inference based on the WinBUGS program, we simulated 2,000 data sets from the linear model with measurement error described in Section 9.4.1. For each data set we used 10,500 simulations based on the WinBUGS program and we discarded the first 500 simulations as burn in.

Figure 9.9 shows every 20th iteration of the Gibbs sampler for one data set, indicating that the mixing properties are comparable to those shown in Figure 9.2. However, this is not always the case and WinBUGS programs typically need 10 to 100 times more simulations than expert programs to achieve comparable estimation accuracy. Of course, the time saved by using WinBUGS instead of writing a program often compensates for the extra computational time.

Figure 9.10 displays the squared error of the posterior mean of the exposure effect β_x using Bayes and naive estimators for the linear model with measurement error introduced in Section 9.4. More precisely, for the d^{th} data set, $d = 1, \ldots, 2000$, denote by $\widehat{\beta}_{x,d}^{(B)}$ the posterior mean of β_x using the WinBUGS program and by $\widehat{\beta}_{x,d}^{(N)}$ the MLE of β_x in a standard linear regression, where $\mathbf{X}_i$ is replaced by $\overline{\mathbf{W}}_i = (\mathbf{W}_{i1} + \mathbf{W}_{i2})/2$. Then, the two boxplots in Figure 9.9 correspond to $(\widehat{\beta}_{x,d}^{(B)} - \beta_x)^2$ and $(\widehat{\beta}_{x,d}^{(N)} - \beta_x)^2$, respectively.

We also calculated the coverage probabilities of β_x by the 90% and 95% equal-tail probability credible intervals obtained from the Bayesian analysis based on MCMC simulations implemented in WinBUGS. The true value of the parameter β_x was covered for 89.5% and 94.6% of the data sets by the 90% and 95% credible intervals, respectively. In contrast, the true value of β_x was never covered by the 95% confidence interval of the naive analysis because of its bias.

9.8.2 More Complex Models

Only minor changes are necessary to fit the quadratic polynomial regression model in Section 9.5.2. Indeed, the only change is that the specification of the mean function of the outcome model becomes

```
meanY[i]<-beta[1]+beta[2]*X[i]+beta[3]*pow(X[i],2)
          +beta[4]*Z[i]
```

while the number of β parameters in the data `nbetas` is changed from 3 to 4. Here `pow(X[i],2)` represents $\mathbf{X}_i^2$.

As discussed in Section 9.5.3, the multiplicative measurement error model is equivalent with an additive measurement error model using a log exposure scale. This can be achieved by the transformations $\mathbf{W}_{i,j}^* = \log(\mathbf{W}_{i,j})$ and $\mathbf{X}_i^* = \log(\mathbf{X}_i)$. From a notational perspective in WinBUGS, there is no need to use the $\mathbf{X}_i^*$ notation instead of the $\mathbf{X}_i$ as long as the data is transformed accordingly. Therefore, the only necessary change is that the mean function of the outcome model becomes

```
meanY[i]<-beta[1]+beta[2]*exp(X[i])+beta[3]*Z[i]
```

where `exp(X[i])` represents $e^{\mathbf{X}_i^*}$ and `W[i,j]` represents $\mathbf{W}_{i,j}^*$.

To fit the segmented regression model in Section 9.5.4, one needs to change the mean function of the outcome model to

```
meanY[i]<-beta[1]+beta[2]*X[i]
          +beta[3]*(X[i]-theta)*step(X[i]-theta)
          +beta[4]*Z[i]
```

where `(X[i]-theta)*step(X[i]-theta)` represents $(\mathbf{X}_i - \theta)_+$ because `step(a)` in WinBUGS is equal to a if $a > 0$ and 0 otherwise. One needs only to add the prior for θ:

```
theta~dnorm(barWbar,prec.theta)
```

where `barWbar` represents the average of all W_{ij} observations and `prec.theta` represents $1/(25\sigma^2_{\overline{\mathbf{W}}})$ and are part of the data.

WinBUGS uses a rather inefficient simulation algorithm for fitting complex measurement error models. This is most probably due to the sampling scheme, which updates one parameter at a time and does not take advantage of the explicit full conditionals of groups of parameters. For example, if $\gamma = (\gamma_1, \gamma_2)^t$ has a full conditional Normal$(\mu_\gamma, \Sigma_\gamma)$ with a very strong posterior correlation, it is much more efficient to sample directly from Normal$(\mu_\gamma, \Sigma_\gamma)$ than to sample γ_1 given γ_2 and the others and then γ_2 given γ_1 and the others.

Therefore, the mixing properties of the Markov chains generated by WinBUGS should be carefully analyzed using multiple very long chains. We also found that simple reparameterizations, such as centering and orthogonalization of covariates, can substantially improve mixing.

While we encourage development, when feasible, of expert programs along the lines described in Sections 9.4 and 9.5, WinBUGS can be a valuable additional tool. The main strengths of WinBUGS are

1. Flexibility: Moderate model changes correspond to simple program changes.
2. Simplicity: Program follows almost literally the statistical model.
3. Robustness: Program is less prone to errors.
4. Operability: Programs can be called from different environments, such as R or MATLAB.

The main weakness of WinBUGS is that chains may exhibit very poor mixing properties when parameters have high posterior correlations. This problem may be avoided by expert programs through the careful study of full conditional distributions.

9.9 Cervical Cancer and Herpes

So far in this chapter, we have assumed that a continuously distributed covariate is measured with error. However, Bayesian analysis is straightforward when a discrete covariate is misclassified.

In this section, we continue the analysis given in Section 8.4 of the cervical cancer data discussed in Section 1.6.10. In particular, we continue the retrospective parameterization in Section 8.4 using $\alpha_{xd} = \Pr(\mathbf{W} = 1|\mathbf{X} = x, \mathbf{Y} = d)$ and $\gamma_d = \Pr(\mathbf{X} = 1|\mathbf{Y} = d)$, $x = 0, 1$ and $d = 0, 1$.

We use beta priors with parameters (a_{xd}, b_{xd}) for the α's and (a_d^*, b_d^*) for the γ's, with the α's and γ's being mutually independent. If we impose the constraints, $\alpha_{x0} = \alpha_{x1}$ for $x = 0, 1$, then we have a four-parameter, nondifferential measurement error model. The log-odds ratio is related to the γ's by

$$\beta = \log\left[\{\gamma_1/(1-\gamma_1)\} / \{\gamma_0/(1-\gamma_0)\}\right].$$

Thus, the posterior distribution of β can be found via transformation from the posterior distribution of the γ's.

If we could observe all the $\mathbf{X}$'s, the joint density of the parameters and all the data would be proportional to

$$\prod_{x=0}^{1}\prod_{d=0}^{1}\Bigg[\alpha_{xd}^{a_{xd}-1}\left(1-\alpha_{xd}\right)^{b_{xd}-1} \qquad (9.32)$$
$$\times \prod_{i=1}^{n}\left\{\alpha_{xd}^{\mathbf{W}_i}\left(1-\alpha_{xd}\right)^{1-\mathbf{W}_i}\right\}^{I(\mathbf{X}_i=x,\mathbf{Y}_i=d)}\Bigg]$$
$$\times \prod_{d=0}^{1}\left[\gamma_d^{a_d^*-1}\left(1-\gamma_d\right)^{b_d^*-1}\prod_{i=1}^{n}\left\{\gamma_d^{\mathbf{X}_i}\left(1-\gamma_d\right)^{1-\mathbf{X}_i}\right\}^{I(\mathbf{Y}_i=d)}\right].$$

We can use (9.4) and (9.32) to note that the posterior distribution of γ_d is a beta distribution with parameters $\sum_{i=1}^n \mathbf{X}_i I(\mathbf{Y}_i = d) + a_d^*$ and $\sum_{i=1}^n (1-\mathbf{X}_i) I(\mathbf{Y}_i = d) + b_d^*$. The posterior distribution of α_{xd} is also a beta distribution but with parameters $\sum_{i=1}^n \mathbf{W}_i I(\mathbf{X}_i = x, \mathbf{Y}_i = d) + a_{xd}$ and $\sum_{i=1}^n (1-\mathbf{W}_i) I(\mathbf{X}_i = x, \mathbf{Y}_i = d) + b_{xd}$. The conditional distribution of a missing $\mathbf{X}_i$, given the $(\mathbf{W}_i, \mathbf{Y}_i)$ and the parameters, is Bernoulli with success probability $p_{1i}/(p_{0i} + p_{1i})$, where

$$p_{xi} = \gamma_{\mathbf{Y}_i}^{x}\left(1-\gamma_{\mathbf{Y}_i}\right)^{1-x}\alpha_{x\mathbf{Y}_i}^{\mathbf{W}_i}\left(1-\alpha_{x\mathbf{Y}_i}\right)^{1-\mathbf{W}_i}.$$

Thus, in order to implement the Gibbs sampler, we need to simulate observations from the Bernoulli and beta distributions, both of which are easy to do using standard programs, so the Metropolis–Hastings algorithm was not needed.

For nondifferential measurement error, the only difference in these calculations is that $\alpha_{x0} = \alpha_{x1} = \alpha_x$, which have a beta prior with parameters (a_x, b_x) and a beta posterior with parameters $\sum_{i=1}^n \mathbf{W}_i I(\mathbf{X}_i = x) + a_x$ and $\sum_{i=1}^n (1-\mathbf{W}_i) I(\mathbf{X}_i = x) + b_d$.

We used uniform priors throughout, so that $a_{xd} = b_{xd} = a_d^* = b_d^* = 1$. We ran the Gibbs sampling with an initial burn-in period of 2,000 simulations, and then recorded every 50th simulation thereafter. The posterior modes were 0.623 and 0.927, respectively, these being very close to the maximum likelihood estimates. Note the large difference between the estimates for $d = 1$ and for $d = 0$, indicating the critical nature of whether or not the error is assumed to be nondifferential.

This example shows the value of validation data—without it, one is forced to assume nondifferential error and may, unwittingly, reach erroneous conclusions because this assumption does not hold. If at all feasible, the collection of validation is worth the extra effort and expense.

9.10 Framingham Data

As an illustration, we consider only those males ages 45+ whose cholesterol values at Exam #3 ranged from 200 to 300, giving a data set of $n = 641$ observations. Recall that $\mathbf{Y}$ is the indicator of coronary heart disease. Initial frequentist analysis of this data set showed no evidence of age or cholesterol effects, so we work only with two covariates, smoking status, $\mathbf{Z}$, and $\mathbf{X} = \log(\text{SBP}-50)$, where SBP is long-term average systolic blood pressure. The main surrogate $\mathbf{W}$ is the measurement of $\log(\text{SBP}-50)$ at Exam #3, while the replicate $\mathbf{T}$ is $\log(\text{SBP}-50)$ measured at Exam #2. Given $(\mathbf{Z}, \mathbf{X})$, $\mathbf{W}$ and $\mathbf{T}$ are assumed independent and normally distributed with mean $\mathbf{X}$ and variance σ_u^2; $\sigma_u^2 = \widetilde{\alpha}_1$ in the general notation of Chapter 8. The distribution of $\mathbf{X}$ given $\mathbf{Z}$ is assumed to be normal with mean $\alpha_0 + \alpha_z\mathbf{Z}$ and variance $\sigma_{x|z}^2$ ($\widetilde{\alpha}_2$ in the general notation). We also assume that $\sigma_{x|z}^2$ is constant, that is, independent of $\mathbf{Z}$. Let $\Theta = (\sigma_u^2, \alpha_0, \alpha_z, \sigma_{x|z}^2)$.

Previous analysis suggested that the measurement error variance is less than 50% of the variance of the true long-term SBP given smoking status. We define $\Delta = \sigma_u^2/\sigma_{x|z}^2$ to be the ratio of these variances and assume $\Delta \in (0, 0.5)$. Restricting the range here makes sense, and we would not credit an analysis that suggested that the measurement error variance is larger than the variance of true long-term SBP given smoking status.

The Bayesian analysis will be based on the original model, so that $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z})$ is treated as being logistic with mean

$$H(\beta_0 + \beta_x\mathbf{X} + \beta_z\mathbf{Z}).$$

The unknown parameters are $(\beta_0, \beta_x, \beta_z, \alpha_0, \alpha_z, \sigma_{x|z}^2, \Delta)$. The first five of these are given diffuse (noninformative) locally uniform priors, the next-to-last has a diffuse inverse Gamma prior, the density functions being proportional to $1/\sigma_{x|z}^2$, and Δ has a uniform prior on the interval between zero and one half.

We use WinBUGS to implement the Bayesian logistic regression model. The WinBUGS model, together with an R file used for data and output manipulation, is provided as part of the software files for this book.

Mixing was very good for β_z, α_0, α_z, $\sigma_{x|z}^2$, σ_u^2, and λ. For these pa-

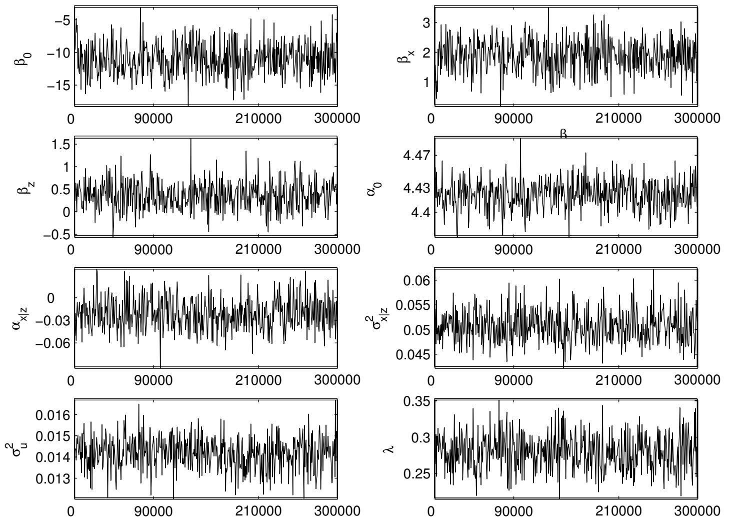


Figure 9.11 *Every 600th iteration of the Gibbs sampler for Framingham example.*

rameters $1,000$ burn-in and $10,000$ simulations were enough for accurate estimation. However, the chains corresponding to β_0 and β_x were mixing very slowly, and we ran $310,000$ iterations of the Gibbs algorithm and discarded the first $10,000$ as burn-in. Figure 9.11 displays every 600th iteration for the model parameters with similar, but less clear patterns, for the unthinned chains.

Table 9.1 compares the inference results for the maximum likelihood analysis, based on the regression calibration approximation with the Bayesian inference based on Gibbs sampling. Clearly, the two types of inferences agree reasonably closely on most parameters. The Bayesian analysis estimates an 8.5% higher effect of SBP $\beta_x = 1.91$ for Gibbs sampling, compared to $\beta_x = 1.76$ for maximum likelihood, but the difference is small relative to the standard errors. Results in Table 9.1 are similar to the likelihood and regression calibration results given in Section 8.5, and the differences are easily due to our use here of only 641 of the 1,615 subjects analyzed in Section 8.5.

9.11 OPEN Data: A Variance Components Model

The OPEN Study was introduced in Section 1.2 and Section 1.5, see Subar, Kipnis, Troiano, et al. (2003) and Kipnis, Midthune, Freedman, et al. (2003)indexsLongitudinal data. Briefly, each participant completed

Parameter	ML. est.	Boot. se	Bayes p. mean	Bayes p. std.
β_0	-10.10	2.400	-10.78	2.542
β_x	1.76	0.540	1.91	0.562
β_z	0.38	0.310	0.40	0.302
α_0	4.42	0.019	4.42	0.019
$10 \times \alpha_z$	-0.19	0.210	-0.20	0.217
$10 \times \sigma^2_{x\|z}$	0.47	0.033	0.51	0.032
$10 \times \sigma^2_u$	0.14	0.011	0.16	0.008
λ	0.30	0.031	0.28	0.025

Table 9.1 *Framingham data. The effects of SBP and smoking are given by β_x and β_z, respectively. The measurement error variance is σ^2_u. The mean of long-term SBP given smoking status is linear with intercept α_0, slope α_z and variance $\sigma^2_{x|z}$. Also, $\lambda = \sigma^2_u/\sigma^2_{x|z}$. "ML" = maximum likelihood, "se" = standard error, "Boot." = bootstrap, "Bayes" =Bayesian inference based on Gibbs sampling implemented in WinBUGS, "p. mean" = posterior mean, and "p. std" = posterior standard deviation.*

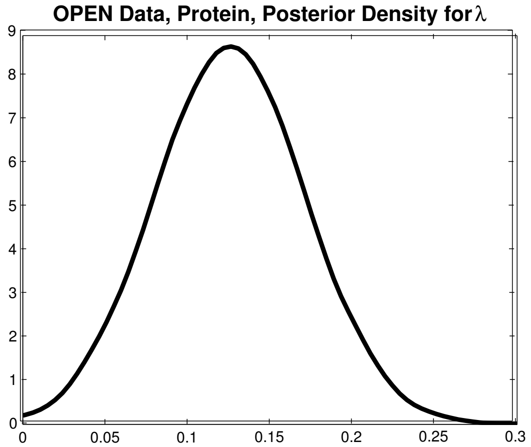


Figure 9.12 *Results of the OPEN Study for Protein intake for females. Plotted is the posterior density of the attenuation λ, defined in this case as the slope of the regression of true intake on a single food frequency questionnaire. The posterior mean is 0.13, with 95% credible interval $[0.04, 0.21]$, roughly in line with results reported previously.*

up to two food frequency questionnaires (FFQ) which measured reported Protein intake, and also up to two biomarkers for Protein intake (urinary nitrogen). Letting $\mathbf{Y}$ denote the logarithm of the FFQ, $\mathbf{W}$ the logarithm of the biomarker and $\mathbf{X}$ the logarithm of usual intake, the variance components model used is

$$\begin{aligned} \mathbf{Y}_{ij} &= \beta_0 + \beta_x \mathbf{X}_i + r_i + \epsilon_{ij}, \qquad (9.33) \\ \mathbf{W}_{ij} &= X_{ij} + U_{ij}, \end{aligned}$$

where $\epsilon_{ij} = \text{Normal}(0, \sigma_\epsilon^2)$, $U_{ij} = \text{Normal}(0, \sigma_u^2)$ and $r_i = \text{Normal}(0, \sigma_r^2)$: the terms r_i is a person-specific bias or equation error, see Section 1.5. In Chapter 11, we note that (9.33) is a linear mixed model with repeated measures. We used a subset of the women in the OPEN study for this analysisindexsLongitudinal data.

The purpose of the OPEN study was to investigate the properties of the FFQ for use in large cohort studies. In regression calibration, Chapter 4, in a cohort study we use the regression of usual intake on the FFQ as the predictor of disease outcome. The slope of this regression is simply

$$\lambda_{\text{regcal}} = \text{cov}(Q, X)/\text{var}(Q).$$

Kipnis, Subar, Midthune, et al. (2003) describe λ_{regcal} as the attenuation factor and note that the regression calibration approximation says that if the true relative risk is R, then the observed relative risk from the use of the FFQ will be $R^{\lambda_{\text{regcal}}}$. For example, a true relative risk of 2 would appear as $2^{.4} = 1.32$ if the attenuation factor were 0.4 and as $2^{.2} = 1.15$ if the attenuation factor were 0.2. It is thus of considerable interest to estimate λ_{regcal}. The WinBUGS code along with the prior distributions used is given in Appendix B.8.2.

We plot the posterior density of λ_{regcal} in Figure 9.12. The posterior mean is 0.13, with 95% credible interval $[0.04, 0.21]$, roughly in line with results reported by Kipnis, Subar, Midthune, et al. (2003). This means that a true relative risk of 2 for Protein intake will be attenuated to a relative risk of $2^{0.13} = 1.09$ when using the FFQ. As Kipnis, et al. state: *"Our data clearly document the failure of the FFQ to provide a sufficiently accurate report of absolute protein … intake to allow detection of their moderate associations with disease.*

Bibliographic Notes

Since the first edition of this book, the literature on Bayesian computation has exploded. The reader is referred to Gelman, Carlin, Stern, & Rubin, Gelman, (2004), Carlin & Louis (2000), and Gilks, Richardson, & Spiegelhalter, (1996) for a thorough introduction. Other important references include two classics, Box & Tiao (1973) and Berger (1985). The latter has an extensive and excellent theoretical treatment. There is also now a statistical package for Bayesian computation, called WinBUGS: we will illustrate the use of WinBUGS in this chapter. The literature now even includes an excellent book devoted exclusively to the Bayesian approach to measurement error modeling, especially for categorical data, see Gustafson (2004).

Good introductions to MCMC are given by Gelman, Carlin, Stern, & Rubin (2004), Carlin & Louis (2003), and Gilks, Richardson, & Spiegelhalter (1996).

The mechanics of stopping the Gibbs sampler and whether one should use one long sequence or a number of shorter sequences are matters of some controversy and not discussed here; however, we note that Gelman & Rubin (1992) and Geyer (1992) give exactly opposite recommendations. There is a large literature on diagnostics for convergence; see Cowles & Carlin (1996), Polson (1996), Brooks & Gelman (1998), Kass, Carlin, Gelman, & Neal (1998), and Mengersen, Robert, & Guihenneuc-Jouyaux (1999). Kass et al. (1998) is an interesting panel discussion of what is actually done in practice by three Bayesian experts, Carlin, Gelman, and Neal: Kass, though also an expert, is the moderator so we do not learn about his views or experiences. This discussion is quite interesting and well worth reading, unless you are already a Bayesian expert yourself, and probably even in that case. It seems that the experts do not use sophisticated convergence diagnostics, because they feel that these can be misleading. However, they all look at trace plots of various parameters, such as Figure 9.2. Carlin and Gelman monitor $\widehat{R}$ (Gelman & Rubin, 1992), which compares the estimated posterior variance from several chains combined to the average posterior variance from the individual chains. $\widehat{R}$ close to 1 means that the chains have mixed. Carlin and Neal also compute autocorrelations of various parameters; high autocorrelations are a sign of slow mixing. Neal also suggests looking at the log posterior density, which will be neither steadily increasing nor steadily decreasing if the chain has converged.

Alternatives to the Metropolis–Hastings algorithm have been proposed, though they seem less used in practice. For example, Smith & Gelfand (1992) discuss the rejection method and the weighted bootstrap method. Ritter & Tanner (1992) and references therein discuss ways of drawing samples from (9.4), including the griddy Gibbs sampler, which effectively discretizes the components of Ω in a clever way; this can be useful since sampling from a multinomial distribution is trivial.

CHAPTER 10

HYPOTHESIS TESTING

10.1 Overview

In this chapter, we discuss hypothesis tests concerning regression parameters when **X** is measured with error. In Section 3.3.1 we argued, in the context of some special cases, that naive tests for the regression coefficients of **Z** are not valid in general when **X** is measured with error and **X** is correlated with **Z**. In particular, we illustrated this in Figure 3.5, where we graphically illustrated a two-group, unbalanced analysis of covariance, showing that if **X** has a different distribution in the two groups, then the treatment effect test is invalid in the naive test, where **W** is simply substituted for **X**. In this chapter, we give a more detailed and thorough account of testing when **X** is measured with error.

To keep the exposition simple, we focus on linear regression. However, the results of Sections 10.2.1, 10.2.3, and 10.5 hold in general, and the results of Sections 10.2.2 and 10.4 hold to a good approximation for all generalized linear models, including logistic regression, whenever the regression calibration approximation is reasonable. More generally, the same can be said of any problem for which the mean and variance of the response depends only upon a linear combination of the predictors, which we assume throughout this chapter. We also assume nondifferential, additive measurement error, $\mathbf{W} = \mathbf{X} + \mathbf{U}$.

10.1.1 Simple Linear Regression, Normally Distributed **X**

In Section 3.2.1 we discussed the effects of measurement error on estimation in the simple linear regression model; see especially equation (3.4). Recall that the model is $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, where **X** has mean μ_x and variance $\sigma_x^2 = 1$, and the error about the regression line ϵ is independent of **X**, has mean zero and variance σ_ϵ^2. Suppose that instead of **X** we observe $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where **U** is independent of **X**, has mean zero, and variance $\sigma_u^2 = 1$. The attenuation is $\sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. As we described, if **X** is normally distributed, the observed regression of **Y** on **W** is the linear model with intercept $\beta_0 + \beta_x \mu_x (1 - \lambda)$, slope $\lambda \beta_x$, and residual variance $\sigma_\epsilon^2 + \lambda \beta_x^2 \sigma_u^2$.

Now consider testing the null hypothesis of no effect due to **X**: H_0 :

$\beta_x = 0$. Since the observed data have slope $\lambda\beta_x$, if the null hypothesis is true, then in the observed data the slope is also zero. In other words, with nondifferential error in this simple setup, no relationship between $\mathbf{Y}$ and $\mathbf{X}$ means no relationship between $\mathbf{Y}$ and the observed $\mathbf{W}$. This has two consequences for the naive test that ignores the measurement error:

- Since the observed data have zero slope under the null hypothesis, the naive test is valid, in the sense that its level (Type I error) is correct.
- Because $\mathbf{X}$ is normally distributed, the observed data actually follow a linear model. Hence, the naive test is efficient in this special case. Of course, this efficiency only holds for normally distributed $\mathbf{X}$.

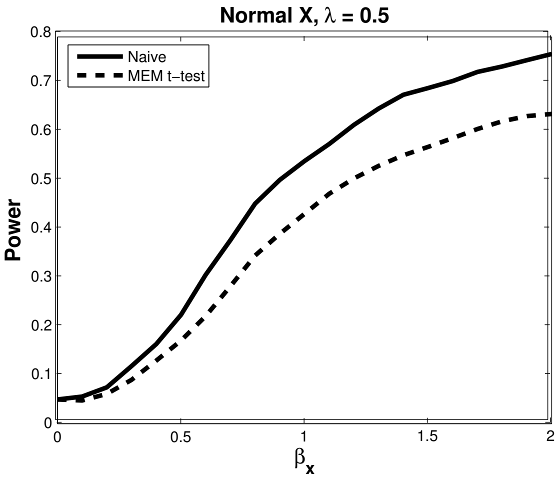


Figure 10.1 *This illustrates the power of a 5%-level test for the null hypothesis of zero slope in a simple linear regression when* $\mathbf{X}$ *is normally distributed,* $n = 20$*, and* $\sigma_x^2 = \sigma_\epsilon^2 = \sigma_u^2 = 1$*. Compared are the naive test that ignores measurement error (solid line) and the test that accounts for measurement error by estimating the standard deviation of the method of moments estimator via the bootstrap (dashed line).*

To illustrate these points, we did a small simulation study similar to that in Section 1.8.1, with $n = 20$ observations and $\mu_x = 0$, $\sigma_x^2 = \sigma_\epsilon^2 = \sigma_u^2 = 1$. In this simulation, σ_u^2 was assumed known. We varied β_x and investigated the power of two tests. The first is the naive test, which is the efficient test because $\mathbf{X}$, $\mathbf{U}$, and ϵ are all normally distributed. The other test is one based upon accounting for measurement error. Specifically, as in Section 3.4.1, we computed the Fuller's corrected estimator (Fuller, 1987, Section 2.5.1), and computed its standard error using 3,000 bootstrap simulations. We then modified the simulations so that the level of each test was exactly 0.05. The comparison of the two methods is displayed in Figure 10.1, where we see that the naive test has the greater power, as predicted by the theory.

The loss of power by the test that corrects for measurement error is due to Fuller's correction and its bootstrap standard error. Using the Fuller correction seems reasonable since it provides an estimator with good finite-sample properties, in particular, less finite-sample bias than simply dividing the naive estimator by $\widehat{\lambda}$. However, if one divides the naive estimator *and* its standard error by $\widehat{\lambda}$, then, though the estimate of β_x may be unstable in small samples, the t-statistic corrected in this way would be the same as the naive t-statistic because the $\widehat{\lambda}$'s would cancel. Thus, this less sophisticated correction would, ironically, result in the naive test and hence be efficient.

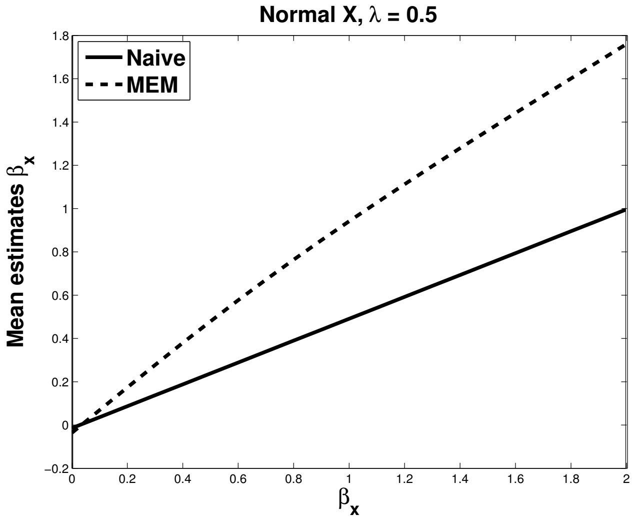


Figure 10.2 *This illustrates the effects of measurement error in a simple linear regression when* $\mathbf{X}$ *is normally distributed,* $n = 20$*, and* $\sigma_x^2 = \sigma_\epsilon^2 = \sigma_u^2 = 1$*, and for different values of* β_x*. The dotted line reflects the true value of* β_x*. Compared are the naive estimate of the slope ignoring measurement error (solid line) and the estimate that accounts for measurement error (dashed line). Note the severe small-sample bias in the naive estimate, as well as the near lack of small-sample bias for the measurement error estimate. The point here is that if estimation and inference about* β_x *are of interest, then measurement error needs to be accounted for.*

Figure 10.1 might cause one to think that measurement error can be safely ignored. This is certainly true *provided* the model is simple linear regression, all random variables are normally distributed, and the only interest is in testing the null hypothesis of zero slope. As a cautionary note, in Figure 10.2 we illustrate once again the effects of measurement error on estimation. In this figure, we show that the naive estimate is very severely biased, when the correction for attenuation with Fuller's modification has almost no small-sample bias.

10.1.2 Analysis of Covariance

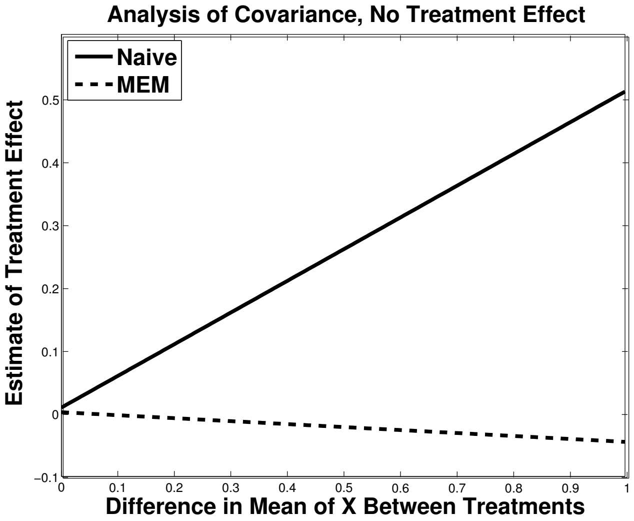


Figure 10.3 *This illustrates the effects of measurement error in unbalanced analysis of covariance, when* $\mathbf{X}$ *is normally distributed,* $n = 20$, *and* $\sigma_x^2 = \sigma_\epsilon^2 = \sigma_u^2 = 1$, *and for different values of* Δ, *the difference in the mean of* $\mathbf{X}$ *between the true groups. Compared are the mean estimates of the treatment effect ignoring measurement error (solid line) and accounting for measurement error (dashed line). Note the increased bias of the naive estimate of treatment effect as the imbalance between the two groups increases. The dashed line shows that the correction results in only a slight negative bias, which is small-sample effect.*

It is, of course, not always true that the Type-I error of the naive test is the nominal 5%. Consider, for example, the analysis of covariance model described in Section 3.3.1 and Figure 3.5. Consider a situation of two-group analysis of covariance where in the first group, $\mathbf{X}$ is normally distributed with mean μ_1 and variance $\sigma_x^2 = 1$, and in the second group, $\mathbf{X}$ is normally distributed with mean μ_2 and variance $\sigma_x^2 = 1$. The measurement error variance is $\sigma_u^2 = 1$, and the residual mean square is $\sigma_\epsilon^2 = 1$. The difference in the mean of $\mathbf{X}$ in the two groups is $\Delta = \mu_2 - \mu_1$, with larger values of Δ reflecting increased imbalance between the two groups. In symbols,

$$\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z} + \epsilon,$$

where $\mathbf{Z}$ is the dummy variable indicating group assignment and the mean of $\mathbf{X}$ given $\mathbf{Z} = z$ is μ_z.

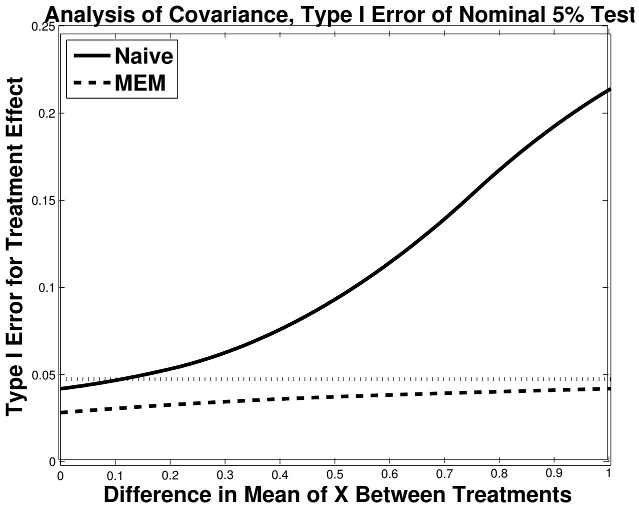


Figure 10.4 *This illustrates the effects of measurement error in unbalanced analysis of covariance, when* $\mathbf{X}$ *is normally distributed,* $n = 20$, *and* $\sigma_x^2 = \sigma_\epsilon^2 = \sigma_u^2 = 1$, *and for different values of* Δ, *the difference in the mean of* $\mathbf{X}$ *between the true groups. Compared are the Type I errors of the naive test for treatment effect ignoring measurement error (solid line) and the Type I error of the test that accounts for measurement error (dashed line). Note the increased level of the naive test as the imbalance between the two groups increases.*

As described in Section 3.3.1, when ignoring measurement error, the effect of measurement error in $\mathbf{X}$ is to bias the estimate of the treatment effect β_z. This is illustrated in Figure 10.3, which is the result of a simulation study with 1,000 replications, where we display the mean estimate of treatment effect β_z as a function of the difference in the mean of $\mathbf{X}$ in the two groups, $\Delta = \mu_2 - \mu_1$, both ignoring and accounting for measurement error. The latter method uses Fuller's modification of the correction for attenuation. Note the severe bias of the naive estimate for larger values of Δ and the corresponding near lack of bias in the correction for attenuation.

The bias in treatment effect when ignoring measurement error also leads to invalid tests, that is, the usual test that ignores measurement error has Type I error greater than the nominal 5%. In Figure 10.4, we plot the Type I error as a function of Δ ignoring and accounting for measurement error: The latter uses a t-test with standard error estimated by $1,000$ bootstrap simulations.

The analysis of covariance illustrates that hypothesis testing in the measurement error context is not fully straightforward. Understanding when the naive test that ignores measurement error is valid and attains its nominal Type I error is thus of considerable importance.

10.1.3 General Considerations: What Is a Valid Test?

Assuming that one or more of the estimation methods described in the previous chapters is applicable, the simplest approach to hypothesis testing forms the required test statistic from the parameter estimates and their estimated standard errors. Such tests are justified whenever the estimators themselves are justified. However, this approach to testing is only possible when the indicated methods of estimation are possible, and thus require either knowledge of the measurement error variance or the presence of validation data or replicate measurements or instrumental variables, etc.

There are certain situations in which naive hypothesis tests are justified and thus can be performed without additional data or information of any kind. Here *naive* means that we ignore measurement error and substitute $\mathbf{W}$ for $\mathbf{X}$ in a test that is valid when $\mathbf{X}$ is observed. This chapter studies naive tests, describing when they are and are not acceptable, and indicates how supplementary data, when available, can be used to improve the efficiency of naive tests.

We use the criterion of asymptotic validity to distinguish between acceptable and nonacceptable tests. We say a test is asymptotically valid if its Type I error rate approaches its nominal level as the sample size increases. Asymptotic validity, which we shorten to *validity*, of a test is a minimal requirement for acceptability.

10.1.4 Summary of Major Results

The main results on the validity of naive tests under nondifferential measurement error are as follows:

- The naive test of no effects due to $\mathbf{X}$ is valid.

- The naive test of no effects due to $(\mathbf{Z}^t, \mathbf{X}^t)^t$ is valid, that is, that none of the covariates affects $\mathbf{Y}$.
- The naive test of no effects due to $\mathbf{Z}$ is not valid in general but is valid under some restrictive assumptions.
- The naive test of no effects due to a specified subvector of $\mathbf{X}$, for example, the first component of $\mathbf{X}$, is not valid in general.
- When $\mathbf{Y}$ follows a generalized linear model (Section A.8) in $\mathbf{Z}$ and $\mathbf{X}$, then we show that the efficient score test of no effects due to $\mathbf{X}$ is easily obtained: One takes the efficient score test when $\mathbf{X}$ is observed and replaces $\mathbf{X}$ by a parametric estimate of $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$. Put another way, a null hypothesis test based on regression calibration is (asymptotically) efficient.

These results are obtained using the regression calibration approximation, which takes the regression model for $\mathbf{Y}$ given $\mathbf{Z}$ and $\mathbf{X}$ and replaces $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$. Recall that throughout this chapter we assume that response depends only upon a linear combination of the predictors, for example, as in a generalized linear model.

10.2 The Regression Calibration Approximation

In linear regression, the mean of the response given the true covariates is $\beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t\mathbf{X}$. Under the additional assumption that the possibly multivariate regression of $\mathbf{X}$ on $\mathbf{Z}$ and $\mathbf{W}$ is linear, that is,

$$E(\mathbf{X} \mid \mathbf{Z},\mathbf{W}) = \alpha_0 + \alpha_z^t\mathbf{Z} + \alpha_w^t\mathbf{W},$$

we have that the observed data also have a linear mean, namely

$$E(\mathbf{Y} \mid \mathbf{Z},\mathbf{W}) = \beta_0 + \beta_x^t\alpha_0 + (\beta_z^t + \beta_x^t\alpha_z^t)\mathbf{Z} + \beta_x^t\alpha_w^t\mathbf{W}. \qquad (10.1)$$

Equation (10.1) is the starting point for our discussion of testing. One of the assumptions of our measurement error model is that α_w^t is an invertible matrix.

A naive analysis of the data fits a linear model as well. We write this model as

$$E(\mathbf{Y} \mid \mathbf{Z},\mathbf{W}) = \gamma_0 + \gamma_z^t\mathbf{Z} + \gamma_w^t\mathbf{W}. \qquad (10.2)$$

It is the correspondence between the naive model (10.2) and the actual model (10.1) that is of interest here.

The assumption made above that α_w^t is invertible is not onerous. In the case of classical multivariate measurement error where $\mathbf{W} = \mathbf{X} + \mathbf{U}$, if the conditional covariance matrix of $\mathbf{X}$ given $\mathbf{Z}$, $\Sigma_{x|z}$, is invertible then

$$\alpha_w^t = \Lambda_z = \Sigma_{x|z}(\Sigma_{x|z} + \Sigma_u)^{-1},$$

so α_w^t is invertible whenever $\Sigma_{x|z}$ is invertible, that is, whenever we do not have complete collinearity of the components of $(\mathbf{X}\,\mathbf{Z})$. This is a minimal assumption for β_x to be estimable even when there is no measurement error.

10.2.1 Testing $H_0 : \beta_x = 0$

Here we show that the naive test of no effect due to any of the predictors measured with error is asymptotically valid, a point illustrated for simple linear regression in Section 10.1.1. The result though holds in general, and not just for linear regression.

A comparison of (10.1) and (10.2) shows that $\beta_x = 0$ implies that $\alpha_w\beta_x = 0$, which in turn implies that $\gamma_w = 0$. The converse is also true, namely that $\gamma_w = 0$ implies that $\beta_x = 0$ because α_w is invertible.

Because $\gamma_w = 0$ if $\beta_x = 0$, it follows that the naive test, that is, the test of $H_0 : \gamma_w = 0$, is a valid test of $H_0 : \beta_x = 0$.

Although $\gamma_w = 0$ only if $\beta_x = 0$, this reverse implication, though perhaps interesting, is not necessary for the validity of the naive test.

10.2.2 Testing $H_0 : \beta_z = 0$

Here we show that in linear regression, the naive tests for effects due to $\mathbf{Z}$ is typically invalid, except under special circumstances, a point illustrated for the analysis of covariance in Section 10.1.2.

Further comparison of (10.1) and (10.2) shows that $\beta_z = 0$ implies that $\gamma_z = 0$, only if $\alpha_z\beta_x = 0$. It follows that the naive test of $H_0 : \beta_z = 0$ is valid if $\mathbf{X}$ is unrelated to $\mathbf{Y}$ in the model (10.7), that is, $\beta_x = 0$, or if $\mathbf{Z}$ is unrelated to $\mathbf{X}$, that is, $\alpha_z = 0$.

In generalized linear models, the naive test is valid when $\mathbf{Z}$ and $\mathbf{X}$ are independent, at least approximately, at the level of the regression calibration approximation. Gail, Wieand, and Piantadosi (1984) and Gail, Tan, and Piantadosi (1988) showed that when the regression calibration approximation fails for logistic regression, then the naive test is no longer even approximately valid.

The general conclusion is that the test of $H_0 : \beta_z = 0$ is invalid, although there are certain situations in which it is valid.

10.2.3 Testing $H_0 : (\beta_x^t, \beta_z^t)^t = 0$

A final comparison of (10.1) and (10.2) shows that $(\beta_x^t, \beta_z^t)^t = 0$ if and only if $(\gamma_w^t, \gamma_z^t)^t = 0$, so the naive test that none of the covariates affects $\mathbf{Y}$ is valid in general.

10.3 Illustration: OPEN Data

In many nutrition studies, the response $\mathbf{Y}$ is binary (disease or not), in which case logistic regression is the likely model choice. If $\mathrm{pr}(\mathbf{Y} = 1|\mathbf{Z},\mathbf{W}) = H(\beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t\mathbf{X})$, then following (10.1) the regression calibration approximation is that

$$\mathrm{pr}(\mathbf{Y} = 1 \mid \mathbf{Z},\mathbf{W}) = H\left\{\beta_0 + \beta_x^t\alpha_0 + (\beta_z^t + \beta_x^t\alpha_z^t)\mathbf{Z} + \beta_x^t\alpha_w^t\mathbf{W}\right\}. \quad (10.3)$$

Some of these concepts can be illustrated numerically in the OPEN data; see Section 1.2. Recall here that $\mathbf{Z}$ is the logarithm of energy (caloric) intake as measured by the doubly labeled biomarker, which we are taking as measured without error.

A standard practice is to take $\mathbf{X}$ to be the logarithm of protein density, which is the percentage of calories coming from protein. Effectively, this is simply the logarithm of the ratio of protein intake to energy intake. The surrogate for $\mathbf{X}$ is $\mathbf{W}$, the logarithm of the ratio of the protein biomarker to energy intake. The interpretation is rather nice: If we change $\mathbf{X}$, then we are changing the relative composition of what we eat.

Using the methods for regression calibration in Section 4.4.2, that is, equation (4.4), we obtain the estimate that $E\widehat{(\mathbf{X}|\mathbf{W},\mathbf{Z})} = \widehat{\alpha}_0 + \widehat{\alpha}_w\mathbf{W} + \widehat{\alpha}_z\mathbf{Z} \approx -0.42 + 0.54\mathbf{W} + 0.06\mathbf{Z}$, and that the estimated correlation between $\mathbf{X}$ and $\mathbf{Z}$ is -0.15. If we inspect (10.1), we see that when we ignore measurement error, the slope in the regression of a response $\mathbf{Y}$ on $(\mathbf{W},\mathbf{Z})$ has an approximate coefficient $\beta_z + 0.06\beta_x$ for $\mathbf{Z}$. This suggests that if there is no real energy effect ($\beta_z = 0$), then since the observed data should manifest a slope of only approximately $0.06\beta_x$, we are unlikely to conclude incorrectly that there is an energy effect when we ignore measurement error, unless β_x is large and hence $\mathbf{X}$ is a very strong predictor of the response.

10.4 Hypotheses about Subvectors of β_x and β_z

There are situations in which interest focuses on testing for effects due to some subset of the predictors measured with error, or due to some subset of the error-free covariates. That is, if $\mathbf{X} = (\mathbf{X}_1^t, \mathbf{X}_2^t)^t$, $\beta_x = (\beta_{x,1}^t, \beta_{x,2}^t)^t$, and $\mathbf{Z} = (\mathbf{Z}_1^t, \mathbf{Z}_2^t)^t$, $\beta_z = (\beta_{z,1}^t, \beta_{z,2}^t)^t$, then we may be interested in testing $H_0 : \beta_{x,1} = 0$ or $H_0 : \beta_{z,1} = 0$.

We have already seen that for testing $H_0 : \beta_z = 0$, the naive test is not valid in general, and it follows from similar reasoning that the same is true of naive tests of $H_0 : \beta_{z,1} = 0$. Therefore, we restrict attention to naive tests of $H_0 : \beta_{x,1} = 0$.

Suppose now that $\beta_x^t\mathbf{X} = \beta_{x,1}^t\mathbf{X}_1 + \beta_{x,2}^t\mathbf{X}_2$ and that

$$E(\mathbf{X}_1 \mid \mathbf{Z},\mathbf{W}_1,\mathbf{W}_2) = \alpha_{1,0} + \alpha_{1,z}^t\mathbf{Z} + \alpha_{1,w_1}^t\mathbf{W}_1 + \alpha_{1,w_2}^t\mathbf{W}_2;$$

$$E(\mathbf{X}_2 \mid \mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2) \quad = \quad \alpha_{2,0} + \alpha^t_{2,z}\mathbf{Z} + \alpha^t_{2,w_1}\mathbf{W}_1 + \alpha^t_{2,w_2}\mathbf{W}_2, \tag{10.4}$$

where $\mathbf{W} = (\mathbf{W}_1^t, \mathbf{W}_2^t)^t$ is partitioned as is $\mathbf{X}$.

With these changes (10.1) becomes

$$\begin{aligned} E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) &= \beta_0 + \beta^t_{x,1}\alpha_{1,0} + \beta^t_{x,2}\alpha_{2,0} \\ &+ (\beta^t_z + \beta^t_{x,1}\alpha^t_{1,z} + \beta^t_{x,2}\alpha^t_{2,z})\mathbf{Z} + (\beta^t_{x,1}\alpha^t_{1,w_1} + \beta^t_{x,2}\alpha^t_{2,w_1})\mathbf{W}_1 \\ &\quad + (\beta^t_{x,1}\alpha^t_{1,w_2} + \beta^t_{x,2}\alpha^t_{2,w_2})\mathbf{W}_2, \end{aligned} \tag{10.5}$$

and in a naive analysis of the data the mean model

$$E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) = \gamma_0 + \gamma^t_z\mathbf{Z} + \gamma^t_{w_1}\mathbf{W}_1 + \gamma^t_{w_1}\mathbf{W}_2 \tag{10.6}$$

is fit to the observed data.

Comparing (10.5) and (10.6) shows that $\beta_{x,1} = 0$ implies that $\gamma_{w_1} = 0$ only if $\alpha_{2,w_1}\beta_{x,2} = 0$. It follows that the naive test of $H_0 : \beta_{x,1} = 0$ is valid only if $\alpha_{2,w_1}\beta_{x,2} = 0$. If $\mathbf{X}_2$ is related to $\mathbf{Y}$, then $\beta_{x,2}$ is nonzero. If $\mathbf{X}_2$ is related to $\mathbf{W}_1$ in (10.4), then α_{2,w_1} is nonzero. This is the case whenever some components of $\mathbf{X}_1$ are correlated with some components of $\mathbf{X}_2$.

For example, consider the NHANES study introduced in Chapter 1 and discussed in more detail in Chapter 4. Let $\mathbf{X}$ be the vector of true total caloric intake (TC $= \mathbf{X}_1$) and saturated fat (SF $= \mathbf{X}_2$), and let $\mathbf{Z}$ denote nondietary variables. The naive test for a SF effect simply substitutes observed TC and SF intake for true TC and SF intake, and it is a valid test provided there is no risk of breast cancer due to TC ($\beta_{x,1} = 0$) or when the regression of true SF intake on observed SF, observed TC and non-dietary variables has no component due to TC ($\alpha_{2,w_1} = 0$).

In general, the conclusion is that the test of $H_0 : \beta_{x,1} = 0$ is invalid, although there are certain situations in which it is valid.

10.4.1 Illustration: Framingham Data

The Framingham Heart Study was introduced in Section 1.6.6 and described in more detail in Sections 5.4.1, 6.5, 7.2.3, 8.5, and 9.10. Here we consider two variables measured with error, namely transformed systolic blood pressure ($\mathbf{X}_1$) and the logarithm of cholesterol ($\mathbf{X}_2$). The variables $\mathbf{Z}$ measured without error in this example are age and smoking status: Age was normalized to have mean zero and variance one.

Using the replicates of blood pressure and cholesterol, we find an estimate of the measurement error covariance matrix using equation (4.3) as follows: The measurement error variance for transformed systolic blood pressure was 0.0126, that for transformed cholesterol was 0.0085, and the correlation of the measurement errors was estimated as 0.0652, that is,

essentially zero. The variances of transformed blood pressure and cholesterol were 0.0525 and 0.0316, respectively, with a correlation of 0.0966. In other words, both transformed blood pressure and transformed cholesterol and their measurement errors are essentially independent. The correlations of observed blood pressure and cholesterol with age and smoking status were also modest. When we used the regression calibration formula (4.5), we found the following regressions:

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2) &\approx 1.0686 + (0.0131, -0.0041)\mathbf{Z} \\ &\qquad + 0.7459\mathbf{W}_1 + 0.0076\mathbf{W}_2; \\ E(\mathbf{X}_2 \mid \mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2) &\approx 1.4298 + (0.0022, 0.0013)\mathbf{Z} \\ &\qquad + 0.0059\mathbf{W}_1 + 0.7310\mathbf{W}_2. \end{aligned}$$

As is seen here, effectively regression calibration of transformed systolic blood pressure on all the variables is essentially the same as regression calibration using the blood pressure measurements alone, and similarly for cholesterol. In particular, we see that effectively, $\alpha_{1,z} \approx 0$, $\alpha_{1,w_2} \approx 0$, $\alpha_{2,z} \approx 0$ and $\alpha_{2,w_1} \approx 0$, so that in practice the naive test for systolic blood pressure is very nearly valid.

10.5 Efficient Score Tests of $H_0 : \beta_x = 0$

In this section, we assume that $\mathbf{Y}$ given $\mathbf{Z}$ and $\mathbf{X}$ follows a generalized linear model (Section A.8). In particular, the mean and variance functions for these models are in the form

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) &= m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = m_{\mathbf{Y}}(\beta_0 + \beta^t_z\mathbf{Z} + \beta^t_x\mathbf{X}); \quad (10.7) \\ \mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) &= \sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta) \\ &= \sigma^2 g^2(\beta_0 + \beta^t_z\mathbf{Z} + \beta^t_x, \theta). \quad (10.8) \end{aligned}$$

We show that the naive score test of $H_0 : \beta_x = 0$, while asymptotically valid in general, is not generally an efficient score test. However, we do find a test that is asymptotically equivalent to the efficient score test and show that under certain conditions this test is equal to the naive score test.

Recall that the naive test simply substitutes $\mathbf{W}$ for $\mathbf{X}$. We show that if a parametric model for $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ is appropriate, say $E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) = m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \alpha)$, and if $\widehat{\alpha}$ is a $n^{1/2}$-consistent estimator of α, then the test that substitutes $m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \widehat{\alpha})$ for $\mathbf{X}$ is asymptotically an efficient score test. It must be emphasized that this result about substituting $m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \widehat{\alpha})$ for $\mathbf{X}$ requires the assumption of a generalized linear model.

The validity of naive null tests for predictors measured with error, and the efficiency for generalized linear models of tests which replace $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$, was shown by Tosteson and Tsiatis (1988). For the special case of models with canonical link functions, the efficiency of tests that replace $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$ follows from the form of the efficient score for generalized linear measurement error models given in Stefanski and Carroll (1987).

It follows from these results that the only time the naive test of $H_0 : \beta_x = 0$ in generalized linear models is equivalent to the efficient score test occurs when $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$ is independent of $\mathbf{Z}$ and linear in $\mathbf{W}$. Moreover, Tosteson and Tsiatis (1988) showed that the asymptotic relative efficiency (ARE) of the naive test to the efficient score test is always less than 1, unless the two tests are equivalent. They also showed that, for the special case where $\mathbf{X}$ is univariate and $\mathbf{Z}$ is not present, this ARE is $\{\text{corr}\,(E(\mathbf{X}|\mathbf{W}),\mathbf{W})\}^2$. Thus, the naive test can be arbitrarily inefficient if $E(\mathbf{X}|\mathbf{W})$ is sufficiently nonlinear in $\mathbf{W}$.

The mathematical arguments supporting these statements are given in the following subsection. This subsection is fairly technical and can be omitted on first reading.

10.5.1 Generalized Score Tests

To define a generalized score test of $H_0 : \beta_x = 0$, let $H_i(\alpha)$ be any random vector depending on $(\mathbf{Z}_i, \mathbf{X}_i, \mathbf{W}_i)$ and the parameter α and having the same dimension as $\mathbf{X}_i$. Possible choices of $H_i(\alpha)$ are discussed later. Define

$$\mathcal{L}(\beta_0,\beta_z,\alpha,\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} H_i(\alpha)d_i(\beta_0,\beta_z,\theta)\left\{\mathbf{Y}_i - m_{\mathbf{Y}}(\beta_0+\beta_z^t\mathbf{Z}_i)\right\}, \tag{10.9}$$

where d_i used here and c_i used below are defined by

$$\begin{aligned} d_i(\beta_0,\beta_z,\theta) &= m'_{\mathbf{Y}}(\beta_0+\beta_z^t\mathbf{Z}_i)/g^2(\beta_0+\beta_z^t\mathbf{Z}_i,\theta) \\ c_i(\beta_0,\beta_z,\theta) &= d_i(\beta_0,\beta_z,\theta)m'_{\mathbf{Y}}(\beta_0+\beta_z^t\mathbf{Z}_i). \end{aligned}$$

Our test statistic is based on $\mathcal{L}$ with the parameters β_0, β_z, α, and θ replaced by estimators. Also define

$$\begin{aligned} C_1(\beta_0,\beta_z,\alpha,\theta) &= n^{-1}\sum_{i=1}^{n} H_i(\alpha)H_i^t(\alpha)c_i(\beta_0,\beta_z,\theta); \\ C_2(\beta_0,\beta_z,\alpha,\theta) &= n^{-1}\sum_{i=1}^{n} H_i(\alpha)(1,\ \mathbf{Z}_i^t)^t c_i(\beta_0,\beta_z,\theta); \\ C_3(\beta_0,\beta_z,\alpha,\theta) &= n^{-1}\sum_{i=1}^{n} (1,\ \mathbf{Z}_i^t)^t(1,\ \mathbf{Z}_i^t)c_i(\beta_0,\beta_z,\theta); \\ D(\beta_0,\beta_z,\alpha,\theta) &= C_1 - C_2C_3^{-1}C_2^t, \end{aligned}$$

where in the last equation the dependence of C_1, C_2, and C_3 on $(\beta_0,\beta_z,\alpha,\theta)$ has been suppressed for brevity.

Let $\widehat{\theta}$ be any $n^{1/2}$-consistent estimate of the variance parameter θ; see Section A.7 or Carroll and Ruppert (1988, Chapter 3) for some methods of estimating θ. If α is unknown, for example, when

$$H_i(\alpha) = E(\mathbf{X}|\mathbf{Z},\mathbf{W}) = m(\mathbf{Z},\mathbf{W},\alpha),$$

then we assume a $n^{1/2}$-consistent estimator of α. Methods of estimating α are discussed in Chapter 4. The quasilikelihood and variance function (QVF) estimates of (β_0,β_z), $(\widehat{\beta}_0,\ \widehat{\beta}_z)$, satisfy

$$0 = \sum_{i=1}^{n}(1,\ \mathbf{Z}_i^t)^t d_i(\widehat{\beta}_0,\widehat{\beta}_z,\widehat{\theta})\left\{\mathbf{Y}_i - m_{\mathbf{Y}}(\widehat{\beta}_0+\widehat{\beta}_z^t\mathbf{Z}_i)\right\}.$$

With $\dim(\mathbf{Z})$ denoting the dimension of $\mathbf{Z}$, define

$$\widehat{\sigma}^2 = \{n-1-\dim(\mathbf{Z})\}^{-1}\sum_{i=1}^{n}\frac{\left\{\mathbf{Y}_i - m_{\mathbf{Y}}(\widehat{\beta}_0+\widehat{\beta}_z^t\mathbf{Z}_i)\right\}^2}{g^2(\widehat{\beta}_0+\widehat{\beta}_z^t\mathbf{Z}_i,\widehat{\theta})}.$$

We consider test statistics of the form

$$\widehat{\sigma}^{-2}\mathcal{L}^t(\widehat{\beta}_0,\widehat{\beta}_z,\widehat{\alpha},\widehat{\theta})D^{-1}(\widehat{\beta}_0,\widehat{\beta}_z,\widehat{\alpha},\widehat{\theta})\mathcal{L}^t(\widehat{\beta}_0,\widehat{\beta}_z,\widehat{\alpha},\widehat{\theta}). \tag{10.10}$$

When $\mathbf{X}$ is observable, then setting $H_i(\alpha) = \mathbf{X}_i$ in (10.9) results in (10.10) being the usual score test statistic of $H_0 : \beta_x = 0$. The naive score test statistic is obtained by setting $H_i(\alpha) = \mathbf{W}_i$ in (10.9). We show in this section that when $E(\mathbf{X}|\mathbf{Z},\mathbf{W}) = m(\mathbf{Z},\mathbf{W},\alpha)$, then setting $H_i(\alpha) = m(\mathbf{Z}_i,\mathbf{W}_i,\alpha)$ in (10.9) results in a test statistic that is asymptotically equivalent to the efficient score test statistic.

We now show that under the hypothesis $H_0 : \beta_x = 0$, the test statistic in (10.10) is asymptotically chi-square with degrees of freedom equal to the common dimension of $H_i(\alpha)$, $\mathbf{X}_i$ and β_x. It follows from Carroll and Ruppert (1988, Chapter 7) that to order $o_p(1)$, under H_0

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_0-\beta_0\\ \widehat{\beta}_z-\beta_z\end{pmatrix} \approx \frac{C_3^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}1\\ \mathbf{Z}_i\end{pmatrix}d_i\left\{\mathbf{Y}_i - m_{\mathbf{Y}}(\beta_0+\beta_z^t\mathbf{Z}_i)\right\},$$

where the dependence of C_3 and d_i on the parameters has been suppressed. Since $E(\mathbf{Y}_i|\mathbf{Z}_i,\mathbf{W}_i) = E(\mathbf{Y}_i|\mathbf{Z}_i,\mathbf{X}_i) = m_{\mathbf{Y}}(\beta_0+\beta_z^t\mathbf{Z}_i)$ under the null hypothesis, it is straightforward to show that to order $o_p(1)$,

$$\begin{aligned} \mathcal{L}^t(\widehat{\beta}_0,\widehat{\beta}_z,\widehat{\alpha},\widehat{\theta}) &\approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i \\ &\times \left\{\mathbf{Y}_i - m_{\mathbf{Y}}(\beta_0+\beta_z^t\mathbf{Z}_i)\right\}\left\{H_i(\alpha) - C_2C_3^{-1}\begin{pmatrix}1\\ \mathbf{Z}_i\end{pmatrix}\right\}, \end{aligned} \tag{10.11}$$

and $\mathcal{L}^t(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\alpha}, \widehat{\theta})$ is hence asymptotically multivariate normal with mean zero and covariance matrix $\sigma^2 D(\beta_0, \beta_z, \alpha, \theta)$. In (10.11) $d_i = d_i(\beta_0, \beta_z, \theta)$. It follows that (10.10) has the indicated chi-square distribution.

It remains to show that for generalized linear models, substituting $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ for $H_i(\alpha)$ in (10.10) results in a test that is asymptotically equivalent to the efficient score test. The argument is adapted from Tosteson and Tsiatis (1988).

The density or mass function of a generalized linear model uses the exponential family density given by (A.41). Write $\xi = g(\eta)$ with $\eta = \beta_0 + \beta_x^t x + \beta_z^t z$. Using the assumption of nondifferential measurement error (conditional independence so that $\mathbf{Y}$ and $\mathbf{W}$ are independent given $\mathbf{X}$ and $\mathbf{Z}$), the density or mass function of the observed data is

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{Z},\mathbf{W}}(y|z,w) &= \int f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w)d\mu(x) \\ &= \int \exp\left[\frac{yg(\eta) - \mathcal{C}\left\{g(\eta)\right\}}{\phi} + c(y,\phi)\right] f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w)d\mu(x). \end{aligned}$$

Write $h(y,z) = \exp\left(\left[yg(\beta_0 + \beta_z^t z) - \mathcal{C}\left\{g(\beta_0 + \beta_z^t z)\right\}\right]/\phi\right)$. Since $c(y,\phi)$ does not depend on β_x, the likelihood score used in construction of the efficient score statistic is

$$\begin{aligned} &\frac{\partial}{\partial \beta_x} \log\left\{f_{\mathbf{Y}|\mathbf{Z},\mathbf{W}}(y|z,w)\right\}\Big|_{\beta_x=0} \\ &= \frac{1}{h(y,z)} \frac{\partial}{\partial \beta_x} \int f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w)d\mu(x)\Big|_{\beta_x=0} \\ &= \frac{1}{h(y,z)} \Bigg[\int f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) \\ &\qquad \times \frac{\partial}{\partial \beta_x} \log\left\{f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x)\right\} d\mu(x)\Bigg]_{\beta_x=0} \\ &= \int f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) \frac{\partial}{\partial \beta_x} \log\left\{f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x)\right\}\Big|_{\beta_x=0} d\mu(x) \\ &= g'(\beta_0 + \beta_z^t z)\left[y - \mathcal{C}'\left\{g(\beta_0 + \beta_z^t z)\right\}\right] \\ &\qquad \times \int (x/\phi) f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w)d\mu(x) \\ &= \frac{1}{\phi}\left[y - \mathcal{C}'\left\{g^2(\beta_0 + \beta_z^t z)\right\}\right] \\ &\qquad g'(\beta_0 + \beta_z^t z)E(\mathbf{X}|\mathbf{Z} = z, \mathbf{W} = w). \end{aligned} \tag{10.12}$$

If $\mathbf{X}$ were observable, the only difference in these calculations would be that $\mathbf{X}$ would replace $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ in (10.12). Hence, the efficient score test for the observed data is obtained by substituting $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ for $\mathbf{X}$.

Bibliographic Notes

For the case studied above in Section 10.5.1 there is a parametric model, $E(\mathbf{X}|\mathbf{Z},\ \mathbf{W}) = m(\mathbf{Z}, \mathbf{W}, \alpha)$. As mentioned before, $n^{1/2}$-consistent estimation of α is possible by the methods in Chapter 4. It is also possible to constructed asymptotically efficient or nearly efficient score tests based on nonparametric estimates of $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$. Stefanski and Carroll (1990a, 1991) constructed semiparametric tests that achieve full or nearly full efficiency when $\mathbf{W}$ is unbiased for $\mathbf{X}$ and its measurement error variance is known or independently estimated. Sepanski (1992) used nonparametric regression techniques to construct efficient tests when there exists an independent validation data set or an independent data set containing an unbiased instrumental variable.

The ROC curve is commonly used to assess the ability of a marker to diagnose the presence of a disease or other condition, for example, in Reiser (2000), serum creatine kinease is used to diagnose when a woman is a carrier of DMD (Duchenne muscular dystrophy). Reiser (2000) discusses the estimation of ROC curves when the marker is measured with error.

CHAPTER 11

LONGITUDINAL DATA AND MIXED MODELS

This chapter is concerned with mixed models and longitudinal/clustered data structures, ones that are more complex than simple random sampling. That is, in previous chapters we have described situations in which the observed data are $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i)$ for individuals $i = 1, ..., n$.

Actually, we have already described a simple example of such a more complex data structure, namely the OPEN data as analyzed in Section 9.11. As seen there, we had repeated measures $\mathbf{Y}_{ij}$ on each individual, rather than just a single observation $\mathbf{Y}_i$. Repeated measures are a type of clustered data and can be analyzed using mixed model technology.

This chapter is meant to give the reader an overview of some of the developments in mixed models with covariate measurement error. The linear mixed model (LMM) has, of course, been the format for most of the advances, but more recent developments have focused on nonlinear mixed models. Book-length treatments of mixed models are given by many authors, including Verbeke and Molenberghs (2000); McCulloch and Searle (2001); Ruppert, Wand, and Carroll, (2003); and Demidenko (2004).

11.1 Mixed Models for Longitudinal Data

11.1.1 Simple Linear Mixed Models

Longitudinal data arise when a sample of subjects is followed over a period of time and, for each subject, some or all variables are measured at multiple time points. The subjects are often called *clusters*. Longitudinal data are a special type of clustered data, one where time is an important component. Both longitudinal data and clustered data models are often analyzed by mixed model technology.

Mixed effects models are a natural extension of linear and generalized linear models for modeling clustered data. The simplest example of a mixed model is the mixed balanced one-way ANOVA model, where a single variable, call it $\mathbf{Y}$, is measured at J time points for each of I subjects. Thus, the data are $\mathbf{Y}_{ij}$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The model is

$$\begin{cases} \mathbf{Y}_{ij} & = & \mu + b_i + \epsilon_{ij}; \\ b_i & \sim & \text{Normal}(0, \sigma_b^2); \\ \epsilon_{ij} & \sim & \text{Normal}(0, \sigma_\epsilon^2). \end{cases} \qquad (11.1)$$

This model can be viewed as a compromise between two fixed effect models: one that assumes equal intercepts for all subjects ($\sigma_b^2 = 0$) and one that assumes a different intercept for each subject (effectively, $\sigma_b^2 = \infty$), and the estimate of b_i is a weighted average of $\overline{Y}_{\cdot\cdot}$, the grand average of all Y_{ij}, and $\overline{Y}_{i\cdot}$, the average of all Y_{ij} in the i^{th} sample. The mixed model with $0 < \sigma_b^2 < \infty$ allows each subject to have a different intercept but assumes that the intercepts are similar, with the degree of similarity increasing as σ_b^2 decreases.

An attractive property of model (11.1) is that observations in the same subject are correlated with correlation coefficient $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$. In a mixed model framework, the random coefficients b_i, $i = 1, \ldots, I$ are called *random effects*, their variance is called a *variance component*, and ρ is called the *within-subject*, or *within-cluster*, correlation.

Note how the OPEN data model (9.33) is a generalization of (11.1). It has the random effect (there called r_i and referred to as person-specific bias), but instead of a mean common to all individuals, it has a linear regression mean structure.

11.1.2 The General Linear Mixed Model

The general linear mixed model is (no surprise!) a generalization of the simple mixed model (11.1). In the general linear mixed model, the mean $\mu + b_i$ for each individual is replaced by a regression with random effects. Specifically, keeping to the notation of this book,

$$Y_{ij} = \beta_0 + \mathbf{X}_{ij}^t \beta_x + \mathbf{Z}_{ij}^t \beta_z + \mathbf{A}_{ij}^t \mathbf{b}_i + \epsilon_{ij}, \qquad (11.2)$$

where the random effects $\mathbf{b}_i$ that vary between subjects are assumed to have a normal distribution with mean zero and covariance matrix $D(\theta)$, depending on a parameter θ: in symbols, $\mathbf{b}_i \sim \text{Normal}\{0, \mathbf{D}(\theta)\}$. In addition, the ϵ_{ij} are mutually independent with mean zero and variance σ_ϵ^2. The regression parameters β_x and β_z that are constant between subjects are called *fixed effects*. The parameters in θ are called *variance components* or, more generally when $\mathbf{D}(\theta)$ is not diagonal, *covariance components*.

In what follows, except for the example described in Section 11.8.1, the covariates $\mathbf{Z}_{ij}$ and $\mathbf{A}_{ij}$ are assumed to be observed without error. A major reason for distinguishing between them, rather than putting them into a single vector $\mathbf{Z}_{ij}$ as done elsewhere in this book, is that the regression coefficients of $\mathbf{A}_{ij}$, namely $\mathbf{b}_i$, are random effects. Often many, if not all, covariates without error are in both $\mathbf{Z}_{ij}$ and $\mathbf{A}_{ij}$; see the examples in Section 11.2.3.

Note that marginally, the mean is our old friend $\beta_0 + \mathbf{X}_{ij}^t\beta_x + \mathbf{Z}_{ij}^t\beta_z$, but the random effects $\mathbf{b}_i$ induce correlations among the observations within an individual. Thus, the variance of Y_{ij} is $\mathbf{A}_{ij}^t D(\theta)\mathbf{A}_{ij} + \sigma_\epsilon^2$, while the covariance between Y_{ij} and Y_{ik} is $\mathbf{A}_{ij}^t D(\theta)\mathbf{A}_{ik}$.

11.1.3 The Linear Logistic Mixed Model

Mixed models are, of course, not confined to the linear model. For example, suppose that the response Y_{ij} is binary. Then the linear logistic mixed model is the natural modification of (11.2),

$$\text{pr}(Y_{ij} = 1|\mathbf{b}_i) = H(\beta_0 + \mathbf{X}_{ij}^t\beta_x + \mathbf{Z}_{ij}^t\beta_z + \mathbf{A}_{ij}^t\mathbf{b}_i), \qquad (11.3)$$

where, as usual, $H(\cdot)$ is the logistic distribution function.

The major differences between the linear mixed model (11.2) and the logistic mixed model (11.3) are (a) computation and (b) in interpretation of the fixed effects. Using the probit approximation to the logistic distribution function (Section 4.8.2), we see that marginally,

$$\text{pr}(Y_{ij} = 1) \approx H\left\{\frac{\beta_0 + \mathbf{X}_{ij}^t\beta_x + \mathbf{Z}_{ij}^t\beta_z}{(1 + \mathbf{A}_{ij}^t D(\theta)\mathbf{A}_{ij}/2.9)^{1/2}}\right\}. \qquad (11.4)$$

Because the $\mathbf{A}_{ij}$ can depend upon $\mathbf{Z}_{ij}$, the interpretation of, for example, β_x as the effect of changing $\mathbf{X}_{ij}$ is no longer correct; see Heagerty and Kurland (2001) for discussion.

11.1.4 The Generalized Linear Mixed Model

These ideas extend naturally to more complex models for longitudinal data and generate the flexible class of generalized linear mixed models (GLMMs). In a generalized linear mixed model, given the random effects $\mathbf{b}_i$, the responses $\mathbf{Y}_{ij}$ are assumed to have a distribution (normal, binomial, etc.), whose mean is given as $\mu_{ij,x}^{\mathbf{b}_i}$, where for some function $g(\cdot)$,

$$g(\mu_{ij,x}^{\mathbf{b}_i}) = \beta_0 + \mathbf{X}_{ij}^t\beta_x + \mathbf{Z}_{ij}^t\beta_z + \mathbf{A}_{ij}^t\mathbf{b}_i. \qquad (11.5)$$

Here $\mu_{ij,x}^{\mathbf{b}_i}$ is the expected value of $\mathbf{Y}_{ij}$, the j^{th} measured response on the i^{th} subject and $\mathbf{X}_{ij}$, $\mathbf{Z}_{ij}$, and $\mathbf{A}_{ij}$ are covariate vectors of dimension p_1, p_2, and q, respectively.

In the linear mixed model, $g(\cdot)$ is the identity function, while in the logistic mixed mode, $g(\cdot)$ is the inverse of the logistic distribution function, etc.

11.2 Mixed Measurement Error Models

The generalized linear mixed measurement error model (GLMMeM) model of Wang, Lin, Gutierrez, et al. (1998) starts with (11.5), but now allows for measurement error in the $\mathbf{X}_{ij}$.

In a GLMMeM, $\mathbf{X}_{ij}$ is not observed; instead one observes $\mathbf{W}_{ij}$ that is related to $\mathbf{X}_{ij}$. The analytic closed-form bias calculations in Wang, Lin, Gutierrez, et al. (1998) are obtained under the assumption of classical additive errors, that is,

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij}, \tag{11.6}$$

where the $\mathbf{U}_{ij}$ are independent Normal$(0, \Sigma_u)$, but (11.6) was not needed by these authors for numerical bias calculations, estimation, or inference.

11.2.1 The Variance Components Model Revisited

It is instructive to consider the OPEN study data model (9.33). In our random effects notation, we have that $\mathbf{X}_{ij} \equiv \mathbf{X}_i$, $\mathbf{Y}_{ij} = \beta_0 + \beta_1 \mathbf{X}_i + b_i + \epsilon_{ij}$, $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$. As it transpires, the observed data become a combination of two linear mixed models, and the unobserved $\mathbf{X}_i$ become random effects. To see this, let μ_x be the population mean of $\mathbf{X}$ and let σ_x^2 be the population variance of $\mathbf{X}$. Let $\mu_y = \beta_0 + \beta_x \mu_x$, and write $\Delta_{xi} = \mathbf{X}_i - \mu_x$. Then the observed data consist of two linear mixed models:

$$\begin{aligned} Y_{ij} &= \mu_y + (\beta_x \Delta_{xi} + b_i) + \epsilon_{ij}; \\ W_{ij} &= \mu_x + \Delta_{xi} + U_{ij}. \end{aligned}$$

The complication is that the two linear mixed effect models are correlated because Δ_{xi} occurs in both of them. It is an interesting exercise to combine these two linear mixed models into a single linear mixed model, although the notation is nasty.

There are two main points to this exercise:

- With effort, a linear mixed model with measurement error in covariates can first be turned into two linear mixed models with correlated components, and then with notational wizardry be turned into a single, albeit more complex, linear mixed model.

- Although it is a little difficult to see from this example, the fact that the variance of $\mathbf{X}$ shows up in the random effects here means that when handling a GLMMeM model, the variance structure of the covariates measured with error must be taken into account. The next subsection makes this more explicit.

11.2.2 General Considerations

An important general principle is that, under the assumption of additive error and a normal structural model, the effect of measurement error on a GLMM relating $\mathbf{Y}_{ij}$ to $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ is to create a new GLMM relating $\mathbf{Y}_{ij}$ to $\mathbf{W}_{ij}$ and $\mathbf{Z}_{ij}$. Stated differently, under the assumptions of additive error and a normal structural model, a GLMMeM in the true covariates becomes a GLMM in the observed covariates. The important consequence of this principle is that analytic expressions for bias and bias-correction can be found by comparing the parameters of the GLMMeM to those of the GLMM.

A more precise expression of this principle is given by Wang, Lin, Gutierrez, et al. (1998) Suppose that $\mathbf{X}_{ij}$ is scalar, and define the vector $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})^t$, and define $\mathbf{Z}_i$, $\mathbf{W}_i$, and $\mathbf{U}_i$ similarly. Suppose that

$$\mathbf{X}_i = \mathbf{1}_i \eta_0 + \mathbf{Z}_i \boldsymbol{\eta}_z + \boldsymbol{e}_{xi},$$

where $\mathbf{1}_i$ is an $n_i \times 1$ vector of ones and $\boldsymbol{e}_{xi}$ given $\mathbf{Z}_i$ is Normal$(0, \Sigma_{xxi})$. Also define $\Lambda_i = \Sigma_{xxi}\{\Sigma_{xxi} + \text{cov}(\mathbf{U}_i)\}^{-1}$. Then

$$\mathbf{X}_i = (I_i - \Lambda_i)(\mathbf{1}_i \eta_0 + \mathbf{Z}_i \boldsymbol{\eta}_z) + \Lambda_i \mathbf{W}_i + \mathbf{b}_i^*, \tag{11.7}$$

where $\mathbf{b}_i^*$ is independent of $\mathbf{b}_i$ and $\mathbf{W}_i$. It follows from (11.7) that

$$\mathbf{X}_{ij} = \alpha_{0j} + \boldsymbol{\eta}_z^t \mathbf{Z}_i^t \boldsymbol{\alpha}_{zj} + \mathbf{W}_i^t \boldsymbol{\alpha}_{wj} + \boldsymbol{C}_{ij}^t \mathbf{b}_i^* \tag{11.8}$$

for some α_{0j}, $\boldsymbol{\alpha}_{zj}$, $\boldsymbol{\alpha}_{wj}$, and $\boldsymbol{C}_{ij}$.

It follows from (11.5) and (11.8) that the observed data $(\mathbf{Y}_i | \mathbf{W}_i, \mathbf{Z}_i)$ follow the GLMM with mean

$$\begin{aligned} g(\mu_{ij,w}^{\mathbf{b}_i^*}) &= (\beta_0 + \alpha_{0j}\beta_x) + \mathbf{W}_i^t \boldsymbol{\alpha}_{wj} \beta_x + (\boldsymbol{\eta}_z^t \mathbf{Z}_i^t \boldsymbol{\alpha}_{zj} \beta_x + \mathbf{Z}_{ij}^t \beta_z) \\ &\quad + (\mathbf{A}_{ij}^t \mathbf{b}_i + \boldsymbol{C}_{ij}^t \beta_x \mathbf{b}_i^*). \end{aligned} \tag{11.9}$$

Note specifically how the variance structure of the $\mathbf{X}_{ij}$ becomes an important consideration when properly handling a GLMMeM.

11.2.3 Some Simple Examples

To illustrate LMMs, GLMMs, and GLMMeMs, this section contains several simple, hypothetical examples. Suppose that on the j^{th} yearly visit of the i^{th} subject to a clinic we observe the subject's systolic blood pressure $\mathbf{Y}_{ij}$ and age $\mathbf{Z}_{ij}$, and that there is a linear relationship between these two variables with a subject-specific intercept and slope. Then a suitable LMM is

$$\mathbf{Y}_{ij} = \beta_0 + \mathbf{Z}_{ij}\beta_z + \mathbf{A}_{ij}\mathbf{b}_i + \epsilon_{ij}, \tag{11.10}$$

where $\mathbf{A}_{ij} = (\,1 \quad \mathbf{Z}_{ij}\,)$ and $\mathbf{b}_i = (\,b_{0,i} \quad b_{z,i}\,)^t$. Here β_0 and β_z are the average intercept and slope across the population of all potential subjects and $b_{0,i}$ and $b_{z,i}$ are the deviations of the i^{th} subject's intercept and slope from average. The matrix of covariance components is

$$\mathbf{D}(\theta) = \begin{bmatrix} \text{var}(b_{0,i}) & \text{cov}(b_{0,i}, b_{1,i}) \\ \text{cov}(b_{0,i}, b_{1,i}) & \text{var}(b_{1,i}) \end{bmatrix},$$

and $\theta = \{\text{var}(b_{0,i}), \text{var}(b_{1,i}), \text{cov}(b_{0,i}, b_{1,i})\}^t$ contains the unique components of $\mathbf{D}$. Now suppose that $\mathbf{Y}_{ij}$ is also related to a true nutrient intake, $\mathbf{X}_{ij}$, over the previous year. If the regression coefficient for $\mathbf{X}_{ij}$ is a fixed effect, that is, independent of the subject, then the relationship between $\mathbf{Y}_{ij}$, $\mathbf{X}_{ij}$, and $\mathbf{Z}_{ij}$ could be modeled by the LMM

$$\mathbf{Y}_{ij} = \beta_0 + \mathbf{X}_{ij}\beta_x + \mathbf{Z}_{ij}\beta_z + \mathbf{A}_{ij}\mathbf{b}_i + \epsilon_{ij}. \qquad (11.11)$$

If the true intake is unobserved and the observed intake is $\mathbf{W}_{ij}$, then we have a linear mixed measurement error model (LMMeM, a special case of a GLMMeM).

As mentioned above, for additive errors and a normal structural model, measurement error's effect on a GLMM for $\mathbf{Y}_{ij}$, $\mathbf{X}_{ij}$, and $\mathbf{Z}_{ij}$ is to induce a new GLMM relating $\mathbf{Y}_{ij}$ to $\mathbf{W}_{ij}$ and $\mathbf{Z}_{ij}$. To illustrate this principle, we use the fact that if $\mathbf{X}_{ij}$ is Normal(μ_z, σ_x^2) and independent of $\mathbf{Z}_{ij}$, if the $\mathbf{X}_{ij}$ are mutually independent, and if the classical additive error model (11.6) holds, then we have a regression calibration model

$$\mathbf{X}_{ij} = \gamma_0 + \lambda\mathbf{W}_{ij} + \mathbf{b}^*_{ij}, \qquad (11.12)$$

where the $\mathbf{b}^*_{ij}$ are mutually independent and independent of $\mathbf{W}_{ij}$, $\gamma_0 = (1-\lambda)\mu_x$, and λ is the attenuation; see Section 2.2.1. Substituting (11.12) into (11.11), one obtains

$$\mathbf{Y}_{ij} = (\beta_0 + \beta_x\gamma_0) + \mathbf{W}_{ij}\gamma_1\beta_x + \mathbf{Z}_{ij}\beta_z + \mathbf{A}_{ij}\mathbf{b}_i + \epsilon^*_{ij}, \qquad (11.13)$$

where $\epsilon^*_{ij} = \beta_x\mathbf{b}^*_{ij} + \epsilon_{ij}$. Clearly, (11.13) is an LMM.

Comparing the parameters in (11.11) to those in (11.13) gives analytical expressions for the asymptotic biases of the naive estimator, because the naive estimator will be consistent for the parameters in (11.13). For the fixed effects, these biases are the same as discussed in Chapter 3. The naive estimator of the covariance components matrix $\mathbf{D}$ is unbiased, since the random effects part of the model remains $\mathbf{A}_{ij}\mathbf{b}_i$. The only variance component for which the naive estimator is biased is σ^2_ϵ, since the "error" in (11.13) is $\epsilon^*_{ij} = \beta_x\mathbf{b}^*_{ij} + \epsilon_{ij}$, so that the naive estimator is consistent for $\beta_x^2\text{var}(\mathbf{b}^*_{ij}) + \sigma^2_\epsilon$. This is an example of another general principle: Naive estimates of variance parameters typically are either unbiased or biased upward, because the variation included by measurement error is not modeled and so is attributed to the random effects or error.

The unbiasedness of the naive estimator of $\mathbf{D}$ in this example is due to the restrictive assumption that the $\mathbf{X}_{ij}$ are mutually independent; this assumption is called the "homogenous model" by Wang et al. In many examples, there will be within-subject correlation between the $\mathbf{X}_{ij}$, and this correlation causes $\mathbf{D}_{11}$ to be biased upward. Wang, Lin, Gutierrez, et al. (1998) have a "heterogenous" model for use in such examples. As should be clear from our discussion, and as is made explicit in Wang et al. (1998, Section 4), if the homogeneous model is fit when the heterogeneous model holds, then biases occur, both in the fixed effects and in the random effects. Hence, as mentioned previously, in fitting a mixed model with measurement error, it is important to consider the structure of the **X**-variables within each individual.

11.3 A Bias-Corrected Estimator

An early study of measurement error in longitudinal modeling is Tosteson, Buonaccorsi, and Demidenko (1998). These authors assume that for the i^{th} subject, one observes a t-dimensional vector $\mathbf{Y}_i$ of responses, which corresponds to a t-dimensional vector $\mathbf{X}_i$ of true covariate values. The observed covariates values are $\mathbf{W}_i = \mathbf{X}_i + \mathcal{U}_i$, where the measurement error vector $\mathcal{U}_i$ has t iid Normal$(0, \sigma_u^2)$ components. In their example, $\mathbf{Y}_i$ contains five yearly observed plasma beta-carotene levels and $\mathbf{W}_i$ contains the corresponding values of observed beta-carotene intakes measured by food frequency questionnaires (FFQ) given at the same times as the plasma beta-carotene assays. Thus, $\mathbf{X}_i$ is defined as the corresponding true beta-carotene intakes, each over the year prior to the FFQ.

An important feature of their model is that they assume neither validation data nor replication of the measurements. For example, in their application one never observes true intakes of beta-carotene and no subject fills out more that one FFQ at any yearly visit. Of course, one could have some subjects complete two FFQs at some visits, but these would not be true replicates because their errors would be highly correlated. We have seen that measurement error models are usually not identified in the absence of validation or replication data. However, for longitudinal data, the repeated measurements can substitute for replication and, as will be seen, allow parameter identifiability, at least if one is willing to put structure on the mean of the **X**-values. See also Higgins, Davidian, and Giltinan (1997), who noted the same point. The independence of the measurement errors if, of course, an assumption in itself.

The model of Tosteson et al. for $\mathbf{Y}_i$ given $\mathbf{X}_i$ is

$$\mathbf{Y}_i = \mu + \Gamma \mathbf{X}_i + \mathcal{Z} v_i + \epsilon_i, \tag{11.14}$$

where Γ is a parameter matrix, $\mathcal{Z}$ is a known design matrix, and v_i is a Normal$(0, \Omega)$ random effect where Ω is unknown. In addition, ϵ_i is Normal$(0, \sigma_\epsilon^2)$.

They found that the naive estimator of Γ is attenuated by the factor $\Sigma_T(\Sigma_T + \sigma_{\mathbf{u}}^2 I)^{-1}$, but at this level of generality it is apparently not possible to get explicit results for the bias of the naive estimator of the (co)variance component matrix Ω.

Many of their further results assume that $\Gamma = \gamma I$, where I is the $t \times t$ identity matrix and γ is a scalar parameter.

They assume a structural model

$$\mathbf{X}_i = \mu_{\mathbf{X}} + R\phi_i, \tag{11.15}$$

where R is a known $t \times q$ design matrix, $q < t$, and ϕ_i is a Normal$(0, \Omega_T)$ random effect. The assumption that $q < t$ is crucial and implies that the (co)variance component matrix Ω_T has only $q(q+1)/2$, rather than $t(t+1)/2$ unique components. This dimension-reduction identifies the parameters of the model, even though there are no replicate measurements of the components of $\mathbf{X}_i$.

Typical choices of R are $R_1 = (1 \; 1 \; \cdots \; 1)^t$ and

$$R_2 = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & t \end{pmatrix}^t. \tag{11.16}$$

Using R_1 implies that components of $\mathbf{X}_i$ are constant, each equaling $\mu_{\mathbf{X}} + \phi_i$. In this case, the components of $\mathbf{W}_i$ are true replicates. However, this assumption is often suspect, and then R_2 might be more reasonable, since R_2 implies that the components of $\mathbf{X}_i$ follow a linear time trend. Note that R_2 can be used only if $t \geq 3$, since $q = 2$ for R_2.

Under the assumption that $\mathcal{Z} = R$, Tosteson et al. obtained explicit results for the bias of the naive estimator of Ω, the matrix of (co)variance components of the random effects, v_i. In particular, they found that the naive estimator of Ω is positively biased, since it is estimating both Ω and extra variability due to measurement error.

Tosteson et al. reparameterized their model into two parameter vectors: one for the marginal density of $\mathbf{W}_i$ and the other for the conditional density of $\mathbf{Y}_i$ given $\mathbf{W}_i$; this is different from our discussion in Section 11.2.1. Doing this allows all parameters to be estimated by standard methods using currently available software; they used SAS PROC MIXED, though S-PLUS or R could be used. They mentioned that this “bias corrected” estimator is not the MLE, but they conjectured that it is highly efficient. Their conjecture turned out to be false, since they showed in Buonaccorsi, Demidenko, and Tosteson (2000) that maximum likelihood and pseudo likelihood estimators can be considerably more efficient.

11.4 SIMEX for GLMMEMs

Wang, Lin, Gutierrez, et al. (1998) have studied the SIMEX method for estimation of the parameter in a GLMMeM. They found the SIMEX is straightforward to apply and effective for removing measurement error induced bias. They used the quadratic extrapolation function. SIMEX is, of course, applied to some naive estimator, that is, an estimator that would be used if there were no measurement error. For GLMMs there are several possible choice for the naive estimator. Wang et al. (1998) used the corrected penalized quasilikelihood method (CPQL) of Breslow and Lin (1995) and Lin and Breslow (1996).

11.5 Regression Calibration for GLMMs

As we have seen in Chapter 4, regression calibration is a simple and effective method for estimating the parameters in a GLM with covariate measurement error. However, a naive application of regression calibration is not suitable for GLMMeMs (Wang, Lin, and Gutierrez, 1999). The reason for this is that substituting $E(\mathbf{X}|\mathbf{W}, \mathbf{Z})$ for $\mathbf{X}$ in a GLMM correctly specifies that fixed-effects structure, but not the random-effects structure. Therefore, the bias of the naive estimators of variance components is not corrected properly by regression calibration. Wang, Lin, Gutierrez, et al. (1998) stated that since, in general models such as logistic regression, fixed-effects parameters and variance components are not orthogonal, the fixed-effects parameters estimates will also be biased.

Despite these difficulties, Buonaccorsi, Demidenko, and Tosteson (2000) have found regression calibration suitable and, in fact, highly efficient for estimation of fixed-effects in *linear* mixed models, a special case in which fixed-effects parameters and variance components *are* orthogonal. Moreover, in the context of linear mixed models, they showed how one can correct the bias of the regression calibration estimates of the variance components. The “corrected regression calibration” method equals the pseudomaximum likelihood estimator discussed in Section 11.6.

It is worth reiterating the point made in Section 11.2.4, namely, that regression calibration requires that the within-subject correlation structure of the $\mathbf{X}$-values be properly specified.

11.6 Maximum Likelihood Estimation

Buonaccorsi, Demidenko, and Tosteson (2000) continued the study of the linear mixed models in Tosteson, Buonaccorsi, and Demidenko (1998) and compared the bias-corrected estimator in Tosteson et al. (1998) with the maximum likelihood and pseudomaximum likelihood estimators. They partitioned the parameters into two vectors: θ_1, which contains the parameters in the model for $[\mathbf{Y}|\mathbf{X},\mathbf{Z}]$, and θ_2, which contains the parameters in the model for $[\mathbf{W}|\mathbf{X},\mathbf{Z}]$. The likelihood for the observed data is the product of the likelihood $f(\mathbf{Y}|\mathbf{W},\mathbf{Z};\theta_1,\theta_2)$ for $\mathbf{Y}$ given $(\mathbf{W},\mathbf{Z})$ and the likelihood $f(\mathbf{W}|\mathbf{Z};\theta_2)$ for $\mathbf{W}$ given $\mathbf{Z}$. Maximum likelihood maximizes the product $f(\mathbf{Y}|\mathbf{W},\mathbf{Z};\theta_1,\theta_2)f(\mathbf{W}|\mathbf{Z};\theta_2)$. Pseudomaximum likelihood (Gong and Samaniego, 1981) estimates θ_2 by maximizing $f(\mathbf{W}|\mathbf{Z};\theta_2)$ and then maximizes $f(\mathbf{Y}|\mathbf{W},\mathbf{Z};\theta_1,\theta_2)$ over θ_1 with θ_2 held fixed at this prior estimate. They showed that the pseudomaximum likelihood estimator equals their corrected regression calibration estimator mentioned in Section 11.5.

In a study of efficiency, Buonaccorsi, Demidenko, and Tosteson (2000) showed that the pseudomaximum likelihood has nearly the same efficiency as full maximum likelihood, but the bias-corrected estimator in Tosteson et al. (1998) has a noticeably lower efficiency.

11.7 Joint Modeling

As discussed in Section 7.3.3.4, joint modeling (Wang, Wang, and Wang, 2000) refers to the use of subject-specific random-effects parameters from a mixed model as covariates in a second model. Typically, the random-effects parameters serve as a summary of a series of measurements thought to be related to the outcome in the second model. For example, researchers have investigated child-specific linear trends in BMI (body mass index) between 3 and 5 years of age and related these parameters to adult obesity. Wang, Wang, and Wang (2000) found that both the initial BMI at age 3 and the slope of the linear trend between 3 and 5 years of age had a significant effect on the risk of adult obesity. Clearly, relating the risk of adult obesity to the subject-specific intercept and slope of BMI is a more insightful analysis than relating the risk directly to numerous measurements of BMI taken on each individual. The intercepts and slopes are comparable across individuals, while the BMI measurements themselves may not be taken at the same age for all individuals and therefore may not be directly comparable.

In another application of joint modeling, Li, Zhang, and Davidian (2004) presented an example where progesterone levels (PDG) in women, as well as a number of baseline covariates, are related to bone mineral density (BMD) in the hip. PDG varies over the menstrual cycle. Although subjects vary in cycle length, the cycle were standardized to 28 days. During the 28-day standard cycle, log PDG stays constant during the first 14 days and then rises linearly for 7 days before decreasing linearly at the same rate for the remaining 7 days. Thus, the pattern of PDG fluctuation can be described by two parameters: the intercept, which is the baseline level during the first 14 days, and the slope, which is the linear rate of increase or decrease during the last 14 days. Although this general pattern is constant across women, the intercept and slope parameters are subject-specific. Li et al. used these parameters as covariates in a model where the response $\mathbf{Y}$ is absence of osteopenia, which is defined as BMD above the 33rd percentile. However, the intercept and slope for any subject are unknown and only longitudinal PDG measurements are available, so the intercept and slope are estimated with error. Li et al. (2004) use several of the estimators discussed in this section and find that the subject-specific intercept is not related to the absence of osteopenia ($p \approx 0.5$, depending slightly upon the method) but the subject-specific might be ($0.07 \le p \le 0.11$ for the various methods). As in the previous example, regressing the absence of osteopenia on the intercept and slope is a better summary of the data than relating the absence of osteopenia directly to the PDG values.

A number of estimators have been developed for joint modeling. Wang, Wang, and Wang (2000) proposed a pseudoexpected estimating equation estimator (EEE) (Wang and Pepe, 1999), a regression calibration estimator (RC), and a refined regression calibration estimator. They found that the RC estimator was biased in nonlinear models. The EEE estimator performs well but requires numerical integration. The refined RC estimator does not require numerical integration and its performance is close to that of the EEE estimator. Li, Zhang, and Davidian (2004) proposed two functional estimators, the sufficiency estimator and the conditional score estimator, both based upon Stefanski and Carroll (1987); see Section 7.3.3.4. Li, Zhang, and Davidian (2005) studied two flexible structural estimators, maximum likelihood and maximum pseudolikelihood, using the seminonparametric (SNP) structural model. Full maximum likelihood requires numerical integration, and Li et al. used Gauss–Hermite quadrature, though Monte Carlo integration could be used.

11.8 Other Models and Applications

11.8.1 Models with Random Effects Multiplied by $\mathbf{X}$

Previously, we have assumed that the random effects $\mathbf{b}_i$ have covariate vectors $\mathbf{A}_{ij}$ that are observed exactly. This need not be the case, of course.

Liang, Wu, and Carroll (2003) considered the varying coefficient linear mixed model, where random effects multiply $\mathbf{X}$ as well as on $\mathbf{Z}$. While they worked in great generality and included the use of regression splines (Section 12.2.2 and Chapter 13), their essential idea can be seen in the varying coefficient model, namely,

$$Y_{ij} = \beta_{0i} + \mathbf{X}_{ij}^t \beta_{xi} + \mathbf{Z}_{ij}^t \beta_{zi} + \epsilon_{ij}.$$

This is a simple linear regression model, where the regression lines depend on the individual or cluster. If we define $\mathbf{A}_{ij} = (1, \mathbf{X}_{ij}^t, \mathbf{Z}_{ij}^t)^t$, then the model becomes a linear mixed model:

$$\begin{aligned} Y_{ij} &= \beta_0 + \mathbf{X}_{ij}^t \beta_x + \mathbf{Z}_{ij}^t \beta_z + b_{i0} + \mathbf{X}_{ij}^t b_{ix} + \mathbf{Z}_{ij}^t b_{iz} + \epsilon_{ij} \\ &= \beta_0 + \mathbf{X}_{ij}^t \beta_x + \mathbf{Z}_{ij}^t \beta_z + \mathbf{A}_{ij}^t \mathbf{b}_i + \epsilon_{ij}. \qquad (11.17) \end{aligned}$$

If we now substitute (11.12) into (11.17), we see that the observed data no longer follow a standard mixed model, because now the random effects $\mathbf{b}_i$ in (11.17) are multiplied by the induced random effects $\mathbf{b}_{ij}^*$ in (11.12). Liang et al. (2003) fit this model using regression calibration.

11.8.2 Models with Random Effects Depending Nonlinearly on X

Higgins, Davidian, and Giltinan (1997) and Wu (2002) describe an nonlinear mixed effects model where the random effects themselves depend on covariates measured with error. The general form of the model is that

$$\begin{aligned} \mathbf{Y}_{ij} &= m_{\mathbf{Y}}(\mathbf{Z}_{ij}, \mathcal{B}_{ij}) + \epsilon_{ij}; \\ \mathcal{B}_{ij} &= d(\mathbf{X}_{ij}, \beta_x, \mathbf{b}_i), \end{aligned}$$

for known functions $m_{\mathbf{Y}}(\cdot)$ and $d(\cdot)$. Note here how the random effects $\mathcal{B}_{ij}$ depend on a subject-level random effect $\mathbf{b}_i$ as well as the true but unobserved covariates $\mathbf{X}_{ij}$. It is assumed that $\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij}$, where the measurement errors $\mathbf{U}_{ij}$ are independent with variance σ_u^2. As in Section 11.3, there are no replicate data to understand the measurement error properties, so a model is used along the lines of (11.15). For example, Wu assumed that $\mathbf{X}_{ij} = \mathbf{Z}_{ij}^*(\mu_{\mathbf{X}} + \alpha_i)$, where $\mathbf{Z}_{ij}^*$ are observed and α_i are independent random effects. Higginset al. used a regression calibration approach to fit the models, while Wu used the EM-algorithm.

11.8.3 Inducing a True-Data Model from a Standard Observed Data Model

Throughout this book, we have taken a common approach:

- Begin with a model for the data as if $\mathbf{X}$ could be observed, the so-called underlying true-data model.
- Specify the error model properties.

- Measurement error then induces a model for the observed data. This observed data model may be more or less standard.
- Use this observed data model to estimate the underlying true-data model.

Zidek, Le, Wong, et al. (1998) and Zidek, White, Le, et al. (1998) took exactly the opposite approach, in a clustered-data situation. Specifically, they specified a standard GLMM for the *observed* data, and then, by Taylor series expansion and the like attempt to approximate the underlying true-data model. The advantage of their approach is that standard models for observed data are, at least in principle, easier to check. The disadvantage of course is that it is easily possible that the underlying true-data model derived by their approach may be nearly unrecognizable, except if approximations are made to do so.

Fairly roughly, in their particular instance, Zidek, Le, Wong, et al. (1998) started with a model in which they specified the mean $E(\mathbf{W}_{ij}|\mathbf{Z}_i)$ and $\text{cov}(\mathbf{W}_{ij}, \mathbf{W}_{ik}|\mathbf{Z}_i) = G_{ik}$, where $\mathbf{Z}_i$ is the collection of observed $\mathbf{Z}$-values for the i^{th} person. Let $\mathbf{W}_i$ be the collection of observed $\mathbf{W}$-values for the i^{th} person. In our parlance, this is basically specifying the regression of $\mathbf{X}$ on $\mathbf{Z}$ as well as the measurement error variance. They then assumed a model for the observed data: For some parameter α, a known function $\mathcal{G}$, and letting $\mathbf{A}_{ij}$ depend on $(\mathbf{Z}_{ij}, \mathbf{W}_{ij})$, their model is

$$\begin{aligned} \mathcal{V}_{ij} &= \beta_0 + \mathbf{W}_{ij}^t \beta_w + \mathbf{Z}_{ij}^t \beta_z + \mathbf{A}_{ij}^t(\mathbf{W}_{ij}, \mathbf{Z}_{ij})\mathbf{b}_i; \\ E(\mathbf{Y}_{ij}|\mathbf{W}_i, \mathbf{Z}_i, \mathbf{b}_i) &= m_{\mathbf{Y}}(\mathcal{V}_{ij}); \qquad (11.18) \\ \text{cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{ik}|\mathbf{W}_i, \mathbf{Z}_i, \mathbf{b}_i) &= \mathcal{G}(\alpha, \mathcal{V}_{ij}). \qquad (11.19) \end{aligned}$$

With all the nonlinearities in this model, the hard part clearly is to go from this model for the observed data to the underlying true-data model. If one were willing to make the assumption that the random effects $\mathbf{b}_i$ were independent of the $\mathbf{W}$-values and the $\mathbf{Z}$-values, this can be done directly by numerical integration. Zidek et al. instead used various clever Taylor series approximations to obtain and approximate version of the underlying true-data model.

11.8.4 Autoregressive Models in Longitudinal Data

Schmid, Segal, and Rosner (1994) have an interesting early discussion of measurement error in the longitudinal linear mixed model when data within a person have an autoregressive error structure. Specifically, they allowed covariates $(\mathbf{Z}_{i,\text{ps}}, \mathbf{X}_{i,\text{ps}})$ that are person-specific but do not vary with time, and covariates $(\mathbf{Z}_{ij,\text{tv}}, \mathbf{X}_{ij,\text{tv}})$ that vary with time, so that in our notation $\mathbf{X}_{ij} = (\mathbf{X}_{i,\text{ps}}, \mathbf{X}_{ij,\text{tv}})$, $\mathbf{W}_{ij} = (\mathbf{W}_{i,\text{ps}}, \mathbf{W}_{ij,\text{tv}})$ and $\mathbf{Z}_{ij} = (\mathbf{Z}_{i,\text{ps}}, \mathbf{Z}_{ij,\text{tv}})$. As seen, for example, in Section 11.2.3, one must specify correctly the variance components structure of the measurement errors in $\mathbf{W}_{ij}$ and the variance components structure of the true covariates $\mathbf{X}_{ij}$ in order to obtain valid inferences. Schmid et al. found that when this is done, the maximum likelihood estimate is nearly unbiased and has good inference properties such as confidence intervals.

11.9 Example: The CHOICE Study

The Choices for Healthy Outcomes in Caring for ESRD (CHOICE) study is a multicenter prospective study to investigate treatment choices and outcomes of dialysis care among patients beginning dialysis. Its rationale and design have been reported by Powe, Klag, Sadler, et al. (1996). Briefly, the CHOICE study recruited 1,041 incident dialysis patients from 81 DCI (Dialysis Clinic Inc.) clinics between 1995 and 1998. Eligibility criteria for CHOICE study included ability to provide informed consent for participation, age older than 17 years, and ability to speak English or Spanish. Patients were enrolled a median of 45 days from initiation of chronic dialysis (98% within 4 months), 54% of the cohort had diabetes at baseline and 51% of the cohort died by December 1, 2001.

Dialysis population is subject to high risk of inflammation. The white blood cell (WBC) count and the C-reactive protein (CRP) are both inflammatory markers but may reflect different physiologic changes. How WBC correlates with CRP after the initiation of dialysis remains unknown. We use the CHOICE data to describe the longitudinal association between WBC and CRP and examine whether dynamic changes are different between hemo and peritoneal dialysis.

Because dialysis patients have a high death rate, a complete analysis would require the joint modeling of the survival and biomarker processes. However, for illustration purposes we focus only on subjects who have died during the study. We further restrict our data set to those subjects who have at least one WBC and one CRP measure during the same visit. Our subset of the data contained 373 subjects from 28 dialysis clinics. The analysis of these data is complicated by the expected nonlinear longitudinal trajectory of the biomarkers after initiation of dialysis, clustering of observations within subject and of subjects within clinics, and measurement error in CRP. Information on the CRP measurement error is available from a blinded duplicate CRP assay conducted on 42 subjects.

11.9.1 Basic Model

The following mixed model was used for fitting the data:

$$\left\{\begin{array}{lll} \mathbf{Y}_{ijc} & \sim & \text{Normal}\{f_{D(i)}(t_{ij}) + \mathbf{Z}_i^t\beta_z + \mathbf{X}_{ij}\beta_x + r_{ic}, \sigma_\epsilon^2\} \\ \mathbf{W}_{ij} & \sim & \text{Normal}(\mathbf{X}_{ij}, \sigma_u^2) \\ r_{ic} & \sim & \text{Normal}(s_c, \sigma_r^2) \\ s_c & \sim & \text{Normal}(\mu_s, \sigma_s^2), \end{array}\right. \tag{11.20}$$

where $\mathbf{Y}_{ijc}$ is the (log) white blood cell count for subject i at the visit j at the clinic c. The temporal effect is captured by the functions $f_{D(i)}(t_{ij})$, where $D(i)$ is the dialysis type indicator for subject i. Both functions, $f_0(t)$ and $f_1(t)$, were modeled as linear regression splines, that is, piecewise linear functions, with three knots at $5, 10, 15$ months after initiation of chronic dialysis. The two covariates, $\mathbf{Z}_i$, that were not subject to measurement error were age at baseline and sex. CRP was subject to measurement error and true (log) CRP was denoted by $\mathbf{X}_{ij}$, while observed (log) CRP was denoted by $\mathbf{W}_{ij}$. The subject-level random effect r_{ic} is assumed to have a normal distribution with mean equal to the clinic level mean s_c and variance σ_r^2, while the clinic means were assumed to have a normal distribution with mean equal to the overall mean and variance σ_s^2. A useful measure to describe the sources of residual variability is $R = \sigma_r^2/(\sigma_r^2 + \sigma_s^2)$, which is the fraction of within and between clinic variability attributable to within clinic variability.

11.9.2 Naive Replication and Sensitivity

In order to understand the measurement error variance σ_u^2, we need some version of a replication study. The subtlety here is that there are two sources of measurement error. The first is that what we measure is short-term (log) CRP, and CRP at a specific time does not fully characterize the average CRP, much as the OPEN protein biomarker measured short-term protein intake. The second source of measurement error is the assay variability, that is, the error that the laboratory makes in biochemically measuring a sample. In most cases, the assay variability is small relative to the variability of the short-term measurement around its long-term average, and if only assay variability is taken into account, little correction for measurement error will be made.

Suppose, for example, that all one considers is laboratory variability. Then a replication model is

$$\left\{\begin{array}{lll} \mathbf{W}_{lk}^{(r)} & \sim & \text{Normal}(\mathbf{X}_l^{(r)}, \sigma_u^2) \\ \mathbf{X}_l^{(r)} & \sim & \text{Normal}(\mu_x^{(r)}, \sigma_x^2), \end{array}\right. \tag{11.21}$$

where (r) indicates that these data come from a separate replication study. Here $\mathbf{W}_{lk}^{(r)}$, $k = 1, 2$ are the two lab results of the same sample from subject $l = 1, \ldots, 42$. The joint analysis of models (11.20) and (11.21) was done using Bayesian inference based on MCMC sampling, as described in Chapter 9. What we can expect, of course, is that laboratory/assay variability will be small relative to biological variability and the variability of (log) CRP in the population, so that little correction for measurement error will occur.

To show this, Table 11.1 displays posterior medians and 95% credible intervals for several parameters of interest, based on 100,000 simulations from the joint distribution after discarding 10,000 burn-in samples. Not surprisingly, there is a strong, statistically significant correlation between (log) WBC and (log) CRP. Also, even after adjusting for CRP the effect of age is statistically significant and indicates smaller WBC corresponding to older subjects. Sex was not statistically significant. Another interesting finding is that most of the residual variability ($\sim 95\%$) is due to between-subject variability, with roughly 5% variability being due to between-clinic variability. Reflecting good measurement calibration of the (log) CRP, the posterior mean of the reliability parameter was 0.9994 with a confidence interval $[0.9989, 0.9997]$. This reliability is **of course misleading** because it reflects only the precision of measuring (log) CRP in a given blood sample and does not incorporate potential short term biological variability of (log) CRP. Because no direct data were available to assess the biological measurement error due to using one blood sample to represent the short term (log) CRP average, we conducted a sensitivity analysis using several smaller levels of reliability. Table 11.1 shows results if the biological reliability of (log) CRP were $\lambda = 0.9$ and $\lambda = 0.8$. Interestingly, none of the parameter inferences changes significantly, with the exception of the (log) CRP parameter, which changes by 18% from 0.074 to 0.087.

The top plot in Figure 11.1 displays the posterior means of $f_0(t)$ (solid line) and $f_1(t)$ (dashed line) for $t \leq 30$ months, corresponding to hemo and peritoneal dialysis, respectively. The bottom plot displays the posterior mean and 95% pointwise confidence intervals of $f_1(t) - f_0(t)$.

		CRP	Sex	Age	R
$\lambda > .999$	Point est.	.074	-.044	-.0027	.953
	Std. Err.	.005	.031	.0012	.033
$\lambda = .9$	Point est.	.080	-.045	-.0029	.953
	Std. Err.	.005	.031	.0012	.033
$\lambda = .8$	Point est.	.087	-.045	-.0029	.952
	Std. Err.	.006	.031	.0012	.033

Table 11.1 *Estimates and standard errors from the CHOICE data Bayesian analysis using the longitudinal model (11.20) that also accounts for clustering. "CRP" is the (log) CRP and "R" is $R = \sigma_r^2/(\sigma_r^2 + \sigma_s^2)$, which is the fraction of within and between clinic variability attributable to within clinic variability. Standard errors are obtained from the simulation algorithm. Different values of λ indicate the corresponding reliability, with $\lambda > .999$ corresponding to data that takes into account only laboratory measurement error and $\lambda = .8, .9$ corresponding to hypothetical levels of biological reliability.*

11.9.3 Accounting for Biological Variability

As described in Section 11.3, one way to get at the biological variability (measurement error) in longitudinal studies is by assuming a simple and reasonable model for the variable observed without error. Moreover, measurement error variance is identifiable as long as the number of degrees of freedom in the measurement error model is smaller than the number of observations per subject. We illustrate this methodology here by assuming the following model,

$$\left\{\begin{array}{lcl} \mathbf{Y}_{ijc} & \sim & \text{Normal}\{\mathbf{Z}_i^t\beta_z + \mathbf{X}_{ij}\beta_x + r_{ic}, \sigma_\epsilon^2\} \\ r_{ic} & \sim & \text{Normal}(s_c, \sigma_r^2) \\ s_c & \sim & \text{Normal}(\mu_s, \sigma_s^2), \end{array}\right. \tag{11.22}$$

where information about the unobserved process $\mathbf{X}_{ij}$ and measurement error variance is obtained from the model

$$\left\{\begin{array}{lcl} \mathbf{W}_{ij} & \sim & \text{Normal}(\mathbf{X}_{ij}, \sigma_u^2) \\ \mathbf{X}_{ij} & = & f_{D(i)}(t_{ij}) + v_i \\ v_i & \sim & \text{Normal}(\mu_v, \sigma_v^2). \end{array}\right. \tag{11.23}$$

Here the true unobserved (log) CRP process is assumed to be a dialysis-specific regression spline with three knots and subject specific random intercepts. The rest of the variability in the observed (log) CRP is assumed to be measurement error with variance σ_u^2. Note that σ_u^2 is estimable from the model without using the replicated lab data, which was not used in this model.

Table 11.2 reports posterior means and standard errors for several parameters of interest. Interestingly, the point estimator relating (log)

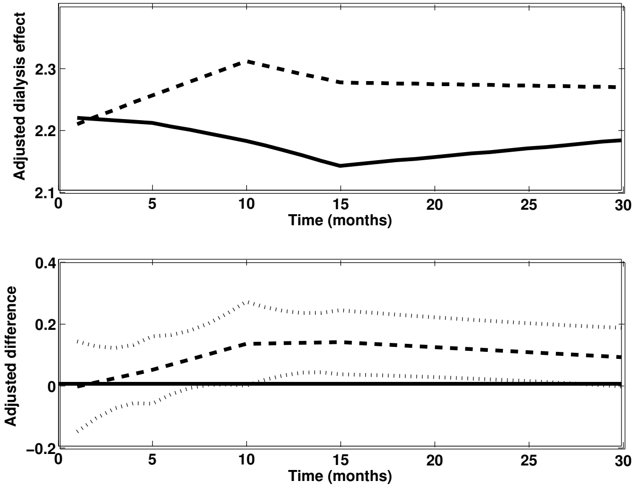


Figure 11.1 *Comparison of trajectories of (log) WBC for hemo and peritoneal dialysis patients adjusted for (log) CRP, age and sex. Top: adjusted population (log) WBC trajectories for Hemodialysis (solid line) and peritoneal dialysis (dashed line). Bottom: difference between adjusted population (log) WBC trajectories (peritoneal-hemo) with pointwise* 95% *confidence intervals.*

WBC to (log) CRP when biological measurement error is taken into account is 0.101, or 36% larger than the estimator based only on laboratory measurement error reported in Table 11.1. The standard error of the (log) CRP estimator has also increased from 0.005 to 0.014, or 180%. This is essentially due to much larger estimated biological measurement error variance, with a posterior mean equal to 0.91, which corresponds to a posterior reliability of 0.51 for the (log) CRP data. Results were practically unchanged for the other parameters.

Bibliographic Notes

Wang and Davidian (1996) were among the first authors to study the effects of measurement error on variance component estimators. They studied Berkson models and found that even a modest amount of measurement error could seriously bias the estimates of intrasubject variability.

As mentioned earlier, Higgins, Davidian and Giltinan (1997) and Tosteson, Buonaccorsi, and Demidenko (1998) discovered that if a mismeasured covariate **X** is observed longitudinally, then a structural model for **X** with dimension less than the number of **X**-observations per subject allows all parameters to be identified. Li, Shao, and Palta (2005) have another interesting application of this important concept, and they used a structural model different from that of Tosteson et al.; see also Wu (2002).

	CRP	Sex	Age	R
Point est.	.101	-.043	-.0028	.954
Std. err.	.014	.032	.0013	.032

Table 11.2 *Estimates and standard errors from the CHOICE data Bayesian analysis using the longitudinal model (11.22) with biological measurement error model (11.23) that also accounts for clustering. "CRP" is the (log) CRP and "R" is* $R = \sigma_r^2/(\sigma_r^2 + \sigma_s^2)$, *which is the fraction of within and between clinic variability attributable to within clinic variability. Standard errors are obtained from the simulation algorithm.*

The functional estimators for joint modeling in Li, Zhang, and Davidian (2004) are extended to multivariate longitudinal data by Li, Wang, and Wang (2005).

Ko and Davidian (2000) studied a two-component nonlinear model for longitudinal data. In the first component of the model, a vector of responses on a subject depends on covariates and a subject-specific parameter. In the second component of the model, the subject-specific parameters depend on covariates, random effects, and fixed effects. Some of the covariates in the second component are measured with error. Ko and Davidian presented an example from an AIDS clinical trial that shows the flexibility of this methodology.

CHAPTER 12

NONPARAMETRIC ESTIMATION

In this chapter, we give an overview of two nonparametric estimation problems that are of interest in their own right and also arise as secondary problems in regression calibration and hypothesis testing. The first problem is the estimation of the density of a random variable **X**, while the second is the nonparametric estimation of a regression, both when **X** is measured with error.

12.1 Deconvolution

12.1.1 The Problem

The fundamental problem of deconvolution is that of estimating the density of **X** when $\mathbf{W} = \mathbf{X} + \mathbf{U}$ is observed and the density of **U** is known. Closely related is the problem of estimating the regression function, $m(w) = E(\mathbf{X} \mid \mathbf{W} = w)$, when only $\mathbf{W} = \mathbf{X}+\mathbf{U}$ is observed and the density of **U** is known. The latter estimation problem is encountered in both regression calibration (Chapter 4) and hypothesis testing (Chapter 10).

There are at least three reasons for trying to understand the density function of **X**. Suppose that **X** is a continuous, scalar random variable, and that there are no covariates **Z** measured without error.

- Sometimes, the distribution of the latent **X** is of intrinsic interest, for example, in nutritional epidemiology, where **X** represents the usual intake of foods. In this case, let $f_X(x)$ be its density function. Then the distribution function is $\mathrm{pr}(\mathbf{X} \le c) = \int_{-\infty}^{c} f_X(x)dx$.
- When **X** is unobservable, likelihood methods (Chapter 8) require a model for the density of **X**. Regression calibration (Chapter 4) consists of the usual analysis but with **X** replaced by

$$\begin{aligned} m(\mathbf{W}) &= E(\mathbf{X}|\mathbf{W}) = \frac{1}{f_W(\mathbf{W})}\int x f_X(x) f_{w|x}(\mathbf{W}|x)dx \\ &= \frac{1}{f_W(\mathbf{W})}\int x f_X(x) f_U(\mathbf{W}-x)dx. \end{aligned} \tag{12.1}$$

- In Section 10.5, it was shown that when testing for the effect of the covariate measured with error, replacing **X** with an estimate of its regression $m(\mathbf{W})$ on **W** yields the hypothesis test with the highest local power (asymptotically).

Estimating the density function, f_X, of **X** is thus critical.

12.1.2 Fourier Inversion

The density function f_W is the convolution of f_X and f_U,

$$f_W(w) = \int f_X(x) f_U(w-x)dx, \tag{12.2}$$

and we thus refer to the problem of estimating f_X in the absence of parametric assumptions as *deconvolution*.

When both f_W and f_U are known, f_X is recovered by Fourier inversion. Letting ϕ_a denote the characteristic function of the random variable **A**, for example, $\phi_w(t) = \int e^{itw} f_W(w)dw$, we have that $\phi_x(t) = \phi_w(t)/\phi_u(t)$. Then by Fourier inversion,

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\phi_x(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\phi_w(t)}{\phi_u(t)}dt.$$

Even if, as we will now suppose, the density function f_U of **U** is known, the problem is complicated by the fact that the density of **W** is unknown and must be estimated. For the deconvolution problem under these assumptions, estimators with known rates of convergence were first obtained by Stefanski and Carroll (1986, 1990c), Carroll and Hall (1988) and Liu and Taylor (1989).

12.1.3 Methodology

We now describe a solution to the deconvolution problem. Statisticians have studied kernel density estimates of f_W of the form

$$\widehat{f}_w(w) = (nh)^{-1}\textstyle\sum_{i=1}^n K\left\{(\mathbf{W}_i - w)/h\right\},$$

where $K(\cdot)$ is a density function and h is the bandwidth, both chosen by the user. The function $\widehat{f}_w$ is itself a density function, with characteristic function $\widehat{\phi}_w$. It has long been known that for estimation of $f_W(w)$ the choice of kernel is relatively unimportant, and ease of use commonly dictates the choice of $K(\cdot)$, for example, the standard normal density or a density with bounded support.

It transpires that for commonly used kernels, the estimated density $\widehat{f}_X(x)$ cannot be deconvolved, in that the integral encountered in Fourier inversion is not defined. Stefanski and Carroll (1986, 1990c) showed that for certain smooth kernels, Fourier inversion of $\widehat{f}_X(x)$ is possible; see also Stefanski (1989). With an appropriately smooth kernel, the estimator,

$$\widehat{f}_x(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\widehat{\phi}_w(t)}{\phi_u(t)}dt,$$

exists, and for suitable choice of bandwidth is consistent for $f_X(x)$. The *deconvoluting kernel density estimator*, $\widehat{f}_x(x)$, integrates to one but is not always positive. It has the alternative representation

$$\widehat{f}_x(x) = (nh)^{-1}\textstyle\sum_{i=1}^n K_*\{(\mathbf{W}_i - x)/h, h\},$$

where the deconvoluting kernel is

$$K_*(t,h) = \frac{1}{2\pi}\int e^{ity}\frac{\phi_K(y)}{\phi_u(y/h)}dy.$$

The deconvoluting kernel density estimator has pointwise mean squared error

$$\begin{aligned} \text{MSE} &= E\left\{\widehat{f}_x(x) - f_X(x)\right\}^2 \\ &\sim ch^4 + (2\pi hn)^{-1}\int\left\{\frac{\phi_K(t)}{|\phi_u(t/h)|}\right\}^2 dt; \\ \text{where}\qquad c &= (1/4)\int x^2K(x)dx\int\left\{f_X^{''}(x)\right\}^2 dx. \end{aligned}$$

12.1.4 Properties of Deconvolution Methods

The best bandwidth, in the sense of minimizing MSE asymptotically, and the best MSE depend on the error density through its characteristic function ϕ_u. It is well known that in the absence of measurement error ($\mathbf{U} \equiv 0$), when f_X has two continuous derivatives the best MSE converges to zero at the rate $n^{-4/5}$. However, for nondegenerate **U**, convergence rates are much slower in general. The best rate of convergence depends on the tail behavior of $|\phi_u(t)|$, with lighter tails resulting in slower rates of convergence. The tail behavior of $|\phi_u(t)|$ is in turn related to the smoothness of $f_U(u)$ at $u = 0$, with smoother densities having characteristic functions with lighter tails.

For example, if **U** is normally distributed, then

$$|\phi_u(t)| = \exp(-\sigma_u^2 t^2/2)$$

is extremely light tailed, and the mean squared error converges to 0 at a rate no faster than the exceedingly slow rate of $\{\log(n)\}^{-2}$. The implication is that with normally distributed errors, it is not possible to estimate the actual value of $f_X(x)$ well. However, detailed analyses by Wand (1998) indicate that, for lower levels of measurement error, deconvolving density estimators can perform well for reasonable sample sizes.

If **U** has a more peaked density function than the normal, then $|\phi_u(t)|$ does not diminish to 0 as rapidly, and the deconvoluting kernel density estimator has better asymptotic performance. For example, consider the Laplace distribution with density function $f_U(u) = (1/\sigma_u\sqrt{2})\exp(-\sqrt{2}|u| /\sigma_u)$. In this case $\phi_u(t) = 2/(2+\sigma_u^2 t^2)$, and the optimal mean squared error converges to zero at the rate $n^{-4/9}$, tolerably close to the rate in the absence of measurement error, that is, $n^{-4/5}$.

The fact that smoothness of the error density determines how well f_X can be estimated is a disconcerting nonrobustness result. An open problem, of course, is how to construct deconvolution estimates that are adaptive to the amount of smoothness of the measurement error density.

We note that the slow rate of convergence of $\widehat{f}_X(x)$ is intrinsic to the deconvolution problem, and not specific to the deconvoluting kernel density estimator, which is known to achieve the best rate of convergence in general (Carroll and Hall, 1988; Stefanski and Carroll, 1990c).

However, rates of convergence are not always fully informative with regard to the adequacy of $\widehat{f}_x(x)$ for estimating the basic *shape* of $f_X(x)$. As shown in the examples below, despite the slow pointwise rate, the estimator itself can provide useful information about shape.

In applications, calculation of $\widehat{f}_x(x)$ requires specification or estimation of a bandwidth h. Stefanski and Carroll (1990c) described a bandwidth estimator when the improper sinc kernel, $K(t) = (\pi t)^{-1}\sin(t)$, is used. Stefanski (1990) showed that for a large class of kernels and a large class of error densities that includes the normal densities, the mean squared error is minimized asymptotically by a known sequence of bandwidths — the optimal bandwidth is $h = h_G = \sigma_u \{\log(n)\}^{-1/2}$ for normal (Gaussian) error. For Laplace measurement error and the kernel with characteristic function $\phi_K(t) = (1-t^2)^3$ when $|t| \le 1$ and zero otherwise, Fan, Truong, and Wang (1991) suggested taking $h_L = (1/2)\sigma_u n^{-1/9}$.

Fan and Truong (1993) and Carroll and Hall (2004) also considered the use of the deconvoluting kernel function $K_*(t,h) = \phi(x)\{1-\sigma_u^2(x^2-1)/(2h^2)\}$, which is the deconvoluting kernel when the errors have a Laplace density and the basic kernel function is the standard normal density $\phi(x)$. Carroll and Hall (2004) were more interested in regression function estimation and called their method Taylex; see also Section 12.2.7.

12.1.5 Is It Possible to Estimate the Bandwidth?

Not withstanding the previous comments, the fact that deconvolution is hard theoretically means that things that are hard to do in easy problems

will be nigh well impossible in this context. Specifically, we refer to the problem of actually estimating the bandwidth.

A simple example will suffice to make the point. Later on, we will study the Framingham data; see Section 12.1.9. These data have 1,615 observations with a reliability ratio of about 0.75, so this is hardly a nasty example. We applied both of the default methods described above to these data. Figure 12.1 is the result, and it is amusing. The default deconvolution method appropriate for Laplace errors is so wild that it swamps the default deconvolution method appropriate for Gaussian errors as well as the best-fitting normal approximation. We removed this totally ridiculous estimate and replotted, see Figure 12.2: This is not much better!

The point is that one should be wary, maybe even suspicious, of methods of automatic bandwidth selection in deconvolution kernel methods. We tend to think that the better method is to vary the bandwidth from smallest to larger and stop when the graph becomes reasonably smooth. The best that one can hope to get out of this is a look at shape.

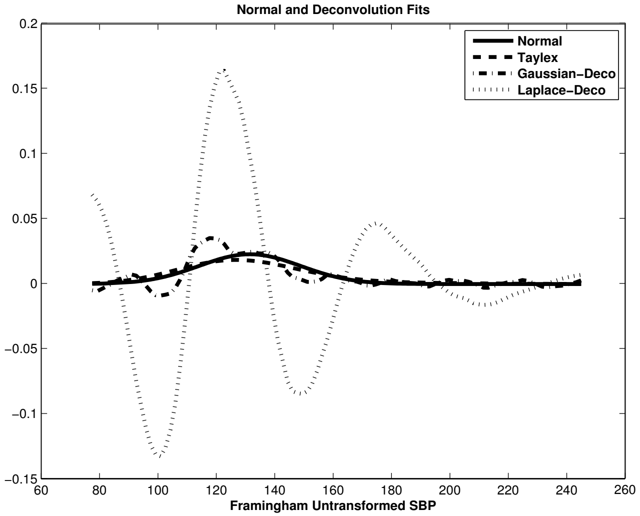


Figure 12.1 *Density estimates of untransformed SBP in Framingham. Four estimates are considered here, but the wildly varying one uses the kernel function with characteristic function $\phi_K(t) = (1-t^2)^3$ when $|t| \le 1$ and zero otherwise, and the default bandwidth $(1/2)\sigma_u n^{-1/9}$. The only real purpose of this figure is to show that automatic bandwidth selection for deconvolution is very hard.*

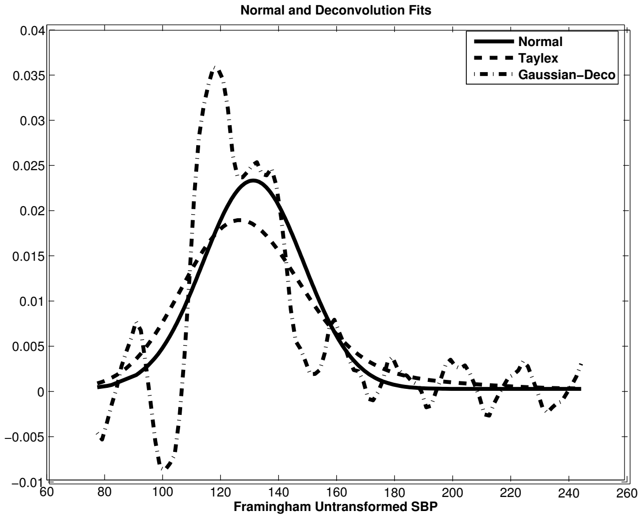


Figure 12.2 *Density estimates of untransformed SBP in Framingham. Three estimates are considered here: the best normal approximation (solid line), the Taylex method (dashed line), and the deconvoluting estimator, which uses the sinc function with a default bandwidth (dot-dashed line).*

12.1.6 Parametric Deconvolution

12.1.6.1 Likelihood Methods

Nonparametric deconvolution is not the only way to estimate the density of $\mathbf{X}$ in an additive model.

If one has a parametric model in mind for $\mathbf{X}$, for example, Weibull, gamma, skew-normal, skew-t, mixtures of normals (Wasserman and Roeder, 1997; Carroll, Maca, and Ruppert, 1999), the SNP (seminonparametric) family (Zhang and Davidian, 2001), etc., then the density/likelihood function for the observed $\mathbf{W}$ is given by the fundamental convolution equation (12.2). Assuming that the integration can be done, for example, numerically for maximum likelihood, via MCMC for Bayes, we can then estimate the unknown parameters and hence obtain an estimate of the density for $\mathbf{X}$.

This simple prescription can be more or less easy, depending on the flexibility of the parametric family involved. After all, the basic fact is that if no assumptions are made, then it is very difficult to assess the density function accurately: This suggests that even flexible parametric methods may have numerical difficulties.

12.1.6.2 Moment Methods

We can also learn something about the first four moments of $\mathbf{X}$ without numerical integration, useful, for example, if one wants to employ the Pearson or Johnson family of densities. Suppose that $\mathbf{W} = \mathbf{X}+\mathbf{U}$, where $\mathbf{U}$ is normally distributed with mean zero and variance σ_u^2. The mean of $\mathbf{W}$ is $\mu_x = E(X)$; the variance of $\mathbf{W}$ is $\sigma_w^2 = \sigma_x^2 + \sigma_u^2$. Let κ_{3x} and κ_{4x} be the skewness and kurtosis of $\mathbf{X}$, being 0 and 3, respectively, if $\mathbf{X}$ is normally distributed. Then the skewness and kurtosis of $\mathbf{W}$ are related to the skewness and kurtosis of $\mathbf{X}$ as follows:

$$\begin{aligned} \kappa_{3w} &= \kappa_{3x}\sigma_x^3/\sigma_w^3; \\ \kappa_{4w} &= (\kappa_{4x}\sigma_x^4 + 6\sigma_x^2\sigma_u^2 + 3\sigma_u^4)/\sigma_w^4, \end{aligned} \tag{12.3}$$

from which the skewness and kurtosis of $\mathbf{X}$ can be extracted.

With replicates, one can push this through even further, making minimal distributional assumptions about $\mathbf{U}$, and then fit a parametric distribution for $\mathbf{X}$ via method of moments. To be specific, suppose that in a sample of size n, one observes replicate observations $\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j}$ ($i = 1, \ldots, n$ and $j = 1, 2$), where it is assumed only that the distribution of the errors is symmetrically distributed about zero, something which often can be achieved by transformation.

Let $\widehat{\mu}_w = \overline{\mathbf{W}}_{\cdot\cdot}$ (the mean), and for $k = 2, 3, 4$ define $s_{w,k}$ to be the sample mean of the terms $\left(\overline{\mathbf{W}}_{i,\cdot} - \widehat{\mu}_w\right)^k$. For $k = 2, 4$ define $s_{u,k}$ to be the sample mean of the terms $\{(\mathbf{U}_{i,1} - \mathbf{U}_{i,2})/2\}^k$. The term $s_{w,k}$ is an estimate of the kth central moment of the $\overline{\mathbf{W}}_{i,\cdot}$'s, while under symmetry $s_{u,k}$ is an estimate of the kth moment of $(\mathbf{U}_{i,1} - \mathbf{U}_{i,2})/2$, which because of symmetry is the same as the kth moment of $(\mathbf{U}_{i1}+\mathbf{U}_{i2})/2 = \overline{W}_{i\cdot} - X_i$.

By equating moments we find the following consistent estimates of the moments of the distribution of $\mathbf{X}$,

$$\begin{aligned} E(\mathbf{X}) = \mu_x &\approx \widehat{\mu}_w; \\ E(\mathbf{X} - \mu_x)^2 &\approx s_{w,2} - s_{u,2}; \\ E(\mathbf{X} - \mu_x)^3 &\approx s_{w,3}; \\ E(\mathbf{X} - \mu_x)^4 &\approx s_{w,4} - s_{u,4} - 6(s_{w,2} - s_{u,2})s_{u,2}. \end{aligned}$$

12.1.6.3 The SNP Family

In the case of additive normally distributed measurement error, the SNP (seminonparametric) distribution has a ready form. The SNP density for $\mathbf{X}$ in the scalar case with $K \le 2$, location μ and scale σ_x is given as

$$f_X(x) = \sigma_x^{-1}\phi\{(x-\mu)/\sigma_x\}G_K\{(x-\mu)/\sigma_x\},$$

where $G_K(x) = (\sum_{j=0}^{K} a_j x^j)^2$. If $K = 0$, this is just the normal density function with mean μ and standard deviation σ_x. For $K > 0$, there are constraints on the a_j that make this a density function. For example, if $K = 1$, we can write $a_0 = \sin(\alpha)$ and $a_1 = \cos(\alpha)$, where $-\pi/2 \le \alpha \le \pi/2$ is a free parameter. Similarly, if $K = 2$, then define $c_1 = \sin(\alpha_1)$, $c_2 = \cos(\alpha_1)\sin(\alpha_2)$ and $c_3 = \cos(\alpha_1)\cos(\alpha_2)$, where (α_1, α_2) are the free parameters such that $-\pi/2 \le \alpha_1, \alpha_2 \le \pi/2$. Let A be the 3×3 matrix with diagonal elements $(1, 1, 3)$, with element $(1, 3)$ and element $(3, 1)$ equal to 1.0, and equal to 0.0 elsewhere. Let B be the symmetric square root of A. Then $(a_0, a_1, a_2)^t = B^{-1}(c_1, c_2, c_3)^t$.

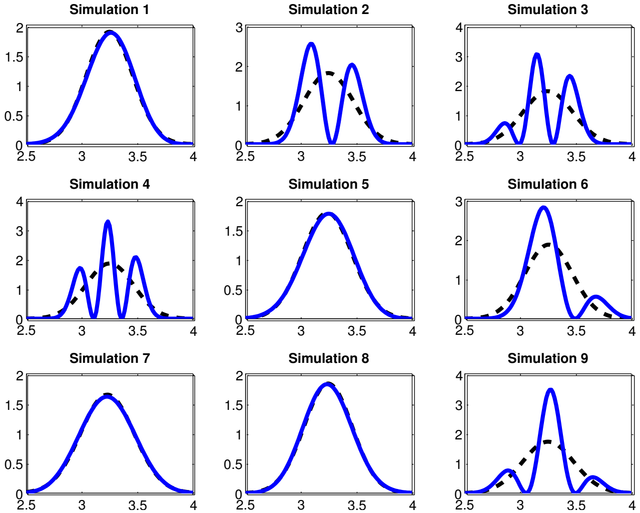


Figure 12.3 *Simulation of the SNP family for parametric deconvolution with* $K = 2$ *and normally distributed* $\mathbf{X}$ *and* $\mathbf{U}$*. The mean and variance of* $\mathbf{X}$ *are* 3.24 *and* 0.052*, respectively, while the variance of* $\mathbf{U}$ *is* 0.171*. The solid line is the SNP fit with* $K = 2$*, while the dashed line is the normal fit. Displayed are nine simulated data sets: The fits should all look normal, but do not.*

Let the measurement error have variance σ_u^2, and make the definitions $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$, $\theta = (\lambda\sigma_u^2/\sigma_x^2)^{1/2}$, $\eta = \lambda^{1/2}(w - \mu)/\sigma_x$, and $\kappa = \lambda^{1/2}\eta$. Then, the density function of the observed $\mathbf{W}$ is given as

$$f_W(w) = (\lambda/\sigma_x^2)^{1/2}\phi(\eta) \int \phi(z)G(\kappa + \theta z)dz, \qquad (12.4)$$

where $\phi(\cdot)$ is the standard normal density function. When $K = 1$, this can be computed exactly as

$$f_W(w) = (\lambda/\sigma_x^2)^{1/2}\phi(\eta) \left\{ (a_0 + a_1\lambda^{1/2}\eta)^2 + \theta^2 a_1^2 \right\}.$$

For $K = 2$,

$$\begin{aligned} f_W(w) \quad = \quad & (\lambda/\sigma_x^2)^{1/2}\phi(\eta)\Big\{ (a_0 + a_1\kappa + a_2\kappa^2)^2 + 3a_2^2\theta^4 \\ & + (\theta a_1 + 2\theta\kappa a_2)^2 + 2a_2\theta^2(a_0 + a_1\kappa + a_2\kappa^2) \Big\}. \end{aligned}$$

For any of $K = 0, 1, 2$, the idea is to use maximum likelihood to estimate the parameters.

Of course, nothing comes for free in deconvolution problems. To see this, we generated nine data sets with $n = 3,145$ observations, mean 3.24, $\mathbf{X}$ normally distributed with variance 0.052, and $\mathbf{U}$ normally distributed with variance 0.171, in line with the NHANES example in Section 12.1.10. This is a lot of measurement error! We fit the SNP distribution with $K = 2$ to do a parametric deconvolution. In this case, remember, $\mathbf{X}$ is normally distributed, but five of the nine fits are very nonnormal, with *one* suggesting a t-density (bottom right, ignore the extra modes) and the others being noticeably multimodal, even though the SNP family with $K = 2$ includes the normal distribution. The point is that one should not overinterpret things like multiple modes when doing deconvolution with a flexible family of distributions.

This example is a little unfair, because SNP fits almost always come equipped with mention of model selection. In Figure 12.4 we have plotted the fits to the same nine simulated data sets as in Figure 12.3, with the exception that we have allowed $K = 0$, the correct model, and $K = 1$ as well as $K = 2$, and let the model be chosen by AIC, which penalizes slightly towards the correct normal model. Here, AIC selected the normal model in eight of the nine data sets.

12.1.7 Estimating Distribution Functions

The pessimistic nature of the results for density estimation with normally distributed error extends to estimating quantiles of the distribution of $\mathbf{X}$, for example, $\text{pr}(\mathbf{X} \le x)$. Here the *optimal* achievable rate of convergence is of the order $\{\log(n)\}^{-3}$, hardly much of an improvement! This casts doubt on the feasibility of estimating quantiles of the distribution of $\mathbf{X}$ without making parametric assumptions.

There are at least two alternatives to a full-blown likelihood analysis. The moment-matching method described previously starts from a model for the density function of $\mathbf{X}$, but makes no assumptions about

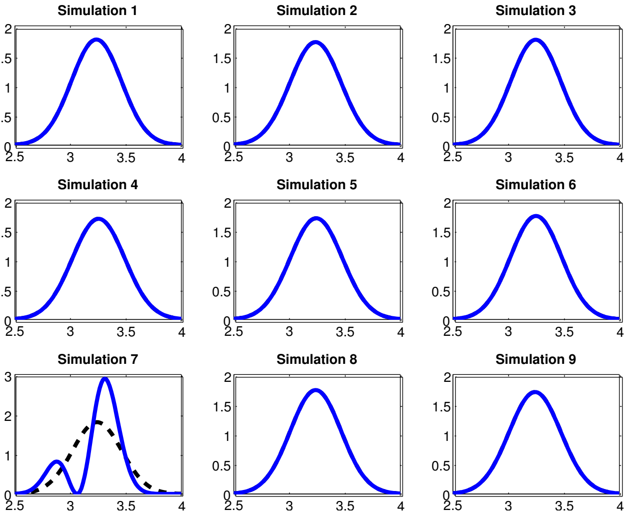


Figure 12.4 *Simulation of the SNP family for parametric deconvolution with K chosen by AIC and normally distributed* $\mathbf{X}$ *and* $\mathbf{U}$*. The mean and variance of* $\mathbf{X}$ *are* 3.24 *and* 0.052*, respectively, while the variance of* $\mathbf{U}$ *is* 0.171*. The solid line is the SNP fit with* $K = 2$*, while the dashed line is the normal fit. In contrast to Figure 12.3, which always used a flexible model, the use of AIC to penalize toward the normal model works reasonably well here.*

the density of $\mathbf{U}$. Its output is an estimated density function that yields estimated quantiles.

Alternatively, with no model for the density of $\mathbf{X}$ but a good model for the error density of $\mathbf{U}$, the SIMEX method can be applied. Previous applications of SIMEX have been to estimated parameters and nonparametric regression estimates, but here the basic input is an empirical distribution function (possibly presmoothed). Stefanski and Bay (1996) studied SIMEX methods for deconvoluting finite-population distribution functions.

12.1.8 Optimal Score Tests

While estimating a density function nonparametrically is difficult in the presence of measurement error, estimating smooth functionals of the unknown density, for example, $m(w) = E(\mathbf{X}|\mathbf{W} = w)$, is often not as difficult.

For estimating $m(w)$, we can simply replace f_X and f_W in (12.1) by their estimators. Stefanski and Carroll (1991) showed that this substitution works, in the sense that the resulting estimate of $m(w)$ when substituted into the score test typically achieves the same local power as if $m(w)$ were a known function.

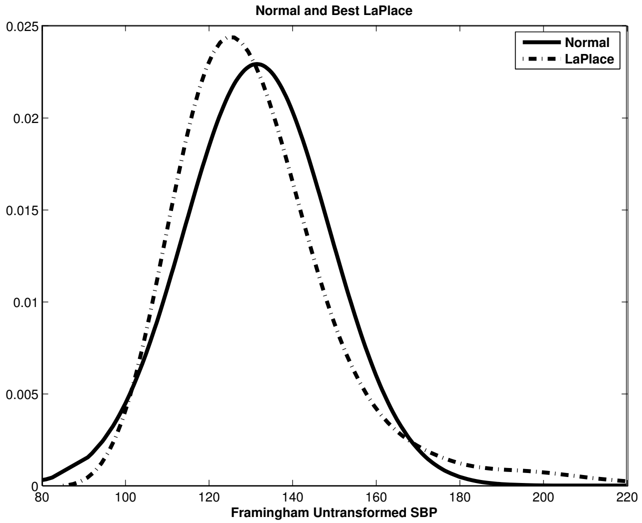


Figure 12.5 *Density estimates of untransformed SBP in Framingham. Two estimates are considered here: the best normal approximation (solid line), the Taylex method (dashed line), with bandwidth chosen by eye to be as small as possible but still be smooth.*

The reason for this is that $m(w)$ is much easier to estimate than f_X, because of the extra integration in (12.1). In fact, with normally distributed measurement errors, the rate of convergence for estimating $m(w)$ is of order $n^{-4/7}$, while for Laplace error the rate is the usual nonparametric one, that is, $n^{-4/5}$ (Stefanski and Carroll, 1991).

12.1.9 Framingham Data

We applied deconvoluting kernel density estimation techniques to the Framingham data, for untransformed systolic blood pressure (SBP), rather than using the transformation $\log(SBP - 50)$. We used SBP at Exam #2 only to estimate the measurement error variance, but deconvolved SBP measured at Exam #3 ($\mathbf{W}$). In the original scale, observed SBP had mean 130.01, variance 395.65, and the estimated measurement error variance was 83.69. This leads to an estimate of the variance for long-term SBP ($\mathbf{X}$) of 311.96.

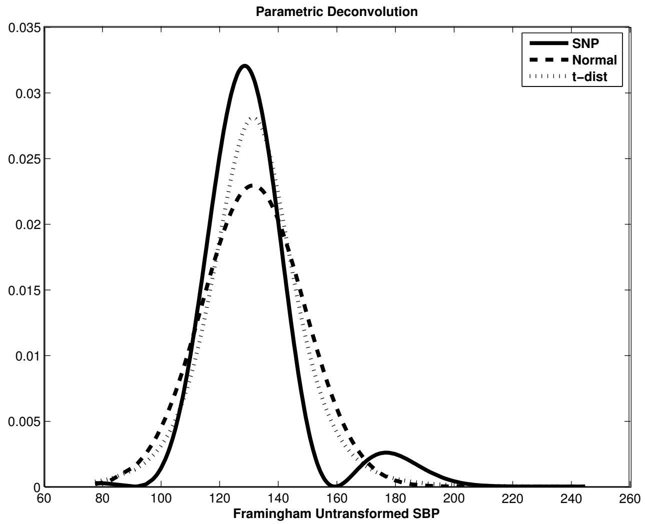


Figure 12.6 *Density estimates of untransformed SBP in Framingham. Three estimates are considered here: the best normal approximation (dashed line), the SNP fit with $K = 2$ (solid line) and the best fitting t-density (dotted line).*

Figure 12.5 shows the best-fitting normal distribution, along with the Taylex deconvoluting kernel fit (Carroll and Hall, 2004), where the bandwidth was chosen by eye to be as small as possible while retaining smoothness. The only real point of interest here is that the deconvoluting kernel fit picks up the skewness inherent in the untransformed SBP data, thus correctly suggesting that the data should be transformed. The parametric deconvolution fit using the SNP distribution is given in Figure 12.6, along with the best-fitting t-density and the best-fitting normal density. Here, the skewness we know to exist is exhibited by the lonely little mode over on the right.

12.1.10 NHANES Data

The NHANES data (Chapter 4) exhibit considerably more measurement error, and consequently deconvolution is much harder. For these data we have earlier derived the variances $\widehat{\sigma}^2_w = 0.223$, $\widehat{\sigma}^2_u = 0.171$ and $\widehat{\sigma}^2_x = 0.052$.

We used the same methods as for the Framingham data. Figure 12.7 gives the best-fitting normal density, along with the deconvolution density estimates for normal and Laplace errors, with bandwidths selected to

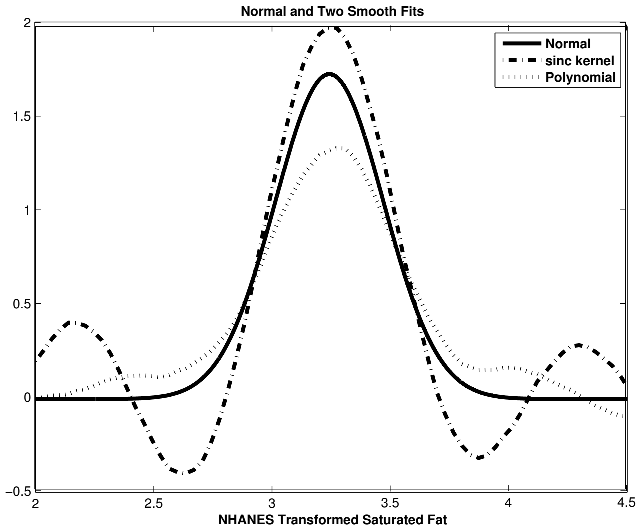


Figure 12.7 *Density estimates of transformed saturated fat in NHANES. Three estimates are considered here: the best normal approximation (solid line), Gaussian deconvolution (dashed line), and Laplace deconvolution (dotted line), both with bandwidth selected by eye. There is a clear suggestion of symmetry in the data, but not much else.*

be smooth but small. The sample skewness of $\mathbf{W}$ is nearly zero (-0.05), and this is reflected in the near symmetry of the plots.

Figure 12.8 gives the best fitting normal, the best-fitting t-density, and the SNP density with $k = 2$ as parametric deconvolution methods. There is a suggestion on the latter two that the data are somewhat like a t-density, here with about 5 degrees of freedom. The sample kurtosis is 3.32, where a kurtosis of 3 applies for the normal distribution. If the kurtosis of $\mathbf{X}$ is denoted by κ_{4x}, then in the additive error model with normally distributed errors the kurtosis for $\mathbf{W}$ is given by (12.3). Substituting sample estimates of $(\kappa_{4w}, \sigma^2_x, \sigma^2_u, \sigma^2_w)$ and solving for κ_{4x}, the kurtosis for $\mathbf{X}$ is estimated to be approximately 9.0, indicating about 5.0 degrees of freedom since the kurtosis of a t-density is $3 + 6/(r - 4)$, where r is the degrees of freedom.

12.1.11 Bayesian Density Estimation by Normal Mixtures

If the reader is unfamiliar with Bayesian estimation, then Chapter 9 should be read before this section.

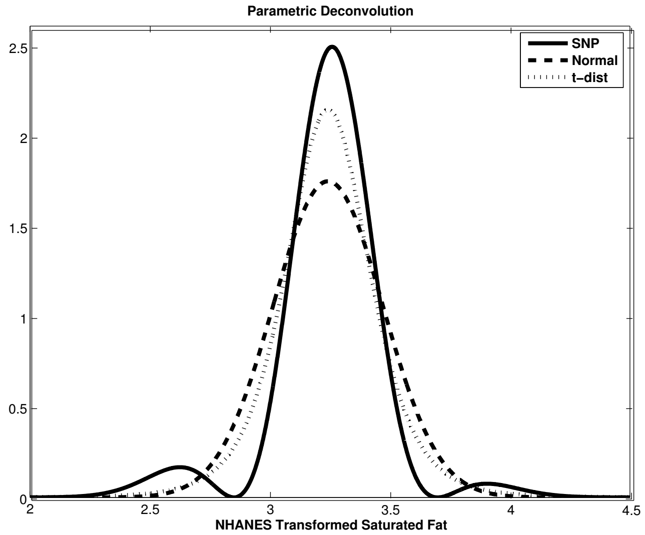


Figure 12.8 *Density estimates of transformed saturated fat in NHANES. Three estimates are considered here: the best normal approximation (dashed line), the SNP fit with $K = 2$ (solid line), and the best fitting t-density (dotted line).*

Wasserman and Roeder (1997) proposed a Bayesian method for nonparametric density estimation. They assume that the density f_X of $\mathbf{X}_1$, $\ldots, \mathbf{X}_n$ is a normal mixture

$$f_X(x) = \sum_{i=1}^{k} p_i \phi\{(x-\mu_i)/\sigma_i\}/\sigma_i, \ p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1, \ k \leq L,$$

where L is a known upper bound for the number of components and ϕ is the Normal$(0,1)$ density. Since any smooth density can be closely approximated by a normal mixture, this method is nonparametric, that is, it is appropriate even if f_X is not exactly a normal mixture. Wasserman and Roeder described how k, $p_1, \ldots, p_k$, $\mu_1, \ldots, \mu_k$, and $\sigma_1, \ldots, \sigma_k$ can be estimated by a Gibbs sampler. Carroll, Maca, and Ruppert (1999) showed that this Gibbs sampler can be applied when $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are observed with measurement error, that is, one observes $\mathbf{W}_1, \ldots, \mathbf{W}_n$ where $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$ and $\mathbf{U}_i$ is Normal$(0, \sigma_u^2)$ for a known σ_u; of course, in practice one substitutes an estimate for σ_u. The only modification needed to accommodate the measurement error is a simple idea used repeatedly in Chapter 9: During the MCMC the unobserved $\mathbf{X}_i$ are sampled from their full conditionals. Given these imputed values, the other steps of the Gibbs sampler are exactly the same as when the $\mathbf{X}_i$ are observed.

12.2 Nonparametric Regression

Nonparametric regression has become a rapidly developing field as researchers have realized that parametric regression is not suitable for adequately fitting curves to all data sets that arise in practice.

Nonparametric regression entails estimating the mean of $\mathbf{Y}$ as a function of $\mathbf{X}$,

$$E(\mathbf{Y}|\mathbf{X} = x_0) = m_{\mathbf{Y}}(x_0), \tag{12.5}$$

without the imposition of $m_{\mathbf{Y}}$ belonging to a parametric family of functions. We focus on the local-polynomial, kernel-weighted regression and spline estimators of $m_{\mathbf{Y}}$.

The most promising approach we know of for nonparametric regression with measurement error is a Bayesian methodology using splines and MCMC introduced by Berry, Carroll, and Ruppert (2002) and its frequentist counterpart, which was introduced by Ganguli, Staudenmayer, and Wand (2005). The algorithm is given also in Ruppert, Wand and Carroll (2003, Chapter 15.3) in a somewhat easier-to-digest form. We will discuss this methodology in detail in Chapter 13. In the present chapter, earlier and simpler estimators will be discussed.

12.2.1 Local-Polynomial, Kernel-Weighted Regression

When $\mathbf{X}$ is observable, the local, order-p polynomial estimator is $\widehat{\beta}_0(x)$, the solution for β_0 to the weighted least squares problem minimizing,

$$\textstyle\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{p} \beta_j (\mathbf{X}_i - x)^j \right\}^2 K_h(\mathbf{X}_i - x). \tag{12.6}$$

Here h is called the *bandwidth*, K is a kernel function such that $\int K(u)\, du = 1$, and $K_h(u) = h^{-1}K(u/h)$. The function $K(\cdot)$ and the bandwidth h are under the control of the investigator, and in practice the latter is the more important.

Problem (12.6) is a straightforward weighted least squares problem, and hence is easily solved numerically. The local least squares estimator of $m_{\mathbf{Y}}(x)$ is then

$$\widehat{m}_{\mathbf{Y}}(x, h) = \widehat{\beta}_0(x), \tag{12.7}$$

while for $j \leq p$, the estimator of the jth derivative of $m_{\mathbf{Y}}(x)$ is $j!\widehat{\beta}_j(x)$. Estimator (12.7) has had long use in time series analysis, and is a special case of the robust, local regression estimators in Cleveland (1979).

Cleveland and Devlin (1988) discussed practical implementation and presented several interesting case studies where local regression data analysis is considerably more insightful than classic linear regression analysis. Ruppert and Wand (1994) described the asymptotic theory of these estimators. R-code is available, see http://web.maths.unsw.edu.au/∼wand.

As in parametric problems, ignoring measurement error causes inconsistent estimation of $m_{\mathbf{Y}}(x)$. The regression calibration and SIMEX methods of Chapters 4 and 5 provide simple means for constructing approximately consistent estimators of $m_{\mathbf{Y}}(x)$ in the case that $\mathbf{W} = \mathbf{X}+\mathbf{U}$, where $\mathbf{U}$ has mean zero. Hastie and Stuetzle (1989) describe an alternative method for an orthogonal regression problem, wherein it is assumed that the conditional variances of $\mathbf{Y}$ and $\mathbf{W}$ given $\mathbf{X}$ are equal; we have already commented (Section 3.4.2) on the general lack of applicability of such an assumption.

In this section, we describe algorithms for nonparametric regression taking measurement error into account.

12.2.2 Splines

Low-degree polynomials are effective for approximating a smooth function in small regions. However, if we wish to approximate a smooth function over a large region, polynomial approximation typically does not work well. One can increase the polynomial order to gain degrees of freedom, but higher-order polynomials can be highly oscillatory and changing the coefficients to increase the accuracy of the approximation in one location changes the polynomial globally and may decrease accuracy elsewhere. The solution to this problem is to piece together a number of low-degree polynomials. A p^{th} degree spline with knots $\kappa_1, \ldots, \kappa_K$ is a piecewise polynomial function s with polynomial form changing at each knot in such a way that the j^{th} derivative of s is continuous everywhere for $j \le p-1$. In practice, generally p is 1, 2, or 3.

There are many ways to parameterize a spline. Usually, one works with a spline basis. Given a fixed degree and knots, one can find $1+p+K$ basis functions, collectively called a *basis*, such that any spline with this degree and knots is a linear combination of these basis functions. There are many, in fact, infinitely many, bases, and we can work with whichever one we like. One basis that is easy to understand is the truncated power function basis. The p^{th} degree truncated power function with knot κ_i is $(x-\kappa_i)_+^p$ which is defined to be zero if $x \le \kappa_i$ and $(x-\kappa_i)^p$ if $x > \kappa_i$. The truncated power basis of degree p and knots $\kappa_1 < \ldots < \kappa_K$ is $\{1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p\}$, and an arbitrary spline with this degree and knots can be written as

$$s(x) = \sum_{k=0}^{p} \beta_k x^k + \sum_{k=1}^{K} b_k (x-\kappa_k)_+^p. \tag{12.8}$$

If a spline is fit by ordinary least squares, then the number and locations of the knots have a tremendous effect on the fit, and there is a large literature on the so-called *adaptive splines* where the knots are chosen to provide the most accurate estimate; see Stone, Hansen, Kooperberg, et al. (1997) and Hansen and Kooperberg (2002). An alternative to adaptive knot selection is to use a large number of knots and to place a penalty on the "roughness" of the fit. For example, cubic smoothing splines use a knot at every unique value of x and penalize the integral of the squared second derivative of the fit. This penalty reduces the curvature of the fit.

Ruppert, Wand, and Carroll (2003) introduce a simple penalty that is convenient for our purposes. They fit $s(x)$ in (12.8) to data $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$ by minimizing

$$\sum_{i=1}^{n} \left[\mathbf{Y}_i - \left\{ \sum_{k=0}^{p} \beta_k \mathbf{X}_i^k + \sum_{i=k}^{K} b_i (\mathbf{X}_i - \kappa_k)_+^p \right\} \right]^2 + \lambda \sum_{k=1}^{K} b_k^2. \tag{12.9}$$

The knots can be equally spaced or spaced so that there are roughly an equal number of the unique values of the $\mathbf{X}_i$ between each pair of adjacent knots. The number of knots is generally between 5 and 20. The exact number of knots has little effect on the fit, because the penalty prevents overfitting (Ruppert, 2002). The spline minimizing (12.9) is called a *penalized spline*. The smoothing parameter λ has a crucial influence on the fit and must be chosen appropriately. Data-based methods for selecting λ include generalized crossvalidation, REML, and Bayesian estimation; see Ruppert, Wand, and Carroll (2003).

12.2.3 QVF and Likelihood Models

Local polynomial nonparametric regression is easily extended to likelihood and quasilikelihood and variance function (QVF) models. The reason is that local polynomial regression can be looked at in two ways that permit immediate generalization. First, as seen in (12.6), local polynomial regression estimation of $m_{\mathbf{Y}}(x_0)$ at a value x_0 is equivalent to a weighted maximum likelihood estimate of the intercept in the model, assuming that $\mathbf{Y}$ is normally distributed with mean $\beta_0 + \beta_1(\mathbf{X} - x_0)$, constant variance, and with the weights $K_h(\mathbf{X} - x_0)$. Thus, in other generalized linear models (logistic, Poisson, gamma, etc.), the suggestion is to perform a weighted likelihood analysis with a mean of the form $h\{\beta_0 + \beta_1(\mathbf{X} - x_0)\}$ for some function $h(\cdot)$.

Extending local polynomial nonparametric regression to QVF models is also routine. As seen in (12.7), local linear regression is a weighted QVF estimate based on a model with mean $\beta_0+\beta_1(\mathbf{X}-x_0)$ and constant variance, and with extra weighting $K_h(\mathbf{X}-x_0)$. The suggestion in general problems is to do the QVF analysis with argument $\beta_0 + \beta_1(\mathbf{X} - x_0)$ and extra weighting $K_h(\mathbf{X} - x_0)$.

12.2.4 SIMEX for Nonparametric Regression

Use of SIMEX in nonparametric regression follows the same ideas as in parametric problems. We require an additive error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $\mathbf{X}$ with variance σ_u^2. Sometimes, a transformation of the original surrogate is required to achieve additivity and homoscedasticity. The SIMEX algorithm for nonparametric regression is as follows:

(a) Fix values for $\lambda \in \Lambda = (0 < \lambda_1 < \ldots < \lambda_M)$.
(b) For $b = 1, \ldots, B$, let ϵ_{ib} be the non-iid pseudoerrors.
(c) Define $W_{ib}(\lambda) = \mathbf{W}_i + \sigma_u\lambda^{1/2}\epsilon_{ib}$.
(d) For $b = 1, \ldots, B$ and $\lambda \in \Lambda$, compute the nonparametric regression estimate (12.7) by regressing $\mathbf{Y}_i$ on $\mathbf{W}_{ib}(\lambda)$. Call the resulting estimate $\widehat{f}(x, b, \lambda)$.
(e) Let $\widehat{f}(x, \lambda)$ be the sample mean of the terms $\widehat{f}(x, b, \lambda)$.
(f) For each x, extrapolate the values $\widehat{f}(x, \lambda)$ as a function of λ back to $\lambda = -1$, resulting in the SIMEX estimator $\widehat{f}(x)$.

12.2.4.1 SIMEX Applied to Local Polynomial Regression

An interesting problem is how best to choose the smoothing parameter. The smoothing parameter determines how one trades off smoothing bias and variance; smaller bandwidths give more variance but less smoothing bias. For the case of local polynomial regression, Carroll, Maca, and Ruppert (1999) applied Ruppert's (1997) EBBS (empirical bias bandwidth selector) method to the naive estimates in step (d), but doing this resulted in final SIMEX estimators that were undersmoothed. This problem was addressed by Staudenmayer and Ruppert (2004), who noticed that the undersmoothing occurred because the SIMEX extrapolant used to estimate $f(x)$ is much more variable than the naive estimators of $f(x)$ in step (d) to which EBBS was being applied. They developed an asymptotic theory for the SIMEX extrapolant so that EBBS could be applied to the SIMEX estimator itself, rather than to the naive estimators fed into SIMEX. With the Staudenmayer and Ruppert bandwidth, the SIMEX estimator is smoother and has smaller mean squared error than SIMEX with the "naive" bandwidth used by Carroll, Maca, and Ruppert (1999).

12.2.4.2 SIMEX Applied to Penalized Splines

Carroll, Maca, and Ruppert (1999) also applied SIMEX to penalized splines and found that the SIMEX/splines estimator performed quite similarly to SIMEX/local polynomial estimator and was inferior to their structural spline estimator, which will be discussed in Section 12.2.6.

12.2.5 Regression Calibration

In 1995, regression calibration made some sense for nonparametric regression, since there was little in the way of a literature. Now there is more, and we do not think regression calibration should be used in this context. At best, in its expanded form, it will be able to capture quadratic functions, but fitting quadratics is not the intent of the field.

12.2.6 Structural Splines

A more sophisticated application of the regression calibration idea was proposed by Carroll, Maca, and Ruppert (1999). The spline regression model

$$\mathbf{Y}_i = \sum_{k=0}^{p} \beta_k \mathbf{X}_i^k + \sum_{i=k}^{K} b_i(\mathbf{X}_i - \kappa_k)_+^p + \epsilon_i, \qquad (12.10)$$

implies that

$$E(\mathbf{Y}_i|\mathbf{W}_i) = \sum_{k=0}^{p} \beta_k E(\mathbf{X}_i^k|\mathbf{W}_i) + \sum_{i=k}^{K} b_i E\{(\mathbf{X}_i - \kappa_k)_+^p|\mathbf{W}_i\}. \qquad (12.11)$$

If we can estimate each of the conditional expectations in (12.11), then we can fit this equation to the $\mathbf{Y}_i$ to estimate the parameters in (12.10). Fortunately, estimation of these conditional expectations is a straightforward application of Bayesian density estimation methodology of Section 12.1.11, as will be explained in the next paragraph. Fitting (12.11) is an example of regression calibration if we think of $\mathbf{X}_i, \ldots, \mathbf{X}_i^p, (\mathbf{X}_i - \kappa_1)_+^p, \ldots, (\mathbf{X}_i - \kappa_K)_+^p$ as a $p + K$ dimensional set of covariates measured with error.

When the MCMC algorithm in Section 12.1.11 is run, each $\mathbf{X}_i$ is imputed from its conditional distribution given $\mathbf{W}_i$ and the parameters. Thus, if for any function g we average $g(\mathbf{X}_i)$ over the imputed values of $\mathbf{X}_i$ from an MCMC sample, then we estimate $E\{g(\mathbf{X}_i)|\mathbf{W}_i\}$. If this is done for $g(x)$ equal to each of $x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_K)_+^p$ and for $i = 1, \ldots, n$, then these quantities can be used as the covariates to fit (12.11).

Since the methodology of Section 12.1.11 is based on a flexible normal mixture model for the density of $\mathbf{X}$, the algorithm described in the present section was called the "structural spline" method by Carroll, Maca, and Ruppert (1999). In their simulation experiment, the structural spline estimates had substantially smaller mean squared errors than SIMEX applied to either local polynomial estimation or penalized splines. However, Berry, Carroll, and Ruppert (2002) found that a fully Bayesian model outperformed the structural spline estimator. Their fully Bayesian approach is described in Chapter 13.

12.2.7 Taylex and Other Methods

12.2.7.1 Globally Consistent Methods via Deconvolution

A globally consistent deconvoluting kernel regression function estimate can be obtained by replacing the kernel in (12.6) with a deconvoluting kernel (Fan and Truong, 1993), resulting in what we refer to as a deconvoluting kernel, local regression estimator.

However, the bandwidth selection problem associated with this approach is by no means trivial, and the rates of convergence for the resulting estimators are the same as for the density estimation problem. Moreover, in a simulation study of Carroll, Maca, and Ruppert (1999), the deconvoluting kernel estimate was applied with the ideal "oracle" bandwidth that minimized the mean squared error, and even then it was inferior to SIMEX and structural splines.

All the deconvolution methods described to this point require that the distribution of the measurement error distribution be known, except up to a scale parameter such as a variance. Schennach (2004) developed a method that allows the measurement error distribution to be completely unknown, as long as there are replicates of $\mathbf{W}$, that is, $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$ for $j = 1, 2$. One can easily see how this is can be in the most important special case of additive symmetric errors $\mathbf{U}_{ij}$, because there is a special trick that applies. In this case, $\overline{\mathbf{W}}_{i\cdot} = \mathbf{X}_i + \overline{\mathbf{U}}_{i\cdot} = \mathbf{X}_i + (\mathbf{U}_{i1} + \mathbf{U}_{i2})/2$, while $(\mathbf{W}_{i1} - \mathbf{W}_{i2})/2 = (\mathbf{U}_{i1} - \mathbf{U}_{i2})/2$. Here is the trick: Because of symmetry of the errors, the distributions $\mathbf{U}_{i2}$ and $-\mathbf{U}_{i2}$ are the same, and thus the distributions of of $(\mathbf{U}_{i1} + \mathbf{U}_{i2})/2$ and $(\mathbf{U}_{i1} - \mathbf{U}_{i2})/2$ are the same. This means that $(\mathbf{W}_{i1} - \mathbf{W}_{i2})/2$ can be used to estimate the density function of $\overline{\mathbf{U}}_{i\cdot}$. Of course, the slow rates of convergence for deconvolution methods do not get any better once one has to estimate the error distribution.

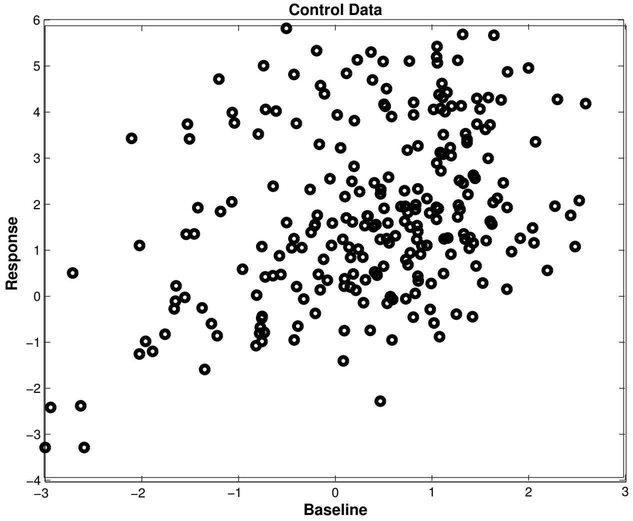


Figure 12.9 *The control data in the baseline change example.*

12.2.7.2 Taylex

Carroll and Hall's (2004) Taylex method (for Taylor Series Expansion) is in the class of deconvolution estimators, with the deconvoluting kernel function $K_*(t,h) = \phi(x)\{1 - \sigma_u^2(x^2-1)/(2h^2)\}$. However, and crucially, they do not advertise the method as *globally* consistent, but merely approximately consistent for relatively small measurement error. They used the following bandwidth selection algorithm. They first fitted local linear regression with local bandwidths computed Ruppert's (1997) EBBS methodology, using the Taylex kernel with $\sigma_u^2 = 0$. Then the mean of the EBBS bandwidths was computed, and the bandwidth actually used was the average EBBS bandwidth multiplied by 0.75. Calculations by Staudenmayer and Ruppert (2004) show that there is a bias term of order $O(\sigma_u^4)$ in the regression estimate, and holding this fixed suggests that the bandwidth should be of order smaller than the usual $h^{-1/5}$. The 0.75 correction is an ad hoc means of accomplishing this.

12.3 Baseline Change Example

We analyze data originally analyzed by Berry, Carroll, and Ruppert (2002). Unfortunately, we do not have permission to discuss the study details here, or to make the data publicly available. The data have been transformed and rescaled and have had random noise added.

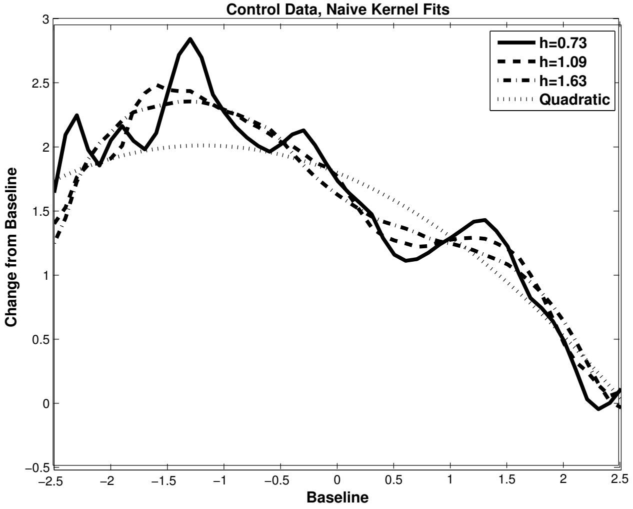


Figure 12.10 *Local quadratic kernel regression fits to the baseline change controls data, along with a quadratic fit. All methods ignore measurement error.*

Essentially, there is a treatment group and a control group, which are evaluated using a scale at baseline ($\mathbf{W}$) and at the end of the study ($\mathbf{Y}$). Smaller values of both indicate more severe disease. The scale itself is subject to considerable error because it is based on a combination of self-report and clinic interview. The study investigators estimate that in their transformed and rescaled form, the measurement error variance is approximately $\sigma_u^2 = 0.35$.

A preliminary Wilcoxon test applied to the observed change from baseline, $Y - W$, indicated a highly statistically significant difference between the two groups.

The main interest focuses on the population mean change from baseline $\Delta(X) = m_{\mathbf{Y}}(X) - X$ for the two groups, and most importantly the difference between these two functions. Here we describe results only for the control group. The data are given in Figure 12.9.

Preliminary nonparametric regression analysis of the data ignoring measurement error indicates possible nonlinearity in the data: A quadratic regression is marginally statistically significant ($p \approx 0.03$). When we corrected the quadratic fits for measurement error (Cheng and Schneeweiss, 1998) and bootstrapped the resulting parameter estimates, both p-values exceeded 0.20, although the fitted functions had substantial curvature. In Figure 12.10, we plot local quadratic kernel estimates with three bandwidths, along with a quadratic regression, all ignoring measurement error. As is typical of kernel methods, the smallest bandwidth has the least smooth character.

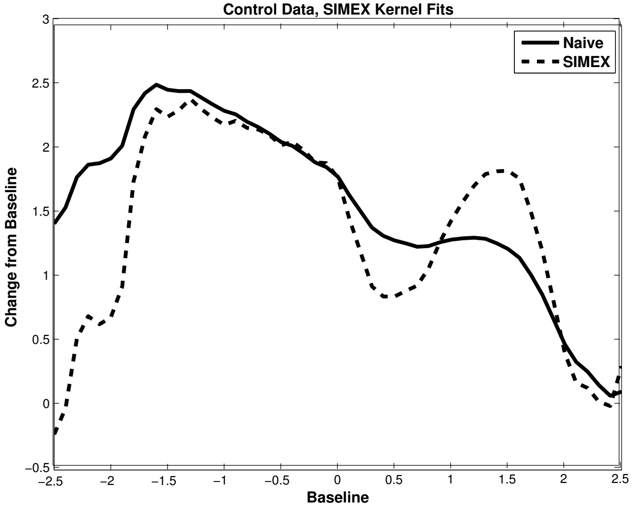


Figure 12.11 *A naive local quadratic fit (solid line) and a SIMEX local quadratic fit (dashed line) for the baseline change controls data.*

We then applied SIMEX to both the local polynomial fits and to the quadratic fit. In Figure 12.11, we used the middle of the three bandwidths and display the naive and SIMEX fits. One can see that the "bumps," or less colloquially the local features, in the naive fit are exaggerated somewhat in the SIMEX fit. Figure 12.12 gives the SIMEX local quadratic fit and the SIMEX global quadratic fit.

12.3.1 Discussion of the Baseline Change Controls Data

Figure 12.12 is difficult to interpret without the scientific context, which we are not at liberty to discuss. However, we can note that the people in the study are controls, that is, not given a treatment. The higher the change from baseline, the more the patient has improved by the end of the study. Since these are untreated patients, what this figure shows is a placebo effect, sometimes a rather strong one. Both the local quadratic SIMEX fit and the global quadratic SIMEX fit suggest that the placebo effect is confined away from those doing best or worst at baseline. In the context of the actual study, this is not so far-fetched. In contrast, the naive fits (Figure 12.10) suggest that those who are doing the worst at baseline (lowest values) have a strong placebo effect.

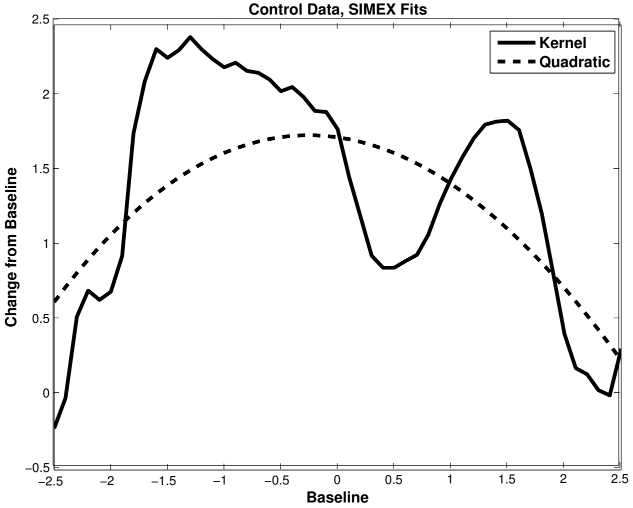


Figure 12.12 *A SIMEX local quadratic fit (solid line) and SIMEX global quadratic fit (dashed line) to the baseline change controls data.*

Bibliographic Notes

The early work of Stefanski and Carroll (1986, 1990c), Carroll and Hall (1988) and Liu and Taylor (1989) has since spawned a considerable literature on deconvolution; see for example Liu and Taylor (1990), Zhang (1990), Fan (1991a,b,c; 1992a), Fan, et al. (1991), Masry and Rice (1992), Fan and Truong (1993), Fan and Masry (1993), and Stefanski (1989, 1990). An interesting econometric application using a modification of these methods is discussed by Horowitz and Markatou (1993). There continues to be interest in the problem of deconvolution (Taupin, 2001; Delaigle and Gijbels, 2004ab, 2005, 2006).

There have been a number of monographs nonparametric regression (for example, Müller, 1988; Härdle, 1990; Hastie and Tibshirani, 1990; Fan and Gijbels, 1996; Bowman and Azzalini, 1997; and Ruppert, Wand, and Carroll, 2003), where it is shown that nonparametric regression techniques have much to offer in applications.

CHAPTER 13

SEMIPARAMETRIC REGRESSION

13.1 Overview

Semiparametric models combine parametric and nonparametric submodels. For example, in a semiparametric regression model the effects of some, but not all, covariates are modeled nonparametrically. In this chapter we will take a Bayesian viewpoint, mostly for the pragmatic reason that we have found that a Bayesian analysis is easiest to implement.

This chapter considers three main topics. In the first, primary topic, the effect of **X** on the response is modeled nonparametrically and the effect of **Z** is modeled parametrically. The second topic considered is the opposite: the effect of **X** on the response is modeled parametrically and the effect of **Z** is modeled nonparametrically. Finally, the third topic considers the case that both **X** and **Z** are modeled parametrically by a nonlinear form, but the distribution of **X** is treated nonparametrically.

13.2 Additive Models

For concreteness, we will focus on the additive model where $\mathbf{Z}_i$ is a vector of observed covariates with linear effects and $\mathbf{X}_i$ is an unobserved scalar covariate with an effect of unknown form. Thus, we will use the model

$$\mathbf{Y}_i = s(\mathbf{X}_i) + \beta_z^t \mathbf{Z}_i + \epsilon_i, \tag{13.1}$$

where s is smooth, but otherwise unknown. Model (13.1) is *additive* because the effects of $\mathbf{X}_i$ and the components of $\mathbf{Z}_i$ are added together; there are no interaction terms that would be functions of two or more covariates. We will use a spline model given by (12.8) and repeated here as

$$s(x) = \sum_{k=0}^{p} \beta_k x^k + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^p. \tag{13.2}$$

As discussed in Section 12.2.2, $\kappa_1 < \ldots < \kappa_K$ are knots and $(x - \kappa_i)_+^p$ is defined to be zero if $x \leq \kappa_i$ and $(x - \kappa_i)^p$ if $x > \kappa_i$. The degree p is typically 1, 2, or 3 and there are between five and 20 knots. The knots are usually at "equally spaced" quantiles of the $\mathbf{W}_i$, for example, at the deciles if $K = 9$.

Model (13.1) is one of many spline models discussed in detail in Ruppert, Wand, and Carroll (2003), which deals mostly with covariates without error but has a chapter on measurement error. Readers not familiar with semiparametric regression may wish to consult that reference. Once the analysis of (13.1) is understood, it can be extended to many other models, for example, ones where the effects of some components of $\mathbf{Z}_i$ are also modeled via splines or where $\mathbf{X}_i$ has several components.

13.3 MCMC for Additive Spline Models

As seen in Chapter 9, sampling from the posterior for a regression model with covariate measurement error is nearly identical to sampling from the same model without covariate measurement error. The main difference is that there is an extra step where the unknown values of the $\mathbf{X}_i$ for nonvalidation data are imputed.

By combining (13.1) and (13.2) we arrive at the model

$$\mathbf{Y}_i = \sum_{k=0}^{p} \beta_k \mathbf{X}_i^k + \sum_{k=1}^{K} b_k (\mathbf{X}_i - \kappa_k)_+^p + \beta_z^t \mathbf{Z}_i + \epsilon_i. \tag{13.3}$$

Clearly, (13.3) is a linear model, albeit a somewhat complicated one. In Section 9.4, we used nearly flat, that is, noninformative priors, for the regression coefficients of a linear model. This strategy will not work well for model (13.3), because of the large number of coefficients. What is needed is a method for shrinking the spline coefficients $b_1, \ldots, b_K$ toward zero to prevent overfitting. This shrinkage can be accomplished by using a more informative prior for $b_1, \ldots, b_K$ while continuing to use a flat prior for the other coefficients. To separate the two types of regression coefficients, we will rewrite (13.3) as

$$\mathbf{Y}_i = \beta_{zx}^t m_1(\mathbf{X}_i, \mathbf{Z}_i) + b_x^t m_2(\mathbf{X}_i) + \epsilon_i, \tag{13.4}$$

where

$$\begin{aligned} m_1(\mathbf{X}_i, \mathbf{Z}_i) &= (1, \mathbf{X}_i, \ldots, \mathbf{X}_i^p, \mathbf{Z}_i^t), \\ m_2(\mathbf{X}_i) &= \{(\mathbf{X}_i - \kappa_1)_+^p, \ldots, (\mathbf{X}_i - \kappa_K)_+^p\}, \\ \beta_{zx} &= (\beta_0, \ldots, \beta_p, \beta_z^t)^t, \\ \text{and } b_x &= (b_1, \ldots, b_K)^t. \end{aligned} \tag{13.5}$$

We will use the prior

$$[\beta_{zx}] = \text{Normal}(0, \sigma_\beta^2 \boldsymbol{I}) \tag{13.6}$$

where σ_β^2 is "large," say 10^6. The prior for b_x is hierarchical: given a prior variance σ_b^2, b_x is Normal$(0, \sigma_b^2 \boldsymbol{I})$, and σ_b^2 is IG$(\delta_{b,1}, \delta_{b,2})$, where IG stands for the inverse-Gamma distribution defined in Section A.3. Using the $[\,\cdot\,]$ notation described in the Guide to Notation, we can write this prior as

$$[b_z|\sigma_b^2] = \text{Normal}(0, \sigma_b^2 \boldsymbol{I}), \text{ and } [\sigma_b^2] = \text{IG}(\delta_{b,1}, \delta_{b,2}). \tag{13.7}$$

As discussed in Section 9.4, for the prior to be noninformative the values of $\delta_{b,1}$ and $\delta_{b,2}$ should be small. For $\delta_{b,1}$, "small" means small relative to the sample size. Since $\delta_{b,2}$ is a prior guess about the variance of the b_k, and since spline coefficients are often quite small, it is crucial that $\delta_{b,2}$ be sufficiently small. Put differently, $\delta_{b,2}$ should be small relative to typical values of the b_k^2. Depending on the application, $\delta_{b,2} = 10^{-8}$ or even smaller may be necessary to prevent overfitting, which in this context is the same as undersmoothing. The reason for this is that the spline coefficients are typically very small, though how small depends upon the number of knots and the scaling of $\mathbf{X}$ (Crainiceanu, Ruppert, and Wand, 2005).

Besides the prior on the regression coefficients, we need a prior on σ_ϵ^2, the variance of $\epsilon_1, \ldots, \epsilon_n$. We also need an error model for $\mathbf{W}|\mathbf{X}$ or a Berkson model for $\mathbf{X}|\mathbf{W}$, and, in the case of an error model, a structural "exposure" model for $\mathbf{X}|\mathbf{W}$. Of course, we need priors on these models as well. Each of these models and priors is chosen in the same manner as in Chapter 9.

The full conditionals are the same as in Section 9.4 for error models or Section 9.7 for Berkson models, except for σ_b^2 which did not appear in models studied in Chapter 9. It is not difficult to show that the full conditional for σ_b^2 is

$$f(\sigma_b^2|\text{others}) = \text{IG}\{\delta_{b,1} + K/2,\ \delta_{b,2} + (1/2)\sum_{k=1}^{K} b_k^2\}. \tag{13.8}$$

Here again, we see that $\delta_{b,2}$ will dominate the posterior if it is large relative to $\sum_{k=1}^{K} b_k^2$.

13.4 Monte Carlo EM-Algorithm

Ganguli, Staudenmayer, and Wand (2005) developed a Monte Carlo EM-algorithm for the structural nonparametric regression problem, in the model

$$\mathbf{Y}_i = \sum_{k=0}^{p} \beta_k \mathbf{X}_i^k + \sum_{k=1}^{K} b_k (\mathbf{X}_i - \kappa_k)_+^p + \epsilon_i. \tag{13.9}$$

Their algorithm is relatively simple, although like the Bayesian approach in Section 13.3, a Metropolis–Hastings step is required; see Section 9.3.

We refer to this approach as the GSW-EM (Ganguli, Staudenmayer, Wand-EM) algorithm. The algorithm is given also in Ruppert et al. (2003, Chapter 15.3), although their σ_v^2 is our σ_u^2 and their σ_u^2 is our σ_b^2.

13.4.1 Starting Values

The GSW-EM algorithm starts with estimates of $\mu_x = E(\mathbf{X})$, $\sigma_x^2 = \text{var}(\mathbf{X})$ and $E(\mathbf{X}|\mathbf{W})$; see, for example, Section 4.4 for the regression calibration calculations. Define $\mathcal{V}$ to be the $n \times (p+1)$ matrix with i^{th} row given as $(1, \mathbf{X}_i, \ldots, \mathbf{X}_i^p)$, and define $\mathcal{Z}$ to be the $n \times K$ matrix with i^{th} row $m_2(\mathbf{X}_i)$ given by (13.5). Let $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{W}$, and $\mathcal{E}$ be the $n \times 1$ vectors with i^{th} element $\mathbf{Y}_i$, $\mathbf{W}_i$, $\mathbf{X}_i$ and ϵ_i, respectively. Then the model can be written as

$$\mathcal{Y} = \mathcal{V}\beta_x + \mathcal{Z}b_x + \mathcal{E},$$

where $\text{cov}(b_x) = \sigma_b^2 I$ and $\text{cov}(\mathcal{E}) = \sigma_\epsilon^2 I$. This is a standard linear mixed model, and replacing $\mathbf{X}_i$ with its regression calibration estimate $E(\mathbf{X}_i|\mathbf{W}_i)$, REML is used to obtain starting values for $(\beta_x, b_x, \sigma_b^2, \sigma_\epsilon^2)$.

13.4.2 Metropolis–Hastings Fact

The density of $(\mathbf{X}_1, ..., \mathbf{X}_n)$ given $b_x, \mathcal{Y}, \mathcal{W}$ is proportional to

$$\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\|\mathcal{Y}-\mathcal{V}\beta_x-\mathcal{Z}b_x\|^2-\frac{1}{2\sigma_x^2}\|\mathcal{V}-\mu_x\mathbf{1}\|^2-\frac{1}{2\sigma_u^2}\|\mathcal{W}-\mathcal{V}\|^2\right\}. \quad (13.10)$$

Here $\|\cdot\|$ is the Euclidean norm; see the Guide to Notation. As a result, generating conditional quasi-random variates from $\mathcal{X}$'s conditional distribution can be done using the Metropolis–Hastings algorithm. This fact is also used by Berry, Carroll, and Ruppert (2002).

13.4.3 The Algorithm

We now specify the complete algorithm:

(1) Set $t = 0$.

(2) Use the Metropolis–Hastings algorithm, applied one element at a time, to draw m samples from the distribution of $\mathbf{X}_i$ given $(b_x, \mathcal{Y}, \mathcal{W})$ evaluated at the current estimates of $(\beta_x, b_x, \sigma_b^2, \sigma_\epsilon^2)$. Call these samples $(X_{1i}, ..., X_{mi})$. This is the most time-consuming step in the EM algorithm. The choice of the number of samples m is a problem that remains difficult to solve. In their calculations, Ganguli, Staudenmayer, and Wand started with $m = 50$ and increased m by 10 for each interaction of the EM algorithm, to a maximum of 500.

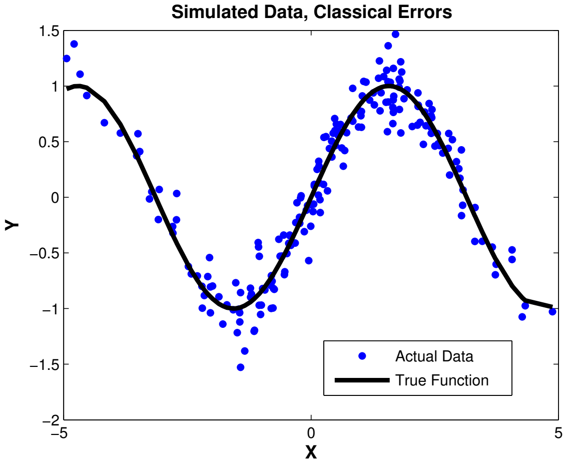


Figure 13.1 *Classical error simulation with sine function. Solid line: true function. The true* **X***-observations are plotted: note how the true function is readily estimated if* **X** *were observable.*

(3) Define

$$P = \begin{bmatrix} \mathcal{V}^t\mathcal{V} & \mathcal{V}^t\mathcal{Z} \\ \mathcal{Z}^t\mathcal{V} & \mathcal{Z}^t\mathcal{Z} + (\sigma_\epsilon^2/\sigma_b^2)I \end{bmatrix}.$$

The value of P is unknown and must be imputed, because we do not observe $\mathcal{V}$.

(4) Define $\mathcal{C} = [\mathcal{V}, \mathcal{Z}]$. With the results from step 2, compute Monte Carlo estimates of the conditional expectations of P, $\mathcal{C}^t\mathcal{Y}$, and $\mathcal{C}^t\mathcal{C}$ given $(\mathcal{Y}, \mathcal{W}, b_x)$. We denote estimates of these quantities by $\widehat{P}$, $\widehat{\mathcal{C}^t\mathcal{Y}}$ and $\widehat{\mathcal{C}^t\mathcal{C}}$. For example, $\widehat{P}$ is computed by defining $\mathcal{V}_j$ to be the same as $\mathcal{V}$ except that the true $\mathbf{X}$-values are replaced by their j^{th} Metropolis–Hastings generated value. Then an upper-left block of $\widehat{P}$ is $m^{-1}\sum_{j=1}^m \mathcal{V}_j^t\mathcal{V}_j$, and upper-right block is $m^{-1}\sum_{j=1}^m \mathcal{V}_j^t\mathcal{Z}$, and so forth.

(5) Holding the estimates from the previous step fixed, run several iterations of an update scheme. For instance, using the standard EM algorithm to compute REML estimates (for example Dempster, Rubin, and Tsutakawa, 1981), the nested updates are as follows. First, update β_x and b_x as $\{\beta_x^t, b_x^t\}^t = \widehat{P}^{-1}\widehat{\mathcal{C}^t\mathcal{Y}}$. Then, update σ_b^2 as $\sigma_b^2 = K^{-1}\{b_x^t b_x + \sigma_\epsilon^2 \text{trace}(\widehat{P}^{-1})$. Finally, update σ_ϵ^2 as follows. Define $\mathcal{D} =$

$\{\beta_x^t, b_x^t\}^t$ and let the current value be $\sigma^2_{\epsilon,\text{curr}}$. Then

$$\sigma_\epsilon^2 = n^{-1}\{\mathcal{Y}^t\mathcal{Y} - 2(\widehat{\mathcal{C}^t\mathcal{Y}})^t\mathcal{D} + \mathcal{D}^t\widehat{\mathcal{C}^t\mathcal{C}}\mathcal{D}\} + n^{-1}\sigma^2_{\epsilon,\text{curr}}\text{trace}(\widehat{\mathcal{C}^t\mathcal{C}}\widehat{P}^{-1}).$$

(6) Update μ_x and σ_x from their current $\mu_{x,\text{curr}}$ and $\sigma^2_{x,\text{curr}}$ using standard point estimates based on the Monte Carlo data from Step 2. Specifically, μ_x is the mean across all the values of components of $(X_{1i}, ..., X_{mi})$. Also, σ_x^2 is the mean across all components of $(X_{ji} - \mu_{x,\text{curr}})^2$.

We terminate the algorithm after plots of the current estimates of the regression function appear to stabilize

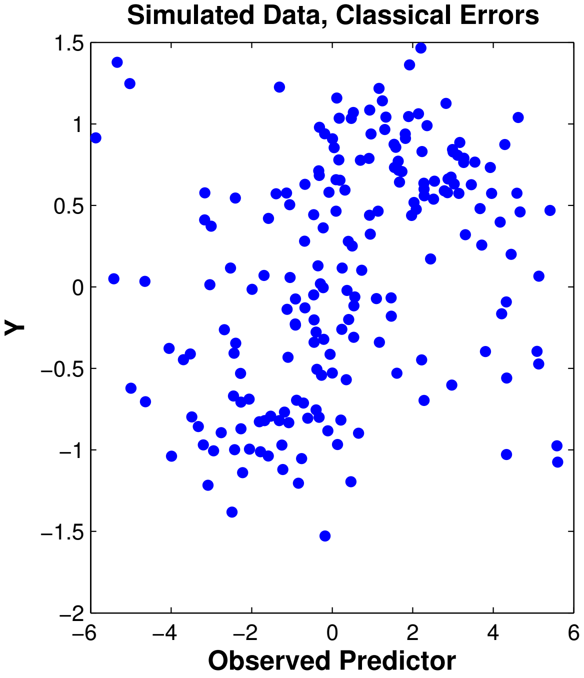


Figure 13.2 *Classical error simulation with sine function. Figure 13.1 gives the true-$\mathbf{X}$ data along with the observed responses, where the sine function is obvious. Here we plot the responses $\mathbf{Y}$ against the observed values $\mathbf{W}$ of the mean of the two mismeasured covariates. Note the lack of features in the data, where the sine function is masked.*

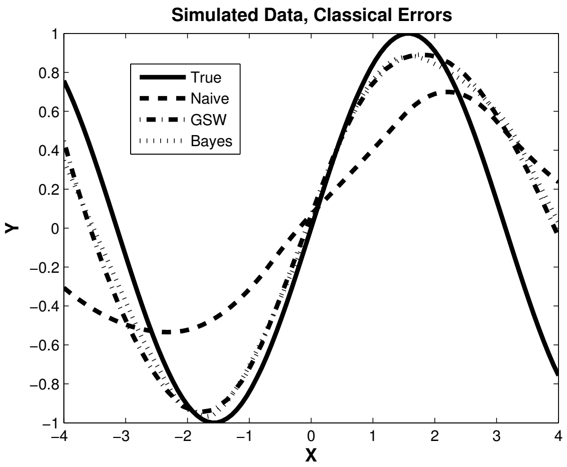


Figure 13.3 *Classical error simulation with sine function. Solid line: true function. Dashed line: naive spline estimate that ignores measurement error. Dot-dashed line: measurement error spline fit by EM-algorithm of Ganguli, et al. (2005). Dotted line: 5-knot Bayesian fit.*

13.5 Simulation with Classical Errors

We simulated data with replicated classical errors, $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$, $j = 1, 2$, and a sinusoidal regression function $\mathbf{Y}_i = \sin(\mathbf{X}_i) + \epsilon_i$. In the simulation $\mathbf{X}_i = \text{Normal}(0.0, 4.0)$, $\mathbf{U}_i = \text{Normal}(0.0, 4.0)$, $\epsilon_i = \text{Normal}\{0, (0.2)^2\}$, and $n = 200$. The actual data and the true regression are given in Figure 13.1, where the true function is apparent.

To see what is going on, we first plot the observed data $(\overline{\mathbf{W}}_i, \mathbf{Y}_i)$. Once again we see the double-whammy of classical measurement error described in Section 1.1: in contrast to the true data in Figure 13.1, the observed data are clearly more variable about any version of a line, hence the loss of power, and the true function is now no longer apparent, hence the loss of features. The idea that one can recreate the true line given in Figure 13.1 from the observed data in Figure 13.2 is daunting.

In Figure 13.3 we see a plot of a Bayes estimate, the true regression function, a naive fit and the EM-algorithm fit of Ganguli, et al. (2005) defined in Section 13.4. The Bayes estimator and the naive fit used five knots at quantiles of the $\mathbf{W}$. The Bayes estimator that had σ_x^2 and σ_ϵ^2 had IG$(1,1)$ priors, μ_x and the β's had Normal$(0, 100)$ priors, and σ_b^2 had an $IG(0.01, 100)$ prior.

The naive fit was a penalized spline fit of $\overline{\mathbf{W}}_i$ to $\mathbf{Y}_i$ and shows clear

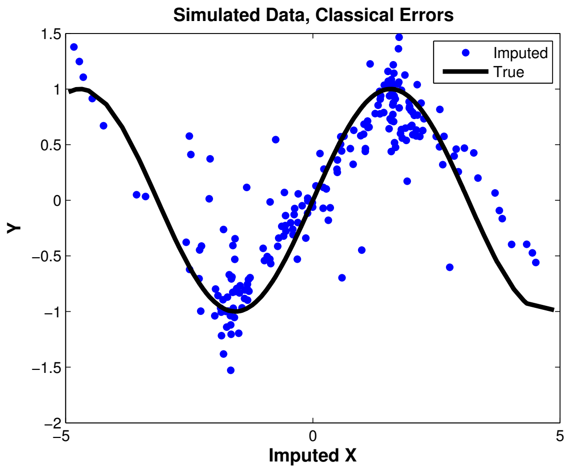


Figure 13.4 *A plot of the imputed* $\mathbf{X}_i$ *versus* $\mathbf{Y}_i$. *The imputed* $\mathbf{X}_i$ *are from the from the last iteration of the Gibbs sampler. Notice the sinusoidal pattern of the true regression function (solid curve labelled "true") is quite evident here, although it was largely hidden in the plot of* $\overline{\mathbf{W}}_i$ *versus* $\mathbf{Y}_i$ *in Figure 13.2.*

bias, both being attenuated toward zeros and having its peaks somewhat mislocated. The Bayes fit, and the EM-algorithm, can virtually recreate the correct function, somewhat remarkable in light of the observed data in Figure 13.2.

The Bayes estimate is the mean of 5,000 iterations of a Gibbs sampler, after a burn-in of 5,000 iterations. Because a spline is a linear model, but with a hierarchical prior for the regression coefficients given by (13.6) and (13.7), the Gibbs sampler was the same as used for linear regression with classical errors in Section 9.4, except that there was an extra step for sampling σ_b^2 using (13.8).

The Gibbs sampler used the naive penalized spline fit to get starting values for the parameters. The starting value for $\mathbf{X}_i$ was $\overline{\mathbf{W}}_i$. The spline was of degree $p = 2$ with five knots placed so that there was approximately an equal number of $\overline{\mathbf{W}}_i$ between each adjacent pair. The knot locations are *not* considered to be unknown parameters and the knots were kept fixed during the Gibbs sampler.

The power of the Gibbs sampler is that it uses both $\mathbf{Y}_i$, $\overline{\mathbf{W}}_i$ and knowledge of the regression function to impute $\mathbf{X}_i$. The result is that often $\mathbf{X}_i$ is estimated with remarkable accuracy. This accuracy can be seen in Figure 13.4, which is a plot of $\mathbf{Y}_i$ against the imputed $\mathbf{X}_i$ from the last iteration of the Gibbs sampler. The shape of the true regression

function is quite obvious, much more so than in the plot of $\mathbf{Y}_i$ against $\overline{\mathbf{W}}_i$.

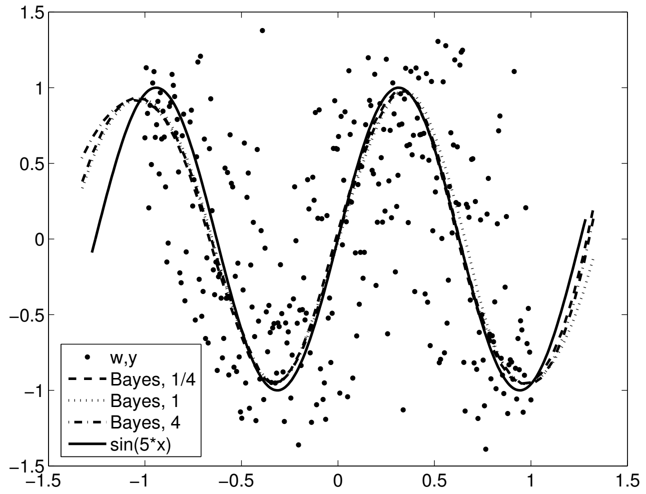


Figure 13.5 *Bayes estimates with Berkson errors. "Bayes, C" is the Bayes estimate for a given* C.

13.6 Simulation with Berkson Errors

Berkson errors were simulated with $\mathbf{W}_i$ equally spaced on $(-1, 1)$, $\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i$, where $\mathbf{U}_i = \text{Normal}\{0, (0.2)^2\}$, $\mathbf{Y}_i = \sin(5X_i) + \epsilon_i$ with $\epsilon_i = \text{Normal}\{0, (0.2)^2\}$, and $n = 300$.

In the case of classical errors, we used knots at quantiles of the $\mathbf{W}_i$. Since the $\mathbf{W}_i$ are more dispersed than the $\mathbf{X}_i$ when the errors are classical, the knots should cover the range of the $\mathbf{X}_i$. In the case of Berkson errors, this might not be true. Therefore, we simulated a set of $\mathbf{X}_i$ by adding $\text{Normal}(0, \tilde{\sigma}_u^2)$ random variates to the $\mathbf{W}_i$. Here, $\tilde{\sigma}_u^2$ is the mean of the prior on σ_u^2. The knots were at quantiles of these simulated $\mathbf{X}_i$. There were 10 knots.

The prior on σ_u^2 was $\text{IG}(1/2, C\sigma_u^2/2)$ where C was 1/4, 1, and 4 corresponding to prior guesses of σ_u^2 equal to 1/4, 1, or 4 times the true value. Since the prior effective sample size is 1, the value of C should not be too important. The Bayes estimates of $s(x) = \sin(5x)$ are shown in Figure 13.5. In Figure 13.6 we see the naive spline fit of $\mathbf{Y}$ to $\mathbf{W}$, the ideal spline fit of $\mathbf{Y}$ to $\mathbf{X}$, the Bayes estimator with $C = 1$, and the true curve.

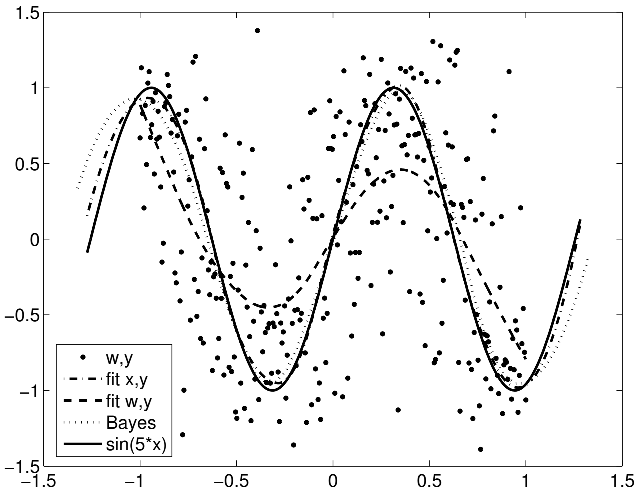


Figure 13.6 *Berkson errors. True curve* (sin(5*x*))*, Bayes estimate with* C *=1 (Bayes), ideal spline fit using true covariates (x-y), and naive spline fit (w-y).*

In Figure 13.7 we see the trace plots for σ_u for the three values of C. One can see that σ_u can be rather accurately estimated despite the lack of replication. It is interesting that σ_u is so well estimated here. In linear regression, σ_u is not identified. For some nonlinear models, such as segmented binary regression, σ_u is theoretically identified but there is so little information about σ_u in the data that, for practical purposes, it is not identified; see the Munich bronchitis example in Section 9.7.3. Here, we have two things going for us when we estimate σ_u:

- The response is continuous not binary.
- The true curve is very nonlinear.

13.7 Semiparametrics: X Modeled Parametrically

The other variant of semiparametric regression is when the effect of $\mathbf{X}$ on $\mathbf{Y}$ is modeled parametrically but at least a component of $\mathbf{Z}$ is modelled nonparametrically. For example, Liang, Härdle, and Carroll (1999) considered the partially linear model

$$E(\mathbf{Y}|\mathbf{X},\mathbf{Z}) = \mathbf{X}^t\beta_x + \theta(\mathbf{Z}),$$

where $\theta(\cdot)$ is an unknown function. Similar to (13.12), the approach is a correction for attenuation. Let Σ_{uu} be the covariance matrix of the measurement errors. An infeasible estimator is

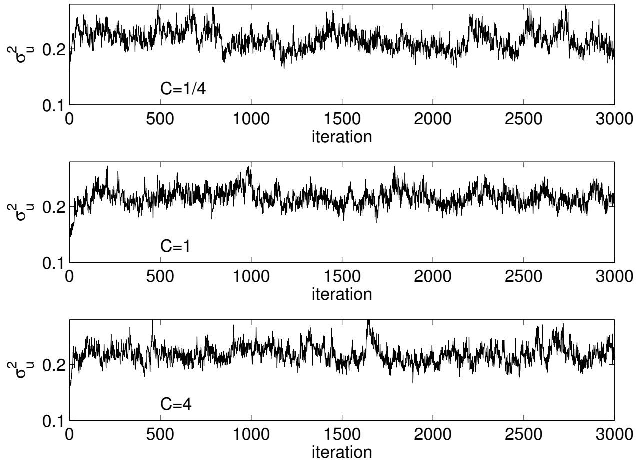


Figure 13.7 *Berkson errors. Estimates of* σ_u^2.

$$\begin{aligned}\widehat{\beta}_{\text{infeas}} &= [n^{-1}\sum_{i=1}^n\{\mathbf{W}_i - E(\mathbf{W}|\mathbf{Z}_i)\}\{\mathbf{W}_i - E(\mathbf{W}|\mathbf{Z}_i)\} - \Sigma_{uu}^t]^{-1}\\ &\quad \times n^{-1}\sum_{i=1}^n\{\mathbf{W}_i - E(\mathbf{W}|\mathbf{Z}_i)\}\{\mathbf{Y}_i - E(\mathbf{Y}|\mathbf{Z}_i)\}. \qquad (13.11)\end{aligned}$$

Liang et al. simply replace $E(\mathbf{W}|\mathbf{Z})$ and $E(\mathbf{Y}|\mathbf{Z})$ with any convenient nonparametric regression, which is feasible because these are all observed quantities. They used kernels because it is easy to prove results for kernels, but splines, etc., can be used as well.

Liang and Wang (2005) considered the partially linear single index model where now $\mathbf{Z}$ is multivariate and

$$E(\mathbf{Y}|\mathbf{X},\mathbf{Z}) = \mathbf{X}^t\beta_x + \theta(\mathbf{Z}^t\beta_z).$$

This is a harder problem, and they developed a clever two-step approach wherein they first estimated β_x without having to worry about β_z, and then updated to get an estimate of β_z. Specifically, $Y - E(\mathbf{Y}|\mathbf{Z}) = \{\mathbf{X} - E(\mathbf{X}|\mathbf{Z})\}^t\beta_x$, which is a partially linear model as described above, so that β_x is easily estimated. They then noted that $E(\mathbf{Y}|\mathbf{Z}) - E(\mathbf{X}^t|\mathbf{Z})\beta_x = \theta(\mathbf{Z}^t\beta_z)$, which is a standard single-index model, for which a host of solutions are known.

13.8 Parametric Models: No Assumptions on X

There has been a great deal of recent activity about parametric response models that attempt to correct for measurement error under minimal assumptions about the latent variable **X** or the measurement error when observing $\mathbf{W} = \mathbf{X}+\mathbf{U}$. In this section, we try to summarize this research concisely, much of it being theoretical in nature. It is probably too soon to tell how this recent literature will affect the practice in the area, because some of the methods are computationally complex or rely on the use of characteristic functions.

Before 1995, there was only one set of literature available that yielded globally consistent estimation, namely, the conditional and corrected methods described in Chapter 7, and that only for special cases, such as linear and logistic regression and under strong parametric assumptions about the measurement error distribution. The newest literature expands the models that can be considered. When replications of **W** are available, assumptions about the error distribution may also sometimes be avoided.

13.8.1 Deconvolution Methods

13.8.1.1 Least Squares Methods

Suppose that the mean of **Y** given **X** is given as $m_{\mathbf{Y}}(\mathbf{X}, \mathcal{B})$. Remember that $f_X(\cdot)$ is the unknown density function of **X**, $f_{\mathbf{W}|\mathbf{X}}(\cdot)$ is the density function of **W** given **X**, and $f_W(\cdot)$ is the density function of **W**. Then, as described in (12.1) in Section 12.1, the mean of **Y** given an observed **W** is

$$m_{\mathbf{Y}|\mathbf{W}}(\mathbf{W}, \mathcal{B}, f_X, f_W, f_{W|X}) = \frac{1}{f_W(\mathbf{W})}\int m_{\mathbf{Y}}(x, \mathcal{B}) f_X(x) f_{W|X}(\mathbf{W}|x)dx.$$

The obvious approach then is to do some form of least squares. Let $\omega(\mathbf{W})$ be a weight function. Then, assuming that all the density functions are known, one way to estimate $\mathcal{B}$ is to minimize

$$\sum_{i=1}^{n} \omega(\mathbf{W}_i)\{Y_i - m_{\mathbf{Y}|\mathbf{W}}(\mathbf{W}_i, \mathcal{B}, f_X, f_W, f_{W|X})\}^2.$$

To implement this idea, one has to produce estimates for the density functions $(f_X, f_W, f_{W|X})$. For example, suppose that $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where **U** is normally distributed with mean zero and known variance σ_u^2, so that $f_{W|X}(w|x) = \sigma_u^{-1}\phi\{(w-x)/\sigma_u\}$, where $\phi(\cdot)$ is the standard normal density function. Let $\widehat{f}_X$ be the deconvoluting kernel density estimator defined in Section 12.1, let $\widehat{f}_W$ be the corresponding regular kernel density estimator. Then it would be tempting to estimate $\mathcal{B}$ by minimizing

$$\sum_{i=1}^{n} \omega(\mathbf{W}_i)\{Y_i - m(\mathbf{W}_i, \mathcal{B}, \widehat{f}_X, \widehat{f}_W, f_{W|X})\}^2.$$

This line of attack has been taken, for example, by Taupin (2001). She allowed the unknown mean function $f(\mathbf{X}, \mathcal{B})$ to have different amounts of smoothness, and showed that if it is smooth in any standard sense, then the algorithm will produce estimates of $\mathcal{B}$ that have parametric rates of convergence, although standard error estimates were not obtained.

13.8.1.2 Partially Linear Models

Liang (2000) used deconvolution methods in the partially linear model where $E(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) = \mathbf{Z}^t\beta + \theta(X)$, where $\theta(\cdot)$ is an unknown function. This is the same model as (13.1) that we analyzed in Sections 13.2 and 13.3 using splines. Liang started with the infeasible "estimator"

$$\begin{aligned}\widehat{\beta}_{\text{infeas}} &= \left[n^{-1}\sum_{i=1}^{n}\{\mathbf{Z}_i - E(\mathbf{Z}|\mathbf{X}_i)\}\{\mathbf{Z}_i - E(\mathbf{Z}|\mathbf{X}_i)\}^t\right]^{-1} \\ &\quad \times n^{-1}\sum_{i=1}^{n}\{\mathbf{Z}_i - E(\mathbf{Z}|\mathbf{X}_i)\}\{\mathbf{Y}_i - E(\mathbf{Y}|\mathbf{X}_i)\}. \qquad (13.12)\end{aligned}$$

He then replaced $E(\mathbf{Z}|\mathbf{X}_i)$ and $E(\mathbf{Y}|\mathbf{X}_i)$ by deconvoluting kernel regression estimators. For normally distributed measurement errors, he derived a limiting normal distribution for the resulting estimate of β, under the restriction that **Z** and **X** were independent.

Zhu and Cui (2003) considered the same model, except that they allowed **Z** to also be observed with error independent of the error in **X**. They did not require that **X** and **Z** be independent. Their method is effectively a correction for attenuation version of the infeasible estimate.

13.8.2 Models Linear in Functions of **X**

Schennach (2004a) considered models of the form

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \mathbf{Z}^t\beta_z + h(\mathbf{X})\beta_x,$$

She noted that $\mathcal{B} = (\beta_z^t, \beta_x^t)^t$ can be recovered as

$$\mathcal{B} = \begin{bmatrix} E(\mathbf{Z}\mathbf{Z}^t) & E\{\mathbf{Z}h^t(\mathbf{X})\} \\ E\{h(\mathbf{X})\mathbf{Z}^t\} & E\{h(\mathbf{X})h^t(\mathbf{X})\}\end{bmatrix}^{-1} [E(\mathbf{Z}^t\mathbf{Y}), E\{h^t(\mathbf{X})\mathbf{Y}\}]^t.$$

Her work is striking because the measurement error distribution need not be specified. Indeed, she allowed for a second measurement, $\mathbf{T} = \mathbf{X} + \mathbf{V}$, where **V** has mean zero and is independent of everything else (her conditions are actually somewhat weaker than this). No assumptions are made about the distribution of either the measurement error **U** or the measurement error **V**. She showed how to go about estimating the moments necessary to identify $\mathcal{B}$, and derived the limiting distribution of her estimator.

While Schennach's approach uses characteristic function, this is not deconvolution in any sense. However, Schennach (2004b) used much the same approach to develop a deconvoluting kernel density estimator without assumptions about the distribution of the measurement errors; see also Section 12.2.7.

13.8.3 Linear Logistic Regression with Replicates

Huang and Wang (2001) considered linear logistic regression when there are replicated error prone measures, so that $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$ is observed for $j = 1, 2$. Although we believe that simple regression calibration is generally quite well suited to this problem, at least for the purpose of having a complete asymptotic theory, there is a need to have consistent, and not just approximately consistent, estimation procedures. Of course, if the measurement errors are normally distributed, we have already described such a methodology; see Chapter 7.

Huang and Wang made no assumptions about the distribution of the measurement errors or about the distribution of **X**. Their notation is somewhat difficult to decipher (sometimes too much generality detracts from otherwise excellent papers), but in the case of two replicates, works like this. Define

$$\begin{aligned}
\Psi_{i1}(\beta_0, \mathcal{B}) &= \sum_{j \neq k} \{\mathbf{Y}_i - 1 + \mathbf{Y}_i \mathrm{exp}(-\beta_0 - \mathbf{Z}_i^t \beta_z - \mathbf{W}_{ij}^t \beta_x)\}(1, W_{ik}^t)^t; \\
\Psi_{i2}(\beta_0, \mathcal{B}) &= \sum_{j \neq k} \{\mathbf{Y}_i + (\mathbf{Y}_i - 1)\mathrm{exp}(\beta_0 + \mathbf{Z}_i^t \beta_z + \mathbf{W}_{ij}^t \beta_x)\}(1, W_{ik}^t)^t.
\end{aligned}$$

For $j = 1, 2$, let $\widehat{\beta}_{0j}$ and $\widehat{\mathcal{B}}_j$ be a solution to $0 = \sum_{i=1}^n \Psi_{ij}(\beta_0, \mathcal{B})$, and let $\widehat{\Lambda}$ be the sample covariance matrix of the terms $\{\Psi_{i1}(\widehat{\beta}_{01}, \widehat{\mathcal{B}}_1)$ and $\{\Psi_{i2}(\widehat{\beta}_{02}, \widehat{\mathcal{B}}_2)$ when combined. Then estimate $\mathcal{B}$ by minimizing the quadratic form

$$\begin{bmatrix} \sum_{i=1}^n \Psi_{i1}(\beta_{01}, \mathcal{B}) \\ \sum_{i=1}^n \Psi_{i1}(\beta_{02}, \mathcal{B}) \end{bmatrix}^t \widehat{\Lambda} \begin{bmatrix} \sum_{i=1}^n \Psi_{i1}(\beta_{01}, \mathcal{B}) \\ \sum_{i=1}^n \Psi_{i1}(\beta_{02}, \mathcal{B}) \end{bmatrix}.$$

Huang and Wang derived the limiting distribution of the estimator and also allowed for the possibility that there are more than two replicates.

13.8.4 Doubly Robust Parametric Modeling

The approaches outlined in this section have been more or less ad hoc, some of them with fairly daunting computational issues.

Tsiatis and Ma (2004) avoided the ad hoc nature of the approach by developing a modern semiparametric approach to the problem. They started with a full parametric model for **Y** given $(\mathbf{X}, \mathbf{Z})$ in terms of a parameter $\mathcal{B}$, and they assumed that the distribution of **W** given **X** is completely known, or is known except for a parameter. Their methodology has the following features:

- They first specify a candidate distribution for **X**, which may or may not be correct.
- Their method, which is not maximum likelihood for this candidate distribution, provides a consistent estimate of the parameter $\mathcal{B}$, no matter what the actual distribution of **X** is.
- Their method is also the most efficient approach among all methods that are consistent in the above sense. Thus, for example, for canonical exponential families with normally distributed measurement error, their method reduces to the conditional score method described in Section 7.2.2.

Tsiatis and Ma applied their methodology to the logistic model that is quadratic in **X**, with impressive improvement in bias reduction compared even to their version of regression calibration (similar but not the same as the one proposed in this book), much less the naive method.

The methodology can be briefly summarized as follows. Suppose that a density function $f^*_{X|Z}(x|z)$ has been hypothesized as the density for **X** given **Z**. Let $\mathcal{S}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathcal{B})$ be the score function if **X** were observable, that is the derivative of the loglikelihood function. Define $\mathcal{S}^*(\mathbf{Y}, \mathbf{W}, \mathbf{Z}, \mathcal{B}) = E^*\{\mathcal{S}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathcal{B})|\mathbf{Y}, \mathbf{W}, \mathbf{Z}\}$, where the superscript means that the expectation is taken with respect to the hypothesized model. Then there is a function $a(\mathbf{X}, \mathbf{Z})$ with the property that it solves the integral equation

$$E\{\mathcal{S}^*(\mathbf{Y}, \mathbf{W}, \mathbf{Z}, \mathcal{B})|\mathbf{X}, \mathbf{Z}\} = E\left[E^*\{a(\mathbf{X}, \mathbf{Z})|\mathbf{Y}, \mathbf{W}, \mathbf{Z}\}|\mathbf{X}, \mathbf{Z}\right]. \quad (13.13)$$

Further define

$$\mathcal{S}_{\mathrm{eff}}(\mathbf{Y}, \mathbf{W}, \mathbf{Z}, \mathcal{B}) = \mathcal{S}^*(\mathbf{Y}, \mathbf{W}, \mathbf{Z}, \mathcal{B}) - E^*\{a(\mathbf{X}, \mathbf{Z})|\mathbf{Y}, \mathbf{W}, \mathbf{Z}\}.$$

Then, Tsiatis and Ma proposed to estimate $\mathcal{B}$ by solving the equation $0 = \sum_i \mathcal{S}_{\mathrm{eff}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i, \mathcal{B})$.

Theoretically, this approach has a great deal to be said for it. One can use best guesses to get something near a maximum likelihood solution, but still have robustness against specifying the model for **X** incorrectly. The practical difficulties are the following:

- First, one has to be able to compute $\mathcal{S}^*(\mathbf{Y}, \mathbf{W}, \mathbf{Z}, \mathcal{B})$, which requires numerical integration, but then of course so too does maximum likelihood.
- More of an issue is actually solving the integral equation (13.13). In their Section 4.2, Tsiatis and Ma (2004) came up with an approximate solution. Specifically, they discretized $\mathbf{X}$, stating that it takes on a finite number of values, and then they specified the probability of these values given $\mathbf{X}$. In their example, they allowed $\mathbf{X}$ to take on 15 different values, and made the probabilities of $\mathbf{X}$ given $\mathbf{Z}$ to be proportional to the density function of $\mathbf{X}$ given $\mathbf{Z}$ at these 15 values. They then solved (13.13) in this discrete setting, which is little more than solving linear equations with somewhat messy input arguments.

The Tsiatis and Ma methodology has considerable potential. However, more numerical work will be needed in order to understand how to discretize, and more important, it would be very useful if multipurpose software could be developed.

Bibliographic Notes

Berry, Carroll, and Ruppert (2002) developed the Bayesian spline methodology for nonparametric regression with covariate measurement error that is the basis for this chapter. Mallick, Hoffman, and Carroll (2002) developed semiparametric methods for Berkson errors in the Nevada Test Site example, although in their case they knew the variance of the Berkson errors. The application of this regression spline methodology to Berkson errors with unknown Berkson error variance is new.

It has long been known that, in parametric problems, the Berkson error variance in the model $\mathbf{X} = \mathbf{W} + \mathbf{U}$ is identified if the true regression function is not linear; see, for example, Rudemo, Ruppert, and Streibig (1989) in Section 4.7.3. These results apply to our approach, which are flexible and nonlinear parametric methods, hence semiparametric. In purely nonparametric regression problems, identifiability of the measurement error variance and hence of the regression function is harder. Delaigle et al. (2006) pointed out that with Berkson errors, if the true regression function is $m_{\mathbf{Y}}(\cdot)$, then what we estimate is $\gamma(\mathbf{W}) = E\{m_{\mathbf{Y}}(\mathbf{W}+\mathbf{U})\}$, and identifiability of the true regression function means we need to be able to recover it from $\gamma(\cdot)$. This can be tricky, since in Berkson models $\mathbf{X}$ is more variable than $\mathbf{W}$. An interesting theoretical issue is whether one can hope to recover the true function $m_{\mathbf{Y}}(\cdot)$ beyond the range of the observed $\mathbf{W}$-values. At first, this seems difficult, if not impossible, but simulation results suggest otherwise, at least partly; see Figure 13.5, where the estimates follow the true function beyond the $\mathbf{W}$-values.

CHAPTER 14

SURVIVAL DATA

Survival analysis has developed from the analysis of life tables in actuarial sciences and has enjoyed remarkable success with modern applications in medicine, epidemiology, and the social sciences. The popularity of survival analysis models, such as the Cox proportional hazards model, is probably surpassed only by the popularity of standard linear regression models.

Survival data are the product of a continuous death process coupled with a censoring mechanism. Typically, the death rate depends on a number of factors, and time to death is only partially observed for those subjects with censored observations. Standard analyses of survival data assume that all covariates affecting survival rates are observed without error. However, in many applications some of the covariates are subject to measurement error or are available without error only for a subsample of the population.

14.1 Notation and Assumptions

Like most research areas in statistics, survival analysis has several standard sets of notations. In this chapter, we will follow notation introduced by Miller (1998). Assume that n subjects are observed over time and their failure times $\mathbf{T}_1, \ldots, \mathbf{T}_n$ are subject to right censoring and $\mathbf{C}_1, \ldots, \mathbf{C}_n$ are the corresponding censoring times. Let

$$\delta_i = I(\mathbf{T}_i < \mathbf{C}_i)$$

be the failure indicator and

$$\mathbf{Y}_i = \min(\mathbf{T}_i, \mathbf{C}_i)$$

be the time to failure or censoring for subject i. Denote by

$$R_i = \{j : \mathbf{Y}_j \geq \mathbf{Y}_i\}, \tag{14.1}$$

the risk set when the event corresponding to subject i occurs. R_i is the index set for those subjects who have not failed and are uncensored at the time the i^{th} subject fails or is censored. The at-risk indicator process for the i^{th} subject is defined as

$$Y_i(t) = I(\mathbf{Y}_i \geq t).$$

We assume that the survival probability for each subject depends on covariates that are subject to measurement error, $\mathbf{X}_i$, as well as on covariates that are not, $\mathbf{Z}_i$. The covariate $\mathbf{X}_i$ is measured through the usual classical measurement error model

$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i, \tag{14.2}$$

where the distribution of $\mathbf{U}_i$ is known or estimable. We also assume that $(\mathbf{T}_i, \mathbf{X}_i, \mathbf{C}_i, \mathbf{U}_i)$ are iid random vectors, $\mathbf{C}_i$ is independent of $(\mathbf{T}_i, \mathbf{X}_i)$, and $\mathbf{U}_i$ is independent of $(\mathbf{T}_i, \mathbf{X}_i, \mathbf{C}_i)$. The observed data are the vectors $(\mathbf{Y}_i, \delta_i, \mathbf{W}_i, \mathbf{Z}_i)$, where $(\mathbf{Y}_i, \delta_i)$ is a proxy observation for $(\mathbf{T}_i, \mathbf{C}_i)$ and $\mathbf{W}_i$ is a proxy observation for $\mathbf{X}_i$.

The distribution of the failure time, $\mathbf{T}_i$, is completely described by the hazard rate

$$\lambda_i(t|\mathbf{X}_i, \mathbf{Z}_i) = \lim_{dt \downarrow 0} \frac{P(t < \mathbf{T}_i < t + dt|\mathbf{X}_i, \mathbf{Z}_i)}{dtP(\mathbf{T}_i > t|\mathbf{X}_i, \mathbf{Z}_i)}$$

and can be interpreted as the instantaneous risk that the time $\mathbf{T}_i$ of an event equals t conditional on no events for subject i prior to time t. The proportional hazards model introduced by Cox (1972) is the most commonly used model for the hazard rate and assumes that

$$\lambda_i(t|\mathbf{X}_i, \mathbf{Z}_i) = \lambda_0(t)\exp(\beta_x^t\mathbf{X}_i + \beta_z^t\mathbf{Z}_i), \tag{14.3}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function that does not depend on the covariate values. The baseline cumulative hazard function is $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. In the standard regression case when $\mathbf{X}_i$ are observed, Cox (1972) suggested that inference on β_x and β_z be based on the log partial likelihood function

$$l(\beta_x, \beta_z) = \sum_{i=1}^n \delta_i \left[\beta_x^t\mathbf{X}_i + \beta_z^t\mathbf{Z}_i - \log\left\{\sum_{j \in R_i} \exp(\beta_x^t\mathbf{X}_j + \beta_z^t\mathbf{Z}_j)\right\}\right], \tag{14.4}$$

which does not depend on $\lambda_0(\cdot)$. An alternative strategy is to use the log of the full likelihood of the model (14.3)

$$\begin{aligned} L(\beta_x, \beta_z) &= \textstyle\sum_{i=1}^n \delta_i \left[\beta_x^t\mathbf{X}_i + \beta_z^t\mathbf{Z}_i + \log\{\lambda_0(\mathbf{Y}_i)\}\right] \\ &\quad - e^{\beta_x^t\mathbf{X}_i + \beta_z^t\mathbf{Z}_i} \int_0^{\mathbf{Y}_i} \lambda_0(s)ds. \end{aligned} \tag{14.5}$$

14.2 Induced Hazard Function

When $\mathbf{X}$ is unobservable and instead we observe a surrogate $\mathbf{W}$, Prentice (1982) introduced the induced hazard function for the Cox regression model as

$$\begin{aligned} \lambda(t|\mathbf{Z}, \mathbf{W}) &= E\left[\lambda(t|\mathbf{X}, \mathbf{Z})\,|T \geq t, \mathbf{Z}, \mathbf{W}\right] \\ &= \lambda_0(t)\exp\left(\beta_z^t\mathbf{Z}\right)E\left\{\exp(\beta_x^t\mathbf{X}|\mathbf{T} \geq t, \mathbf{Z}, \mathbf{W})\right\}. \end{aligned} \tag{14.6}$$

As shown by Prentice (1982) and by Pepe, Self, and Prentice (1989), the difficulty is that the conditional expectation in (14.6) for the observed data depends upon the unknown baseline hazard function λ_0. This dependence is due to the conditioning on $(T \geq t)$. The induced hazard function does not factor into a product of an arbitrary baseline hazard and a term that depends only on observed data and an unknown parameter, and the methodology for proportional hazards regression cannot be applied without modification.

An important simplification occurs when the failure events are rare, that is, when the probability of survival beyond time t, $P(T \geq t)$, is close to 1. The rare-event assumption implies that the hazard (14.6) of the observed data can be approximated by

$$\lambda^*(t|\mathbf{Z}, \mathbf{W}) = \lambda_0(t)\exp\left(\beta_z^t\mathbf{Z}\right)E\left\{\exp(\beta_x^t\mathbf{X}|\mathbf{Z}, \mathbf{W})\right\}. \tag{14.7}$$

A special case that leads directly to regression calibration is when $\mathbf{X}$ given $\mathbf{Z}$ and $\mathbf{W}$ is normally distributed with mean $m(\mathbf{Z}, \mathbf{W}, \gamma)$ and with constant covariance matrix Σ. In this case the approximate hazard function is, from (14.7),

$$\lambda^*(t|\mathbf{Z}, \mathbf{W}) = \lambda_0^*(t)\exp\left\{\beta_x^t m(\mathbf{Z}, \mathbf{W}, \gamma)\right\},$$

where $\lambda_0^*(t) = \lambda_0(t)\exp(0.5\beta_x^t\Sigma\beta_x)$, which is still arbitrary since λ_0 is arbitrary.

14.3 Regression Calibration for Survival Analysis

One of the first applications of regression calibration was proposed by Prentice (1982) for estimating the parameters in a Cox model. The idea of regression calibration is to replace the covariate of interest by its conditional mean $E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) = m(\mathbf{Z}, \mathbf{W}, \gamma)$ and is a first-order bias-correction method.

14.3.1 Methodology and Asymptotic Properties

The procedure starts by estimating $\mathbf{X}$ by $m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \widehat{\gamma})$, where $\widehat{\gamma}$ is an estimator of γ that could be obtained as in Section 4.4. The next step is to maximize the log partial likelihood (14.4), where $\mathbf{X}_i$ is replaced by $\widehat{\mathbf{X}}_i = m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \widehat{\gamma})$. If $\mathbf{X}_i^* = m_{\mathbf{X}}(\mathbf{Z}, \mathbf{W}, \gamma)$, then the approximate regression calibration Cox model assumes that the hazard function for the observed data is

$$\lambda(t|\mathbf{X}_i^*, \mathbf{Z}_i) = \lambda_0^*(t)\exp(\beta_x^{*t}\mathbf{X}_i^* + \beta_z^{*t}\mathbf{Z}_i). \tag{14.8}$$

Under the regularity conditions in Wang, Wang, and Carroll (1997) one can show that the parameter estimator obtained by maximizing (14.4) with $\mathbf{X}_i$ replaced by $\widehat{\mathbf{X}}_i$ is a consistent, asymptotically normal, estimator of β_x^*. Because model (14.8) is just an approximation of the true model, these results are only approximate for the parameter, β_x, of the true model. In practice, model (14.8) is often a good approximation of model (14.6).

A major advantage of regression calibration is that, after fitting a reasonable model for $E(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)$, one can use existent software designed for proportional hazards models, such as R or S–plus (`coxph()` and `survreg()` functions) or SAS (`PHREG` procedure), to produce first-order bias-corrected estimators of the parameters of a Cox model with covariates subject to measurement error.

14.3.2 Risk Set Calibration

Clayton (1991) proposed a modification of regression calibration that does not require events to be rare. If the $\mathbf{X}$'s were observable, and if $\mathbf{X}_i$ is the covariate associated with the i^{th} event, in the absence of ties the usual proportional hazards regression would maximize

$$\prod_{i=1}^{k} \frac{\exp(\beta_x^t \mathbf{X}_i)}{\sum_{j\in R_i}\exp(\beta_x^t \mathbf{X}_j)},$$

where R_i is the risk set (14.1) at the time when failure or censoring of subject i occurs. Clayton suggested using regression calibration within each risk set, R_i, given in (14.1). He assumed that the true values $\mathbf{X}$ within the i^{th} risk set are normally distributed with mean μ_i and variance σ_x^2, and that within this risk set $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is normally distributed with mean zero and variance σ_u^2. Neither σ_x^2 nor σ_u^2 depends upon the risk set in his formulation. Given an estimate $\widehat{\sigma}_u^2$, one can construct an estimate of $\widehat{\sigma}_x^2$ just as in the equations following (4.4).

Clayton thus modified regression calibration by using it within each risk set. Within each risk set, he applied the formula (4.5) for the best unbiased estimate of the $\mathbf{X}$'s. Specifically, in the absence of replication, for any member of the i^{th} risk set, the estimate of the true covariate $\mathbf{X}$ is

$$\widehat{\mathbf{X}} = \widehat{\mu}_i + \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2}(\mathbf{W} - \widehat{\mu}_i),$$

where $\widehat{\mu}_i$ is the sample mean of the $\mathbf{W}$'s in the i^{th} risk set.

As with regression calibration in general, the advantage of Clayton's method is that no new software need be developed, other than calculating the means within risk sets. Formula (4.5) shows how to generalize this method to multivariate covariates and covariates measured without error.

14.4 SIMEX for Survival Analysis

The simulation–extrapolation (SIMEX) procedure proposed by Cook and Stefanski (1994) and presented in detail in Chapter 5 is a general methodology that extends naturally to survival analysis. For simplicity of presentation, we consider the case when only one variable is measured with error.

The essential idea is to simulate new data by adding increasing amounts of noise to the measured values $\mathbf{W}_i$ of the error prone covariate $\mathbf{X}_i$, compute the estimator on each simulated data set, model the expectation of the estimator as a function of the measurement error variance, and extrapolate back to the case of no measurement error. More precisely, if σ_u^2 is the variance of the measurement error, then for each ζ on a grid of points between $[0, 2]$ we simulate

$$\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \sqrt{\zeta}\mathbf{U}_{b,i}, \quad b = 1, \ldots, B, \tag{14.9}$$

where $\mathbf{U}_{b,i}$ are normal, mean zero, independent random variables with variance σ_u^2, and B is the number of simulations for each value of ζ. The measurement error variance of the contaminated observations $\mathbf{W}_{b,i}(\zeta)$ is $(1+\zeta)\sigma_u^2$, and the case of no measurement error corresponds to $\zeta = -1$.

By replacing $\mathbf{X}_i$ with $\mathbf{W}_{b,i}(\zeta)$ in the hazard function (14.3), we obtain

$$\lambda_i\{t|\mathbf{W}_{b,i}(\zeta), \mathbf{Z}_i\} = \lambda_0(t)\exp\{\beta_x^t\mathbf{W}_{b,i}(\zeta) + \beta_z^t\mathbf{Z}_i\}, \tag{14.10}$$

and either the partial likelihood (14.4) or the full likelihood (14.5) could be used to produce estimators $\widehat{\beta}_x^b(\zeta)$ and $\widehat{\beta}_z^b(\zeta)$. For each level of added noise ζ one obtains

$$\widehat{\beta}_x(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\beta}_x^b(\zeta), \quad \widehat{\beta}_z(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\beta}_z^b(\zeta)\ .$$

A quadratic or rational extrapolant, as described in Section 5.3.2, can then be used to obtain the estimated values corresponding to $\zeta = -1$.

For the case of multivariate failure time data, Greene and Cai (2004) have established the consistency and asymptotic normality of the SIMEX estimator when the measurement error variance and an exact extrapolant are known. Li and Lin (2003a) have used SIMEX coupled with the EM algorithm to provide inference for clustered survival data when some of the covariates are subject to measurement error.

The ideas extend to the case in which more than one predictor is prone to measurement error. Suppose $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$ with $\mathbf{U}_i$ independent Normal$(0, \Omega)$, where Ω is a known positive definite $q \times q$ variance matrix. If $\Omega^{1/2}$ is its positive square root, then remeasured data is generated as

$$\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \sqrt{\zeta}\,\Omega^{1/2}\,\mathbf{U}_{b,i}(\zeta),$$

where $\mathbf{U}_{b,i}(\zeta)$ are independent Normal$(0, I_q)$ vectors and ζ is a positive scalar. Note that

$$\mathrm{Cov}\big\{\mathbf{W}_{b,i}(\zeta)\big\} = (1+\zeta)\,\Omega,$$

which converges to the zero matrix as $\zeta \to -1$. After this, the rest of the simulation and extrapolation steps are conceptually similar.

14.5 Chronic Kidney Disease Progression

To illustrate the regression calibration and SIMEX methodologies in survival analysis, we analyze time-to-event data, where the event is detection of primary coronary kidney disease (CKD). Primary CKD could be viewed as the least severe phase of kidney disease and is typically defined in relationship to the estimated glomerular filtration rate (eGFR) of the kidney. Primary CKD is defined as either achievement of followup eGFR < 60 or a post baseline CKD hospitalization or death (Marsh-Manzi, Crainiceanu, Astor, et al., 2005).

Specifically, we are interested in testing whether African-Americans are at higher risk of CKD progression. Data were obtained from the Atherosclerosis Risk in Communities (ARIC) study, a large multipurpose epidemiological study conducted in four U.S. communities (Forsyth County, NC; suburban Minneapolis, MN; Washington County, MD; and Jackson, MS). A detailed description of the ARIC study is provided by the ARIC investigators (1989). In short, from 1987 through 1989, $15,792$ male and female volunteers aged 45 through 64 were recruited from these communities for a baseline and three subsequent visits. For the purpose of this study, all primary CKD events up to January 1, 2003 were included and the time-to-event data were obtained from annual participant interviews and review of local hospital discharge lists and county death certificates.

The estimated glomerular filtration rate (eGFR) is a measure of kidney function and characterizes the different stages of kidney disease. eGFR is subject to measurement error, and the measurement error variance was estimated from a different replication study. We consider Cox models for time-to-CKD events and we include eGFR, an indicator of African-American race, age at baseline visit, and sex as covariates.

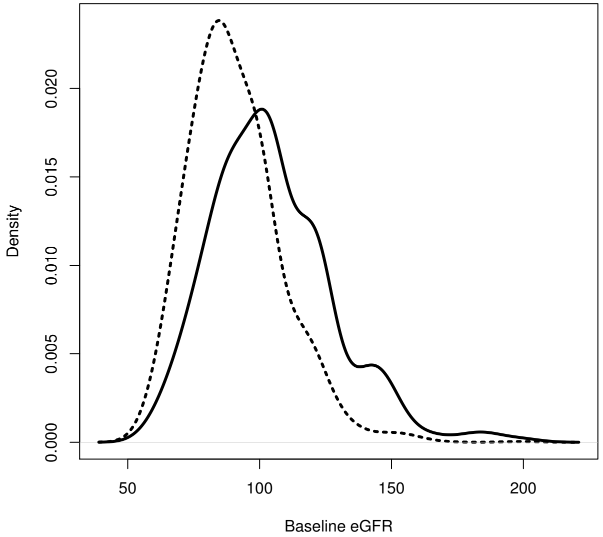


Figure 14.1 *Baseline estimated glomerular filtration rate (eGFR) for African-Americans (solid line) and others (dashed line).*

14.5.1 Regression Calibration for CKD Progression

Regression calibration is a first-order bias-reduction method that works well when the covariates subject to measurement error enter the model linearly. Because eGFR has a nonlinear effect on survival probability, in this section we consider only subjects with baseline eGFR less than 120. This is done only to illustrate regression calibration. One is primarily interested in the relationship between survival and eGFR over the entire range of eGFR, and in Section 14.5.2 we model this nonlinear relationship with a spline and correct for measurement error with SIMEX. Several subjects were omitted from our analyses due to missing data or baseline eGFR values smaller than 60, indicating decreased baseline kidney function. This reduced the number of subjects from $15,792$ to $15,080$ in our full data set and $13,359$ in the reduced data set (eGFR $<$ 120).

Figure 14.1 shows the estimated probability density of baseline eGFR for African-Americans compared to others, indicating better baseline kidney function for African-Americans. We considered a Cox model de-

	eGFR	AA	Age	Sex
Naive	-0.061	0.55	0.075	-0.01
SE	0.0022	0.061	0.0048	0.052
Reg. Cal.	-0.105	0.84	.064	0.024
SE	0.0038	0.064	0.0049	0.052

Table 14.1 *Estimates and standard errors (SE) of risk factors using a reduced ARIC data set* $(13,359$ *subjects) corresponding to* eGFR < 120 *and events observed from first to second visit. Naive, regression on observed eGFR; "AA" is African-American race.*

scribing the time to primary CKD events with covariates eGFR, an indicator of African-American race, sex, and age. Table 14.1 compares results of the naive analysis, which uses the observed eGFR values, with the regression calibration, which uses the means of eGFR conditional on the observed eGFR and the other covariates. Not accounting for measurement error in eGFR would decrease the size of the effect of eGFR by 42% and of the African-American race indicator by 35%, and would increase the effect of age by 17%. The effect of sex on CKD progression was not statistically significant under either the naive or the regression calibration procedure.

The measurement error variance was estimated using data from the Third National Health and Nutrition Examination Survey (NHANES III). Duplicate eGFR measurements were obtained for each of 513 participants aged 45 to 64 with eGFR ≥ 60 from two visits at a median of 17 days apart (Coresh et al., 2002). The estimated measurement error variance was $\widehat{\sigma}_u^2 = 77.56$, was treated as a constant in our analyses, and corresponded to a reliability of 0.80 for eGFR when all subjects with eGFR ≥ 60 were considered and only 0.60 for eGFR of subjects with $60 \leq$ eGFR < 120.

14.5.2 SIMEX for CKD Progression

Given the nonlinear relationship in the full data set between eGFR and the hazard ratio, we fit the following Cox proportional hazard model

$$\lambda_i(t) = \lambda_0(t) \exp\{m_{\mathbf{Y}}(\text{eGFR}_i) + \beta_2 \text{AA}_i + \beta_3 \text{Age}_i + \beta_4 \text{Sex}_i\}, \quad (14.11)$$

	AA	Age	Sex
Naive	0.50	0.070	0.011
SE	0.059	0.0047	0.051
SIMEX	0.63	0.054	0.061
SE	0.062	0.0049	0.052

Table 14.2 *Estimates and standard errors (SE) of risk factors using all subjects with* eGFR > 60 $(15,080$ *subjects) using events up to 2002. Naive is the regression using the observed eGFR; "AA" is African-American race.*

where $m_{\mathbf{Y}}(\cdot)$ is a function of the eGFR, and AA denotes the African-American race. We used a linear spline with four equally spaced knots between eGFR $= 70$ and eGFR $= 165$. More precisely,

$$m_{\mathbf{Y}}(x) = \beta_1 x + \sum_{k=1}^{4} \alpha_k (x - \kappa_k)_+, \quad (14.12)$$

where κ_k, $k = 1, \ldots, 4$ are the knots of the spline and a_+ is equal to a if $a > 0$ and 0 otherwise. In this parameterization, the α_k parameter represents the change in the slope of the log hazard ratio at knot κ_k corresponding to eGFR. The proportional hazard model (14.11) using the linear spline (14.12) with fixed knots to describe the effect of eGFR is linear in the α and β parameters but it is nonlinear in the variable measured with error.

Following the SIMEX methodology described in Chapter 5, we simulated data sets using

$$\text{eGFR}_{b,i}(\zeta) = \text{eGFR}_i + \sqrt{\zeta} U_{b,i}, \quad i = 1, \ldots, n, \quad b = 1, \ldots, B, \quad (14.13)$$

where $U_{b,i}$ are normal, mean zero, independent random variables with variance σ_u^2, where σ_u^2 is the variance of the measurement error associated with eGFR. We used 10 values for ζ on an equally spaced grid between 0.2 and 2 and $B = 50$ simulated data sets for each value of ζ. The entire program was implemented in R and run in approximately 5 minutes on a PC (3.6GHz CPU, 3.6Gb RAM), with more than 99% of the computation time being dedicated to fitting the 500 Cox models, each with $15,080$ observations.

Models were fit by replacing $\text{eGFR}_{b,i}(\zeta)$ for eGFR_i in model (14.13).

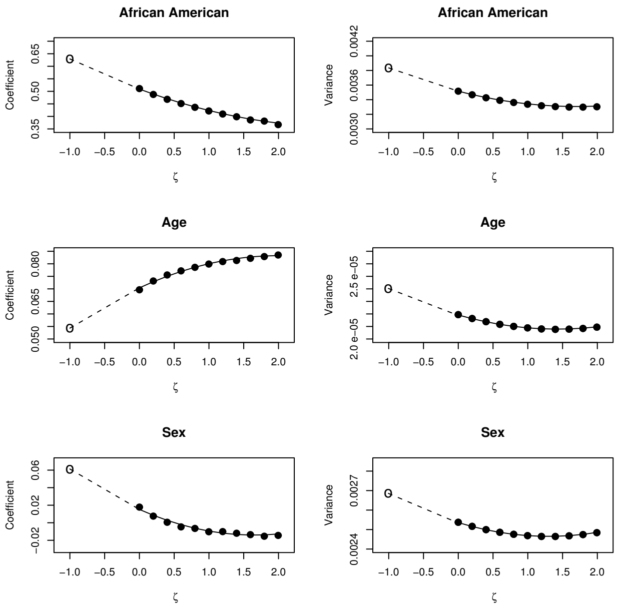


Figure 14.2 *Coefficient and variance extrapolation curves for the ARIC survival modeling. The simulated estimates are based on 50 simulated data sets and are plotted as solid circles. The fitted quadratic extrapolant (solid line) is extrapolated to $\zeta = -1$ (dashed line), resulting in the SIMEX estimate (open circle).*

If $\widehat{\beta}_k^b(\zeta)$, $k = 2, 4$ are the parameter estimates, then

$$\widehat{\beta}_k(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\beta}_k^b(\zeta), k = 2, 3, 4$$

are the estimated effects for noise level $(1+\zeta)\sigma_u^2$. Figure 14.2 displays $\widehat{\beta}_k(\zeta)$ in the left column as filled black circles. The parameter estimates are obtained using a quadratic extrapolant evaluated at $\zeta = -1$, which corresponds to zero measurement error variance. A similar method is applied for the variance of the parameter estimates presented in the right column. The only difference is that we extrapolate separately the sampling and the measurement error variability, as described in Section B.4.1. Table 14.2 provides a comparison between the naive and SIMEX estimates showing that ignoring measurement error would artificially decrease the effect of African-American race by 21% and increase the effect of age by 30%. The effect of sex on progression to primary CKD was not statistically significant under either the naive or the SIMEX procedure.

To obtain the SIMEX estimator of the eGFR effect, we estimated the function $m_{\mathbf{Y}}(\cdot)$ on an equally spaced grid of points x_g, $g = 1, \ldots, G = 100$, between the minimum and maximum observed eGFR. For each level of added noise, $\zeta\sigma_u^2$, the SIMEX estimator at each grid point, x_g, is

$$\widehat{m}_{\mathbf{Y}}(x_g, \zeta) = \frac{1}{B}\sum_{b=1}^{B}\widehat{m}_{\mathbf{Y}}^b(x_g, \zeta),$$

where $\widehat{m}_{\mathbf{Y}}^b(x_g, \zeta)$ is the estimated function at x_g using the b^{th} simulated data set obtained as in (14.13) at the noise level $(1+\zeta)\sigma_u^2$. For every grid point we then used a quadratic linear extrapolant to obtain the SIMEX estimator $\widehat{m}_{\mathbf{Y}}(x_g, \zeta = -1)$. The solid lines in Figure 14.3 represent the estimated function $m_{\mathbf{Y}}(\cdot)$, $\widehat{m}_{\mathbf{Y}}(x_g, \zeta)$, for $\zeta = 0, 0.4, 0.8, 1.2, 1.6, 2$, with higher values of noise corresponding to higher intercepts and less shape definition. The bottom dashed line is the SIMEX estimated curve.

The nonmonotonic shape of all curves is clear in Figure 14.3, with unexpected estimated increase in CKD hazard for very large values of eGFR. Such results should be interpreted cautiously for two reasons. First, the apparent increase may not be statistically significant, since there are only 30 CKD cases with baseline eGFR > 140 and 14 with baseline eGFR > 150. The total number of CKD cases in our data set was $1,605$. Second, eGFR is not a direct measure of the kidney function and is typically obtained from a prediction equation, with creatinine as an important predictor. Creatinine is produced by muscles and is filtered out of the body by the kidney. Thus, lower values of creatinine typically predict better kidney function. However, very low values of creatinine, which would predict very large values of eGFR, could also be due to lack of muscular mass which, in turn, is associated with higher CKD incidence. In short, very low values of creatinine may occur either because the kidney does amazing filtration work or because the subject already has other serious problems and lacks muscular mass. The latter mechanism may actually be the one that is providing the increasing pattern corresponding to eGFR > 140, irrespective of its statistical significance.

14.6 Semi and Nonparametric Methods

Semiparametric models usually refer to a combination of parametric, richly parameterized, and nonparametric models. Survival models with covariate measurement error have three main components that may be modeled semi or nonparametrically:

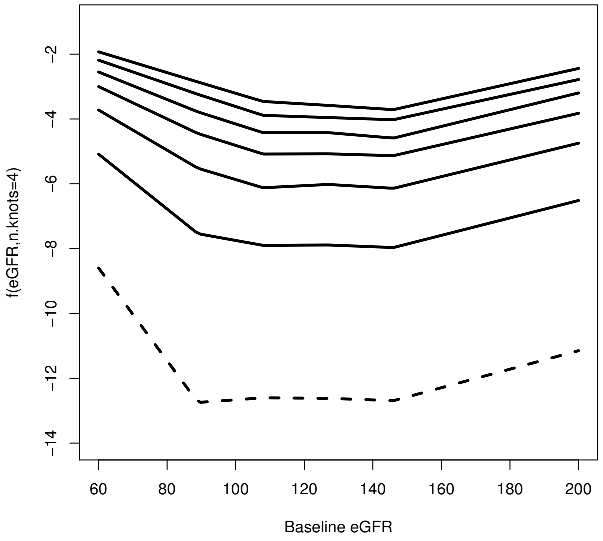


Figure 14.3 *Linear spline fits with* $K = 4$ *knots. Function estimators based on* 50 *simulated data sets corresponding to* $\zeta = 0, 0.4, 0.8, 1.2, 1.6, 2$ *are plotted as solid lines, with larger values of noise corresponding to higher intercepts. The SIMEX estimate is the dashed line.*

1. The conditional expectation $E\{\exp(\beta_x^t \mathbf{X} | \mathbf{T} \geq t, \mathbf{Z}, \mathbf{W})\}$.
2. The distribution function of $\mathbf{X}$.
3. The baseline hazard function $\lambda_0(t)$.

Various parametric and nonparametric methods depend on which component or combination of components is modeled nonparametrically, as well as on the choice between partial or full likelihood function.

14.6.1 Nonparametric Estimation with Validation Data

When the rare-failure assumption introduced in Section 14.2 does not hold, approximating $E\{\exp(\beta_x^t \mathbf{X} | \mathbf{T} \geq t, \mathbf{Z}, \mathbf{W})\}$ by $E\{\exp(\beta_x^t \mathbf{X} | \mathbf{Z}, \mathbf{W})\}$ may lead to seriously biased estimates (Hughes, 1993). One way to avoid this problem is to estimate $E\{\exp(\beta_x^t \mathbf{X} | \mathbf{T} \geq t, \mathbf{Z}, \mathbf{W})\}$ nonparametrically.

Assuming the existence of a validation sample, Zhou and Pepe (1995) proposed a nonparametric estimator of $E\{\exp(\beta_x^t \mathbf{X} | \mathbf{T} \geq t, \mathbf{Z}, \mathbf{W})\}$ when $(\mathbf{Z}, \mathbf{W})$ are categorical. If V and $\bar{V}$ are the sets of indices corresponding to validation and nonvalidation data, respectively, then

$$\widehat{e}_i(t|\beta_x) = \frac{\sum_{j \in V} Y_j(t) I\{\mathbf{Z}_j = \mathbf{Z}_i, \mathbf{W}_j = \mathbf{W}_i\} \exp(\beta_x^t \mathbf{X}_j)}{\sum_{j \in V} Y_j(t) I\{\mathbf{Z}_j = \mathbf{Z}_i, \mathbf{W}_j = \mathbf{W}_i\}} \qquad (14.14)$$

is a simple nonparametric estimator of $E\{\exp(\beta_x^t \mathbf{X}_i | \mathbf{T}_i \geq t, \mathbf{Z}_i, \mathbf{W}_i)\}$. The estimator $\widehat{e}_i(t)$ is easy to calculate and represents the average of $\exp(\beta_x^t \mathbf{X}_j)$ over those subjects in the validation data that are still at risk at time t and share the same observed covariate values $(\mathbf{Z}_j, \mathbf{W}_j)$ with the i^{th} subject. The induced hazard function with the partial likelihood replaced by an estimator is

$$\widehat{\lambda}(t|\mathbf{W}_i, \mathbf{Z}_i) = \lambda_0(t) \exp(\beta_z^t \mathbf{Z}_i) \{\exp(\beta_x \mathbf{X}_i) I(i \in V) + \widehat{e}_i(t|\beta_x) I(i \notin V)\}.$$

Zhou and Pepe (1995) suggested maximizing the following estimator of the log partial likelihood:

$$\mathrm{EPL}(\beta_x, \beta_z) = \sum_{i=1}^{n} \delta_i \left[\log\{\widehat{H}_i(\beta_x, \beta_z)\} - \log\{\sum_{j \in R_i} \widehat{H}_j(\beta_x, \beta_z)\} \right], \qquad (14.15)$$

where $\widehat{H}_i(\beta_x, \beta_z) = \exp(\beta_z^t \mathbf{Z}_i) \{\exp(\beta_x \mathbf{X}_i) I(i \in V) + \widehat{e}_i(\mathbf{Y}_i|\beta_x) I(i \notin V)\}$ is the estimated relative risk of subject i. Because the $\widehat{H}_i(\beta_x, \beta_z)$ is a weighted average of hazard ratios of subjects in the validation sample, standard Cox regression software cannot be used to maximize (14.15). One possible solution would be to maximize (14.15) directly using nonlinear optimization software.

A more serious limitation of the procedure occurs when the conditional distribution $[\mathbf{X}|\mathbf{Z}, \mathbf{W}]$ depends on three or more discrete covariates. In this situation, it is difficult to estimate the conditional distribution $[\mathbf{X}|\mathbf{Z}, \mathbf{W}]$ well from a validation sample that is usually small. This could have severe effects on the variability of the parameter estimate. Also, in practice, the conditional distribution $[\mathbf{X}|\mathbf{Z}, \mathbf{W}]$ often depends on continuous covariates.

A similar nonparametric estimator using kernel smoothing was proposed by Zhou and Wang (2000) when $(\mathbf{Z}, \mathbf{W})$ are continuous. If $K(\cdot)$ is a multivariate kernel function, then the conditional relative risk $\widehat{H}_i(t|\beta_x, \beta_z)$ $= E\{\exp(\beta_x^t \mathbf{X}_i + \beta_z^t \mathbf{Z}_i | \mathbf{T}_i \geq t, \mathbf{Z}_i, \mathbf{W}_i)\}$ can be estimated as

$$\widehat{H}_i(t|\beta_x, \beta_z) = \frac{\sum_{j \in V} Y_j(t) K_h\left\{(\mathbf{Z}_i^t, \mathbf{W}_i^t)^t - (\mathbf{Z}_j^t, \mathbf{W}_j^t)^t\right\} \exp(\beta_x^t \mathbf{X}_j + \beta_z^t \mathbf{Z}_j)}{\sum_{j \in V} Y_j(t) K_h\left\{(\mathbf{Z}_i^t, \mathbf{W}_i^t)^t - (\mathbf{Z}_j^t, \mathbf{W}_j^t)^t\right\}}, \qquad (14.16)$$

where $K_h(\cdot) = K(\cdot/h)$ and h is the kernel bandwidth size. Therefore, for a subject i that is not in the validation data, that is, $i \notin V$, the estimated hazard ratio $\widehat{H}_i(t|\beta_x,\beta_z)$ is a weighted sum of all hazard ratios $\widehat{H}_j(t|\beta_x,\beta_z)$ of subjects in the validation data that are still at risk at time t. The weights assigned to the hazard ratio of each subject depend on the distance between the observed covariates $(\mathbf{Z}_i^t,\mathbf{W}_i^t)^t$ for subject i and $(\mathbf{Z}_j^t,\mathbf{W}_j^t)^t$ in the validation data.

As in the case of discrete covariates, maximizing the function EPL(β_x, β_z) from equation (14.15) can be used to estimate (β_x,β_z). Maximizing EPL(β_x,β_z) requires a separate nonlinear maximization algorithm. However, writing the code should be straightforward because derivatives of $\widehat{H}_i(\beta_x,\beta_z)$ can be calculated as weighted sums of the derivatives of $H_j(\beta_x,\beta_z)$, $j\in V$, with the weights being calculated only once.

The main problem is that the estimators of the induced hazard function will depend heavily on the bandwidth size h of the kernel function. Since typical conditions for asymptotic consistency provide no information about the choice of bandwidth for finite samples, careful data dependent tuning is usually necessary. This problem is especially difficult when there two or more covariates $(\mathbf{Z},\mathbf{W})$, because the typically small validation data would be used for tuning of the smoothing parameters of a multivariate kernel. Another limitation of the methods in this section is that they require a validation sample that, in many applications, is not available.

14.6.2 Nonparametric Estimation with Replicated Data

Huang and Wang (2000) have proposed a nonparametric approach for the case when replicated proxy observations are available for each subject. They assumed that for each variable prone to measurement error, $\mathbf{X}_i$, there are at least two proxy measurements, $\mathbf{W}_{i1},\mathbf{W}_{i2}$ linked to $\mathbf{X}_i$ through a classical additive measurement error model

$$\mathbf{W}_{ij}=\mathbf{X}_i+\mathbf{U}_{ij},\ j=1,2,$$

where $\mathbf{U}_{ij}$ are mutually independent and independent of all other variables. The approach is nonparametric because it does not require specification of the baseline hazard function, λ_0, or the distribution of $\mathbf{U}_{ij}$. However, the proportional hazard function is modeled parametrically.

If $\mathbf{X}_i$ were observed without error, then a consistent, asymptotically normal estimator of (β_x^t,β_z^t) can be obtained by solving the score equation

$$\frac{\partial l(\beta_x,\beta_z)}{\partial\beta_x\partial\beta_z}=0,$$

where $l(\beta_x,\beta_z)$ is the log partial likelihood function defined in (14.4).

The score function is

$$\begin{aligned}\tfrac{\partial l(\beta_x,\beta_z)}{\partial\beta_x\partial\beta_z} \quad &= \quad \textstyle\sum_{i=1}^n \delta_i\Bigg\{[\mathbf{X}_i,\mathbf{Z}_i] \\ &\qquad - \tfrac{\sum_{j\in R_i}[\mathbf{X}_j,\mathbf{Z}_j]\exp(\beta_x^t\mathbf{X}_j+\beta_z^t\mathbf{Z}_j)}{\sum_{j\in R_i}\exp(\beta_x^t\mathbf{X}_j+\beta_z^t\mathbf{Z}_j)}\Bigg\},\end{aligned} \tag{14.17}$$

where $[\mathbf{X},\mathbf{Z}]$ denotes the matrix obtained by binding the columns of $\mathbf{X}$ and $\mathbf{Z}$. The naive approach would replace $\mathbf{X}_i$ with $(\mathbf{W}_{i1}+\mathbf{W}_{i2})/2$, while the regression calibration would replace $\mathbf{X}_i$ with $E[\mathbf{X}_i|\mathbf{W}_{i1},\mathbf{W}_{i2},\mathbf{Z}_i]$. Huang and Wang (2000) proposed replacing the score function (14.17) with

$$\frac{\partial\widetilde{l}(\beta_x,\beta_z)}{\partial\beta_x\partial\beta_z}=\sum_{i=1}^n\delta_i\left\{A_i-\frac{\sum_{j\in R_i}B_j\left(\beta_x,\beta_z\right)}{\sum_{j\in R_i}C_j\left(\beta_x,\beta_z\right)}\right\}, \tag{14.18}$$

where

$$A_i=\frac{[\mathbf{W}_{i1},\mathbf{Z}_i]+[\mathbf{W}_{i1},\mathbf{Z}_i]}{2},$$

$$B_j\left(\beta_x,\beta_z\right)=\frac{[\mathbf{W}_{i1},\mathbf{Z}_j]\exp(\beta_x^t\mathbf{W}_{i2})+[\mathbf{W}_{i2},\mathbf{Z}_j]\exp(\beta_x^t\mathbf{W}_{i1})}{2}\exp(\beta_z^t\mathbf{Z}_j)$$

and

$$C_j\left(\beta_x,\beta_z\right)=\frac{\exp(\beta_x^t\mathbf{W}_{i1})+\exp(\beta_x^t\mathbf{W}_{i1})}{2}\exp(\beta_z^t\mathbf{Z}_j).$$

Huang and Wang (2000) showed that the estimator obtained by maximizing (14.18) is consistent and asymptotically normal. When more than two replicates are available the formulas for A_i, $B_j\left(\beta_x,\beta_z\right)$ and $C_j\left(\beta_x,\beta_z\right)$ are slightly more complicated by taking averages over all replicates instead of just two.

A potential problem with this approach is that the approximate score function (14.18) can be evaluated only for those subjects i that have repeated measurements. Therefore, serious losses of efficiency may occur when replication data are available for only a small subsample. Biased estimators may occur when the subset of subjects with replicated data is not a random subsample of the original population sample.

14.6.3 Likelihood Estimation

Likelihood estimation is a different approach to estimation in the context of survival analysis with measurement error. Under the assumptions in Section 14.1, the likelihood associated with one observation, i, in the

Cox model is

$$\int \{\lambda(\mathbf{Y}_i|x,\mathbf{Z}_i)\}^{\delta_i}\exp\left\{-\int_0^{Y_i}\lambda(u|x,\mathbf{Z}_i)du\right\}f(\mathbf{W}_i,x|\mathbf{Z}_i)dx, \quad (14.19)$$

where $f(w,x|z)=f(w|x,z)f(x)$ is the joint conditional density function of the random variables $\mathbf{W}$ and $\mathbf{X}$ given $\mathbf{Z}$. It is assumed that $\mathbf{X}$ is independent of $\mathbf{Z}$. If $t_1,\ldots,t_m$ are all the unique failure times then the full likelihood function can be written as

$$\begin{aligned}\mathcal{L} \;=\; & \prod_{i=1}^n\Bigg[\int\prod_{j=1}^m\lambda_0(t_j)^{I(\mathbf{Y}_i=t_j)}\exp\{\delta_i(\beta_x^t x+\beta_z^t\mathbf{Z}_i)\}\\ & \times\exp\left[-\exp\{\delta_i(\beta_x^t x+\beta_z^t\mathbf{Z}_i)\}\sum_{j=1}^m\lambda_0(t_j)I(t_j\le\mathbf{Y}_i)\right]\\ & \times f(\mathbf{W}_i|x,\mathbf{Z}_i)f(x|\theta)dx\Bigg],\end{aligned} \quad (14.20)$$

where $\beta_x,\beta_z,\lambda_0(t_1),\ldots,\lambda_0(t_m),\theta$ are treated as unknown parameters. Hu, Tsiatis, and Davidian (1998) assumed that the conditional density $f(w|x,z)$ is known and used parametric, semi and nonparametric models for $f(x|\theta)$.

The parametric model assumes that $\mathbf{X}$ has a normal distribution with parameters $\theta=(\mu_x,\Sigma_x)$, which in many applications is a reasonable assumption. When this assumption is not reasonable one can often find a transformation of the observed proxies, $\mathbf{W}$, that would be consistent with the normality assumption. While Hu, Tsiatis, and Davidian (1998) called this a fully parametric method, the baseline hazard function is not parameterized. Thus, the procedure requires maximization over a large number of parameters and could be considered nonparametric with respect to the hazard function.

An important feature of this methodology is that existent software developed for fitting nonlinear mixed effect models, such as the FORTRAN program `Nlmix` (Davidian and Gallant, 1993) or the R function `nlme`, can be adapted to maximize (14.20). This can be done by treating the unobserved variables $\mathbf{X}_i$ as independent normal random effects and

using the following conditional distribution

$$\begin{aligned}&\prod_{j=1}^m\lambda_0(t_j)^{I(\mathbf{Y}_i=t_j)}\exp\{\delta_i(\beta_x^t\mathbf{X}_i+\beta_z^t\mathbf{Z}_i)\}\\ &\times\exp\left[-\exp\{\delta_i(\beta_x^t\mathbf{X}_i+\beta_z^t\mathbf{Z}_i)\}\sum_{j=1}^m\lambda_0(t_j)I(t_j\le\mathbf{Y}_i)\right]\\ &\times f(\mathbf{W}_i|\mathbf{X}_i,\mathbf{Z}_i).\end{aligned}$$

These functions work well when the integral in equation (14.20) is low dimensional, that is, when the number of variables subject to measurement error is small.

The normality assumption can be further relaxed by considering a more general family of distributions for the unobserved variables $\mathbf{X}_i$. For the case when only one variable is observed with error, Hu et al. (1998) used the semi-nonparametric (SNP) class of smooth densities of Gallant and Nychka (1987),

$$f(x|\theta)=\frac{1}{C(\theta)}\left(1+a_1x+\ldots+a_Kx^K\right)^2\frac{1}{\sigma_x}\exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}, \quad (14.21)$$

where $\theta=(a_1,\ldots,a_K,\mu_x,\sigma_x)$ and $C(\theta)$ is a constant that ensures $\int f(x|\theta)dx=1$. The class of smooth densities (14.21) contains the normal densities as a particular case when $a_1=\ldots=a_K=0$. Because the number of monomials, K, is unknown and can theoretically be very large, the family of distributions (14.21) can be viewed as a nonparametric family. However, in practice, it is very rare that $K\ge 3$ is necessary. When K is small, maximizing (14.20) when $f(x|\theta)$ has the form (14.21) could provide a useful sensitivity analysis to the specification of the marginal distribution of $\mathbf{X}$. Alternatively, the robustness of the specification can be checked using the remeasurement method of Huang, Stefanski, and Davidian (2006) (see also Section 5.6.3). The FORTRAN program `Nlmix` (Davidian and Gallant, 1993) can handle random effect distributions of the type (14.21).

The fully nonparametric approach of Hu et al. (1998) for modeling $f(x|\theta)$ uses a binning strategy similar to histogram estimation. More precisely, a set of locations $x_1,\ldots,x_K$ is fixed, where $K<<n$ is the number of support points of the approximate distribution. The probability mass function of $\mathbf{X}$ is represented as

$$f(x|\theta)=\prod_{k=1}^K p_k^{I(X=x_k)}, \quad (14.22)$$

where $\theta=(K,x_1,\ldots,x_K,p_1,\ldots,p_K)$, $p_k=P(X=x_k)$, $k=1,\ldots,K$ and $\sum_{k=1}^K p_k=1$. While, in principle, one could maximize the likelihood over θ, $\lambda_0(t_1),\ldots,\lambda_0(t_m)$, β_x and β_z this is a not a realistic approach.

Hu, Tsiatis, and Davidian (1998) fixed K to be moderately large ($K = 20$) and $x_1, \ldots, x_K$ equally spaced on the range of observed values of $\mathbf{W}$. For $\theta = (p_1, \ldots, p_K)$, Hu et al. proposed an EM algorithm to maximize (14.20) and provided a simulation study comparing these methods with regression calibration.

Somewhat unexpectedly, regression calibration performs remarkably well even in small samples ($n = 100$) when the distribution of $\mathbf{X}$ is normal and the attenuation factor is moderate or small. The full likelihood analysis using the normal distribution performed well even when the distribution of $\mathbf{X}$ was not normal. From a practical perspective, applying normalizing transformations to the observed $\mathbf{W}$ and using regression calibration may be a very good first step of the analysis. As discussed by Hu et al. (1998), applying a likelihood-based method may be computationally prohibitive for realistic data sets. A reasonable alternative could be to apply these methods to a random subsample of the data as a sensitivity analysis.

One limitation of the methods described in this section is that the distribution of $\mathbf{X}$ is not allowed to depend on observed covariates $\mathbf{Z}$. Another limitation is that the methods are designed for one variable subject to measurement error, and they do not easily generalize to multiple correlated variables. Lastly, the computational burden seems prohibitive for data sets with thousands of observations and multiple covariates.

14.7 Likelihood Inference for Frailty Models

Random effects models have been discussed in Chapter 11. Random effects have also been used in survival analysis with clustered data, but in this context they are called *frailties*. In this section, we will use the notation introduced in Section 14.1 but use a pair of indices (i, j) instead of the single index i, where $i = 1, \ldots, I$ is the cluster index, and $j = 1, \ldots, J$ is the observation index within cluster.

Conditional on the cluster-specific frailty, the proportional hazards function follows a Cox model (14.3):

$$\lambda_{ij}\left(t|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, b_i\right) = \lambda_0(t) \exp\left(\beta_x^t \mathbf{X}_{ij} + \beta_z^t \mathbf{Z}_{ij} + b_i\right), \qquad (14.23)$$

where $\mathbf{X}_{ij}$ are variables subject to measurement error, $\mathbf{Z}_{ij}$ are observed without measurement error, and b_i is the cluster specific frailty. In addition to the standard assumptions for survival analysis we will assume that b_i are iid Normal$(0, \sigma_b^2)$, independent of failure, censoring, and measurement error processes. We assume that the proxy variables $\mathbf{W}_{ij}$ follow a classical additive measurement error model

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij},$$

where $\mathbf{U}_{ij}$ are mean zero measurement error variables, independent of $\mathbf{T}_{ij}$, $\mathbf{C}_{ij}$, $\mathbf{X}_{ij}$ and b_i.

A full likelihood approach was proposed by Li and Lin (2000) for fitting model (14.23), assuming normality of the frailty and $[\mathbf{X}|\mathbf{W}, \mathbf{Z}]$ distributions. They used an EM algorithm to maximize the marginal likelihood by treating the frailties and the covariates observed with error as missing data. The "complete data" for the i^{th} cluster are the observed data $(\mathbf{Z}_{ij}, \mathbf{W}_{ij})$ and the unobserved $(\mathbf{X}_{ij}, b_i)$. The complete data likelihood for this cluster is

$$\mathcal{L}_i(\Theta; \mathbf{X}_{ij}, b_i, \mathbf{Z}_{ij}, \mathbf{W}_{ij}) = \{\lambda_{ij}(t|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, b_i)\}^{\delta_{ij}} \times$$

$$\exp\left\{-\int_0^{Y_i} \lambda(u|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, b_i)du\right\} \phi(b_i, \sigma_b^2)\phi(\mathbf{X}_{ij}\,|\mathbf{W}_{ij}, \mathbf{Z}_{ij}, \theta), \quad (14.24)$$

where $\phi(b_i, \sigma_b^2)$ is the normal density of b_i with mean zero and variance σ_b^2 and $\phi(\mathbf{X}_{ij}|\mathbf{W}_{ij}, \mathbf{Z}_{ij})$ is the conditional normal density of $\mathbf{X}_{ij}$ given $(\mathbf{W}_{ij}, \mathbf{Z}_{ij})$. Here Θ is the vector of all parameters and includes the parameters of the proportional hazard function, (β_x, β_z); the parameter of the random effects model, σ_b^2; the parameter of the conditional distribution $[\mathbf{X}|\mathbf{W}, \mathbf{Z}]$, θ; and the set of all jumps in the integrated baseline hazard function, $\Delta\Lambda_0(t)$.

Of course, the complete data likelihood cannot be used directly for estimation since it contains unobserved data. Instead, one uses the marginal likelihood of the observed data, which is obtained by integrating out the unobserved quantities, that is

$$\mathcal{L}(\Theta) = \prod_{i=1}^{I} \int \left\{\prod_{j=1}^{J} \mathcal{L}_i(\Theta, ; \mathbf{X}_{ij}, b_i, \mathbf{Z}_{ij}, \mathbf{W}_{ij})d\mathbf{X}_{ij}\right\} db_i.$$

The EM algorithm of Li and Lin (2000) usesd Monte Carlo simulations to perform these integrations at the E-step. A full likelihood analysis requires an estimate of the baseline hazard, and Li and Lin used a nonparametric maximum likelihood estimator.

Bibliographic Notes

Substantial methodological and applied research has been dedicated in recent years to survival analysis with covariates subject to measurement error, starting with the seminal paper by Prentice (1982). The regression calibration approach was expanded and refined by Pepe, Self, and Prentice (1989) and Wang, Hsu, Feng, and Prentice (1997). Clayton (1991) used regression calibration within risk sets, thus avoiding the rare disease assumption. For data containing a validation sample, Zhou and Pepe

(1995) and Zhou and Wang (2000) proposed nonparametric estimators of the induced hazard function. For data with at least two replicates, Huang and Wang (2000) proposed a consistent nonparametric estimator based on a modification of the partial likelihood score equation. Augustin (2004) showed that Nakamura's (1992) methodology of adjusting the likelihood can be applied to the Breslow likelihood to provide an exact corrected likelihood. This result circumvented the impossibility result derived by Stefanski (1989) for the partial likelihood. Hu, Tsiatis, and Davidian (1998) have proposed likelihood maximization algorithms for parametric and nonparametric specifications of the distribution of the unobserved variables. Greene and Cai (2004) established the asymptotic properties of the SIMEX estimators for models with measurement error and multivariate failure time data. Hu and Lin (2004) introduced a modified score equation and established the asymptotic properties of the estimators for multivariate failure time data. Li and Lin (2000, 2003a) used the EM algorithm and SIMEX, respectively, to provide maximum likelihood estimators for frailty models with variables observed with error. Song and Huang (2005) compared the conditional score estimation of Tsiatis and Davidian (2001) with Nakamura's (1992) parametric adjustment. Tadesse, Ibrahim, Gentleman, et al. (2005) discussed the Bayesian analysis of times to remission using as covariates gene expression levels measured by microarrays.

Surrogate markers are outcomes that are correlated with the outcome of primary interest, for example CD4 counts have been used as a surrogate marker for survival times in AIDS clinical trials (Dafni and Tsiatis, 1998). The advantage of a surrogate marker is that it can indicate relatively rapidly whether a treatment is effective, for example an AIDS treatment can be judged effective relatively quickly if it causes a significant increase in CD4 counts, whereas effectiveness based on survival times might not be evident until enough deaths have occurred, which might take years. In the meantime, patients would be deprived of a new and effective treatment. A surrogate marker is an outcome, that is a variable depending upon treatment, as well as a covariate for predicting the primary outcome. Dafni and Tsiatis (1998) discussed methodology for handling surrogate outcomes measured with error.

CHAPTER 15

RESPONSE VARIABLE ERROR

15.1 Response Error and Linear Regression

In preceding chapters, we have focused primarily on problems associated with measurement error in predictor variables. In this chapter, we consider problems that arise when a true response is measured with error. Since in previous chapters we have designated **X** as a covariate measured with error, to emphasize that we are not combining response error and covariate error, we will not use **X** in this chapter.

Abrevaya and Hausman (2004) state "Classical measurement error (that is, additive error uncorrelated with the covariates) in the dependent variable is generally ignored in regression analysis because it simply gets absorbed into the error residual." It is interesting to consider this claim.

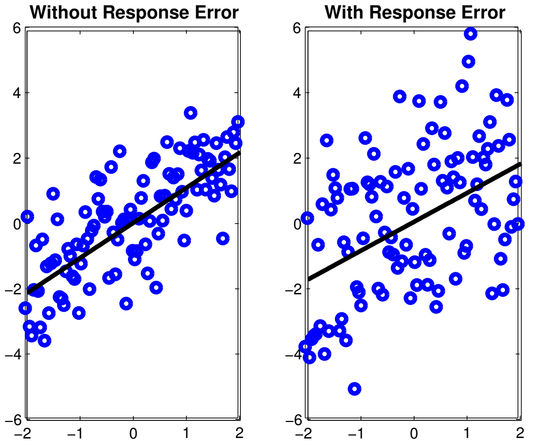


Figure 15.1 *An illustration of response error in linear regression with unbiased classical measurement error. The solid line is the data without response measurement error, while the dashed line is the observed data with response measurement error. Note the increased variability about the line when there is response error.*

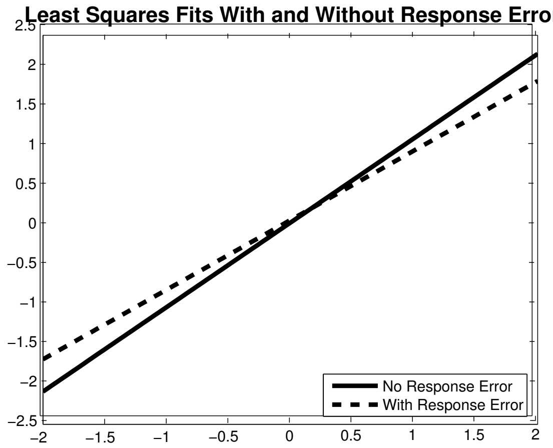


Figure 15.2 *The fitted least squares lines in the data from Figure 15.1. The left panel is the fitted least squares without response measurement error, while the right panel is the fitted least squares line with response measurement error. Note the lack of bias due to this type of response measurement error.*

We generated linear regression data so that the true **Z** values were equally spaced on the interval $[-2, 2]$, the intercept $\beta_0 = 0.0$, and the slope $\beta_z = 1$. The error about the line was $\sigma_\epsilon^2 = 1$, so that $\mathbf{Y} = \text{Normal}(\mathbf{Z}, 1)$. We then added to **Y** normally distributed response error with variance $\sigma_v^2 = 3$, that is we observe $\mathbf{S} = \mathbf{Y} + \mathbf{V}$, where $\mathbf{V} = \text{Normal}(0, 3.0)$. Note that the measurement error in the response is *three times* the error about the line. If the measurement error were this large and had been in the predictors, we know that the effect on the fitted lines would be enormous.

What happens, though, when the measurement error is in the response? The remark of Abrevaya and Hausman is illustrated in Figures 15.1, 15.2, and 15.3. In Figure 15.1, we show a typical set of data generated with and without response error. The obvious feature we see here is that the unbiased measurement error in the response increases the variability of the observe data about the line.

Figure 15.2 delves a little deeper, displaying the actual fitted lines. The remarkable thing here is that, even though the data with response measurement error have four times the variability about the line as the data without response measurement error, the two lines are very similar. Figure 15.3 is the results of 200 simulated data sets, showing that unbiased response measurement error simply increases the variability of the fitted lines.

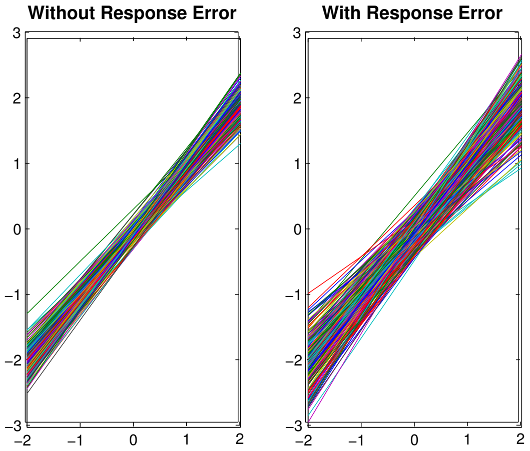


Figure 15.3 *Two hundred simulated data sets without (left panel) and with (right panel) measurement error in the setup of Figure 15.1. Note that response measurement error that is classical and unbiased simply increases the variability of the least squares fitted lines, without affecting bias.*

We now make the following conclusions, the first of which is supported by the exercise with simulated data that we have just undertaken:

- In linear regression with unbiased and homoscedastic response measurement error, the response measurement error increases the variability of the fitted lines without causing bias.
- We now go on to make a far stronger conclusion. In linear or nonlinear regression that has homoscedastic errors about the true line, the only effects of adding unbiased, homoscedastic response measurement error is to increase the variability of the fitted lines and surfaces, and to decrease power for detecting effects. All tests, confidence intervals, etc. are perfectly valid: they are simply less powerful.

The argument for the last very strong statement is perfectly simple. Suppose that without response error, **Y** has mean $m_{\mathbf{Y}}(\mathbf{Z}, \mathcal{B})$ and variance σ^2. Now suppose that we observe **S**, which is just **Y** with additive error σ_v^2. Then the observe response **S** has mean $m_{\mathbf{Y}}(\mathbf{Z}, \mathcal{B})$ and variance $\sigma_{\text{new}}^2 = \sigma^2 + \sigma_v^2$. Thus, the observed data have the same mean and a constant, but larger variance.

There is one caveat. For strongly nonlinear models, the larger response variance has further implications. Inference for nonlinear models is often

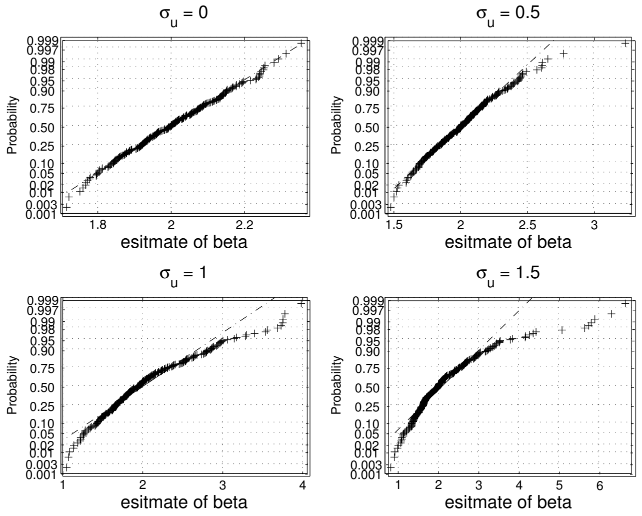


Figure 15.4 *Normal plots of $\widehat{\beta}$ for an exponential regression model with different amounts of measurement error in the response.*

based on approximation of the model by a linear one using a Taylor expansion of the parameter, β, about its true value, β_0, for example

$$\mathbf{Y}_i = m_{\mathbf{Y}}(\mathbf{Z}_i, \beta) + \epsilon_i \approx m_{\mathbf{Y}}(\mathbf{Z}_i, \beta_0) + f'(\mathbf{Z}_i, \beta_0)(\beta - \beta_0) + \epsilon_i.$$

The error in the Taylor approximation decreases to zero as β approaches β_0.

An increase in response variance causes $\widehat{\beta}$ to vary more about β_0, which makes the approximation less accurate. This can be seen in Figure 15.4, which has normal plots of $\widehat{\beta}$ from 250 simulations of the model

$$\mathbf{Y}_i = \exp(-\beta \mathbf{Z}_i) + \epsilon_i + \mathbf{U}_i, \; i = 1, \ldots, 200,$$

where $Z_1 \ldots, Z_{200}$ are equally spaced on [0, 1], $\beta = 2$, $\sigma_\epsilon = 1/4$, and $\sigma_u = 0$, 0.5, 1, and 1.5, respectively, in the four panels starting at the top left. Notice that larger values of σ_u increase not only the variability of $\widehat{\beta}$ but also its skewness. Because the errors are normally distributed, $\widehat{\beta}$ would have an exact normal distribution if the model were linear. Deviation from normality increases with σ_u, because larger values of σ_u increase the effect of the nonlinearity in the model. This problem does not occur, of course, if the model is linear.

15.2 Other Forms of Additive Response Error

15.2.1 Biased Responses

If $\mathbf{S}$ is not unbiased for $\mathbf{Y}$, then regression of it on the observed predictors leads to biased estimates of the main regression parameters. For example, suppose $\mathbf{Y}$ given $\mathbf{Z}$ follows a normal linear model with mean $\beta_0 + \beta_z^t \mathbf{Z}$ and variance σ_ϵ^2, while $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z})$ follows a normal linear model with mean $\gamma_0 + \gamma_1 \mathbf{Y}$ and variance σ_v^2. Here $\mathbf{S}$ is biased, and the observed data follow a normal linear model with mean $\gamma_0 + \beta_0\gamma_1 + \gamma_1\beta_z^t\mathbf{Z}$ and variance $\sigma_v^2 + \gamma_1^2\sigma_\epsilon^2$. Thus, instead of estimating β_z, naive regression ignoring measurement error in the response estimates $\gamma_1\beta_z$.

There is an obvious solution to this problem, namely, to change $\mathbf{S}$ so that it is unbiased, that is use $(\mathbf{S} - \gamma_0)/\gamma_1$. The careful reader will note that when a writer says things are obvious, he/she means something different. Clearly (a better word!), the problem here is to obtain information about (γ_0, γ_1). In a series of papers, Buonaccorsi (1991, 1996) and Buonaccorsi and Tosteson (1993) discussed how to do just this. Here, we give a brief overview of what they proposed.

15.2.1.1 Validation Data

Suppose that validation data are available on a simple random subsample of the primary data. The idea neatly breaks down into a series of steps:

- Use the validation subsample data to obtain estimates of $\mathcal{B}$, the parameters relating $\mathbf{Y}$ and $\mathbf{Z}$, and (γ_0, γ_1): call the former $\widehat{\mathcal{B}}_1$.
- Create an estimated unbiased response as $(\mathbf{S} - \widehat{\gamma}_0)/\widehat{\gamma}_1$ and run your favorite analysis to get a second estimate $\widehat{\mathcal{B}}_2$.
- Estimate the joint covariance matrix of these estimates using the bootstrap, and call it Σ.
- Form the best weighted combination of the two estimates, namely
$$\widehat{\mathcal{B}} = (J^t\Sigma^{-1}J)^{-1}J^t\Sigma^{-1}(\widehat{\mathcal{B}}_1^t, \widehat{\mathcal{B}}_2^t)^t,$$
where $J = (I, I)$ and I is the identity matrix with the same number of rows as there are elements in $\mathcal{B}$.
- Use $(J^t\widehat{\Sigma}^{-1}J)^{-1}$ as the estimated covariance matrix for the combined estimate $\widehat{\mathcal{B}}$.

15.2.1.2 Alloyed Gold Standard

In some (presumably fairly rare) cases, one might not have validation data, but instead, for a random subsample of the primary data, one might have two independent replicate unbiased measurements of $\mathbf{Y}$; call them $(\mathbf{S}_{1*}, \mathbf{S}_{2,*})$. These unbiased replicates are in addition to the biased surrogate $\mathbf{S}$ measured on the main study sample.

In this case, we use the same algorithm as for validation data, with the following changes:

- Use the unbiased response $(\mathbf{S}_{1,*} + \mathbf{S}_{2,*})/2$ to get $\widehat{\mathcal{B}}_1$.
- Estimate (γ_0, γ_1) using measurement error methods (!) as described in Chapter 3, because the replication data follow the model,

$$\begin{aligned} \mathbf{S} &= \gamma_0 + \gamma_1 \mathbf{Y} + \mathbf{V}; \\ \mathbf{S}_{j,*} &= \mathbf{Y} + \mathbf{U}_{j,*} \text{ for } j = 1, 2, \end{aligned}$$

  where $\mathbf{U}_{1,*}$ and $\mathbf{U}_{2,*}$ are independent with mean zero. This is a linear regression measurement error model with response $\mathbf{S}$ and "true covariate" $\mathbf{Y}$ and with replicate measurements $\mathbf{S}_{1,*}$ and $\mathbf{S}_{2,*}$ of $\mathbf{Y}$. The methods reviewed in Chapter 3 are used to estimate (γ_0, γ_1).

15.2.2 Response Error in Heteroscedastic Regression

Weighted least squares and generalized least squares approaches are often used when the data exhibit nonconstant variances, that is are heteroscedastic. We call such models QVF models (Sections 8.8 and A.7), because they combine aspects of quasilikelihood and variance function modeling. Sections 8.8 and A.7 describes what these models are, and how to fit and make inference about them. Of course, we like to think of Carroll and Ruppert (1988) as the authoritative text on the topic.

Briefly, additive unbiased response measurement error in a heteroscedastic regression simply changes the form of the variance function. Once one keeps track of the change, then the methods of Section A.7 apply.

15.2.2.1 A Simple Case

To see this in a simple case, suppose that the regression function one wants to fit to $\mathbf{Y}$ is linear: $\beta_0 + \beta_z \mathbf{Z}$, but that the variance about the true line is $\sigma_\epsilon^2 \mathbf{Z}^\alpha$. If $\mathbf{Y}$ were observed and α were known, one would simply perform a weighted least squares regression with weights $\mathbf{Z}^{-\alpha}$: Section A.7 shows how to estimate α.

Now suppose, however, that instead of observing $\mathbf{Y}$, one observed $\mathbf{S} = \mathbf{Y} + \mathbf{V}$, where $\mathbf{V}$ has mean zero and variance σ_v^2. Then we are simply adding variability, so that $\mathbf{S}$ has the same linear regression function as does $\mathbf{Y}$, but the variance becomes $\sigma_v^2 + \sigma_\epsilon^2 \mathbf{Z}^\alpha$. The form of the variance function has changed by the addition of the response error variance σ_v^2. However, Section A.7 is so general that the methods described there apply to the new model for $\mathbf{S}$.

15.2.2.2 General Case

Luckily, the general case is the simple case, just with more general notation. Now the regression function for $\mathbf{Y}$ is something general like $m_{\mathbf{Y}}(\mathbf{Z}, \mathcal{B})$, and the variance function for $\mathbf{Y}$ is something general like $\sigma_\epsilon^2 g^2(\mathbf{Z}, \mathcal{B}, \theta)$. The rule, though, remains the same: if there is additive, unbiased response error, the regression function remains the same, the variance function changes, and Section A.7 shows how to cope with the change. As before, if we observe $\mathbf{S} = \mathbf{Y} + \mathbf{V}$, where $\mathbf{V}$ has mean zero and variance σ_v^2, then the new variance function is just $\sigma_v^2 + \sigma_\epsilon^2 g^2(\mathbf{Z}, \mathcal{B}, \theta)$.

15.2.2.3 Ignoring Heteroscedasticity

We think it is silly, but many people ignore nonconstant variability and fit unweighted regressions, with the variance for the regression function parameters fixed up by devices like the bootstrap (Section A.9) or the sandwich method (Section A.6).

Why silly?

- Not accounting properly for variability leads to a decrease in efficiency for estimating the regression parameter $\mathcal{B}$. In effect, this means throwing away data just for sport. Few investigators have enough data that they are willing to throw some away to entertain the statistician.
- Not all of statistics is estimating regression parameters. It is often important to understand the variability in order to make inferences about predictions and calibrations. This ihas been demonstrated in striking detail by Carroll (2002) and in a series of examples by Carroll and Ruppert (1988).

15.3 Logistic Regression with Response Error

15.3.1 The Impact of Response Misclassification

In logistic regression, response error is *misclassification.* There are two primary differences with regression models having a continuous response and additive response measurement error:

- Additive measurement error makes no sense. The error occurs when a positive response $\mathbf{Y} = 1$ is transmuted into a negative response $\mathbf{S} = 0$, and vice versa.
- Misclassification is biased response error, and the bias needs to be accounted for.

Thus, consider a logistic regression model that has probability of response

$$\text{pr}(\mathbf{Y} = 1 | \mathbf{Z}) = H(\beta_0 + \beta_z^t \mathbf{Z}), \tag{15.1}$$

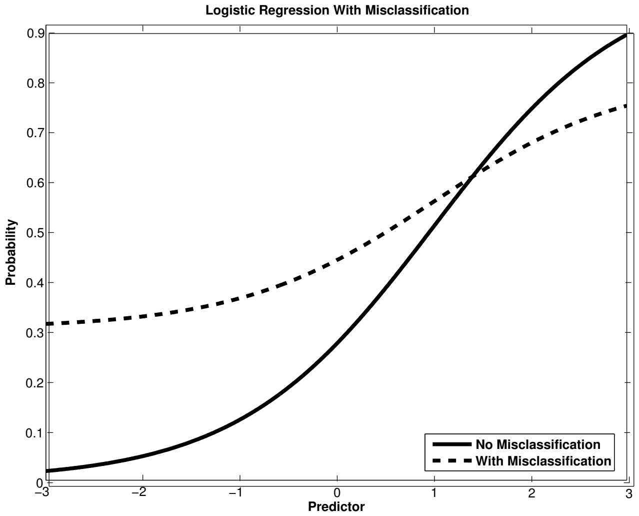


Figure 15.5 *Illustration of the effect of misclassification of the response in logistic regression. Solid line: the true probability of response. Dashed line: the observed probability of response when cases (noncases) are classified incorrectly* 20% *(*30%*) of the time.*

where $H(\cdot)$ is the logistic distribution function. Pretend that misclassification does not depend on $\mathbf{Z}$, and that we classify individuals correctly with probabilities

$$\begin{aligned} \mathrm{pr}(\mathbf{S}=1|\mathbf{Y}=1,\mathbf{Z}) &= \pi_1; \\ \mathrm{pr}(\mathbf{S}=0|\mathbf{Y}=0,\mathbf{Z}) &= \pi_0. \end{aligned}$$

The observed data no longer follow the logistic model (15.1), but instead have the more complex form

$$\mathrm{pr}(\mathbf{S}=1|\mathbf{Z}) = (1-\pi_0) + (\pi_1+\pi_0-1)H(\beta_0+\beta_z^t\mathbf{Z}). \tag{15.2}$$

Figure 15.5 gives an illustration of the impact of misclassification of the response. In this setting, those who actually have a response $\mathbf{Y}=1$ are correctly classified with probability 80%, while those who did not have a response, that is $\mathbf{Y}=0$, are correctly classified with probability 70%. The logistic intercept is -1.0 and the slope is 1.0. One can see that the effect of response misclassification is to bias the true line badly, somewhat along the lines of an attenuation. The difference between the major impact of misclassification of the response here and the near null impact of unbiased response error in linear regression (Figure 15.2) is profound and important.

Also, response misclassification can lead to major biases in parameter estimates. In our case, the true slope is $\beta_z = 1$, but the response misclassification makes logistic regression think that the slope is more along the lines of 0.40, a major difference.

This illustration indicates that response misclassification does need to be accounted for.

15.3.2 Correcting for Response Misclassification

The profound impact of response misclassification in logistic regression has led to the development of many interesting statistical methods, see among many others Palmgren and Ekholm (1982, 1987), Ekholm and Palmgren (1987), Copas (1988), Neuhaus (2002), Ramalho (2002), Prescott and Garthwaite (2002), and Paulino et al. (2003).

15.3.2.1 Unknown Misclassification Probabilities

If one believes the misclassification probabilities are independent of the covariates, then estimating all the parameters $(\pi_1,\pi_0,\beta_0,\beta_z)$ can be done via maximum likelihood or Bayesian approaches. Let the probability model (15.2) be denoted as $\Psi(\mathbf{S},\mathbf{Z},\pi_0,\pi_1,\beta_0,\beta_z)$. Then the loglikelihood function to be maximized is just

$$\begin{aligned} \sum_{i=1}^{n} \big[& \mathbf{S}_i \mathrm{log}\{\Psi(\mathbf{S},\mathbf{Z},\pi_0,\pi_1,\beta_0,\beta_z)\} \\ & + (1-\mathbf{S}_i)\mathrm{log}\{1-\Psi(\mathbf{S},\mathbf{Z},\pi_0,\pi_1,\beta_0,\beta_z)\}\big]. \end{aligned} \tag{15.3}$$

Maximization of this loglikelihood can be done via many devices, including the method of scoring, iteratively reweighted least squares and the EM-algorithm.

The major practical issue is that the classification probabilities are only very weakly identified by the data, that is they are difficult to estimate with any precision, and that difficulty carries over to estimation of the underlying risk function. Copas (1988) states that "accurate estimation of (the misclassification parameters) is very difficult if not impossible unless n is extremely large." This is one of the classic cases where parameters may be identified theoretically but not in any practical sense; see also Section 8.1.2. Copas (1988) and Neuhaus (2002) both basically concluded that the best one can hope to do is a sensitivity analysis for plausible values of the misclassification probabilities.

Paulino et al. (2003), in a slightly different context, addressed the problem of lack of practical identifiability of the misclassification probabilities via the Bayesian route, using informative prior distributions that were developed with the help of a subject-matter expert.

We next describe situations in which there is information about the misclassification probabilities.

15.3.2.2 Known Misclassification Probabilities

In the presumably rare event that the classification probabilities π_1 and π_0 are known, maximizing the loglikelihood (15.3) in (β_0, β_z) is simple using iteratively reweighted least squares.

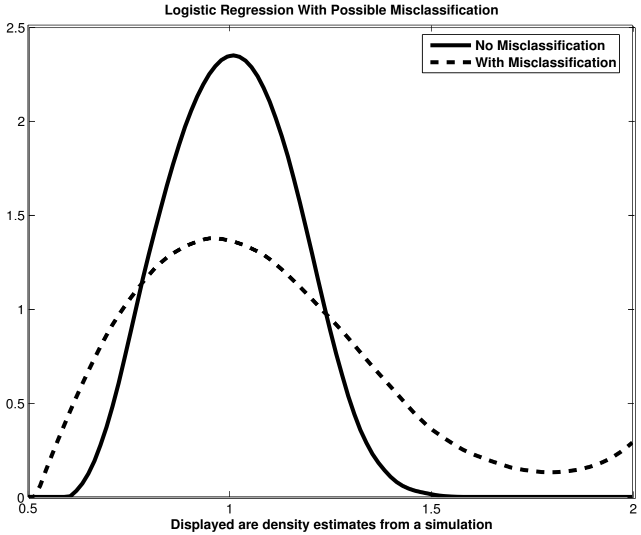


Figure 15.6 *Illustration of the effect of misclassification of the response in logistic regression. Solid line: density estimate of the estimated slopes in a simulation study of the logistic regression when there is no misclassification. Dashed line: density estimate with misclassification, when the misclassification probabilities are known. The observed probability of response with cases (noncases) are classified incorrectly* 20% *(*30%*) of the time. Note the profound loss of information due to response misclassification.*

Figure 15.6 describes a simulation study of the same logistic model described previously when the number of observations is $n = 500$. It contrasts the density function of the logistic regression slope estimator if there were no misclassification (solid line) versus what happens when there is misclassification, but the misclassification probabilities are known. The point of this figure is to note that if the classification probabilities are known, then one can indeed construct an approximately consistent estimate of the true slope (in contrast, if one ignores the misclassification, one thinks that the slope is 0.40, not the correct 1.00), but that the effect of misclassification is to increase greatly the variability of the fitted logistic slope.

15.3.2.3 Validation Data

In some cases, there might be validation data, that is $\mathbf{Y}$ may be observable on a subset of the study. In this case, one can directly estimate the classification probabilities.

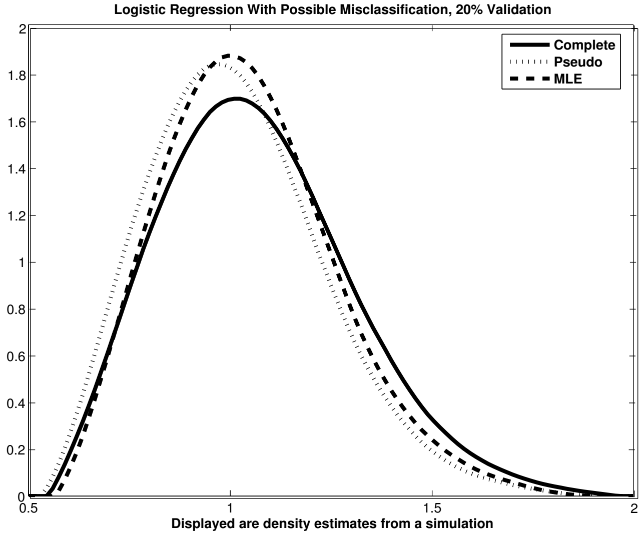


Figure 15.7 *Illustration of the effect of misclassification of the response in logistic regression, when there is* 20% *validation done completely at random. Solid line: density estimate of the slope in the logistic regression for the validation data. Dashed line: density estimate for the MLE. Dotted line: density estimate for pseudolikelihood.*

One possibility is to estimate π_1 (π_0) as the fraction of those in the validation study who are correctly classified among those whose true value is $\mathbf{Y} = 1$ $(\mathbf{Y} = 0)$, pretend that these are known, and then maximize the (now pretend) likelihood (15.3). This approach is called *pseudolikelihood*, a methodology that has a long and honorable history in statistics. There are two major difficulties with such an approach:

- It is invalid, leading to biased estimation and inference, if selection into the validation study depends on the observed values of $\mathbf{S}$ or $\mathbf{Z}$, as might reasonably happen.
- It is inefficient, because in this case a proper likelihood analysis can

be undertaken that uses the observed **Y** values effectively. A detailed derivation of the likelihood function is delayed until Section 15.4 below. See also Prescott and Garthwaite (2002) for a Bayesian treatment.

In Figure 15.7, we display what happens to the complete data estimate, the pseudolikelihood estimate, and the maximum likelihood estimate of the slope when there is 20% randomly selected validation. The complete data estimate, also called the *complete-cases estimate*, uses only the cases that have all variables measured, that is only the validation data. All the methods are consistent estimates, and there is little to choose between them.

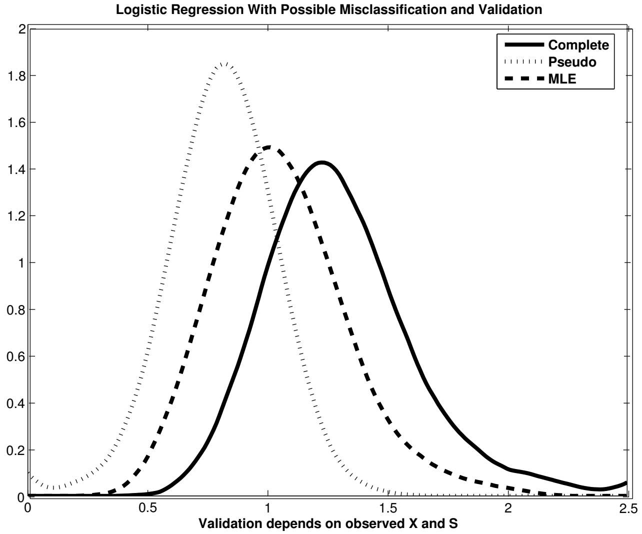


Figure 15.8 *Illustration of the effect of misclassification of the response in logistic regression, when selection into the validation study depends on* **S** *and* **Z**. *Solid line: density estimate of the slope in the logistic regression for the validation data. Dashed line: density estimate for the MLE. Dotted line: density estimate for pseudolikelihood. The actual slope is* 1.0*: note the bias in all but the MLE.*

However, in Figure 15.8 validation is more complex and depends on both **S** and **Z**. If $\mathbf{S} = 1$ and $\mathbf{Z} > 0$, we observe **Y** with probability 0.05. If $\mathbf{S} = 1$ and $\mathbf{Z} \leq 0$, we observe **Y** with probability 0.15. If $\mathbf{S} = 0$ and $\mathbf{Z} > 0$, we observe **Y** with probability 0.20. If $\mathbf{S} =$ and $\mathbf{Z} \leq 0$, we observe **Y** with probability 0.40. This figure shows that only the maximum likelihood estimate is unbiased, with the complete data estimates and the pseudolikelihood estimates incurring substantial bias.

Validation Data			
Z	**S**	**Y**	Count
0	0	0	19
0	0	1	5
0	1	0	7
0	1	1	14
1	0	0	28
1	0	1	27
1	1	0	8
1	1	1	24
Nonvalidation Data			
0	0	–	47

Table 15.1 *GVHD data set. Here* $\mathbf{Y} = 1$ *if the patient develops chronic GVHD and* $= 0$ *otherwise, while* $\mathbf{S} = 1$ *if the patient develops acute GVHD. The predictor* $\mathbf{Z} = 1$ *if the patient is aged 20 or greater, and zero otherwise.*

15.3.2.4 Example

In this section, we present an example where selection into the validation study depends on the mismeasured response. We compare the maximum likelihood estimate with the naive use of the complete data. The latter is not valid and appears to be seriously biased in this case.

Pepe (1992) and Pepe et al. (1994) described a study of 179 aplastic anemia patients given bone marrow transplants. The objective of the analysis is to relate patient age to incidence of chronic graft versus host disease (GVHD). Patients who develop acute GVHD, which manifests itself early in the post transplant period,are at high risk of developing chronic GVHD. Thus, in this example **Y** is chronic GVHD, **S** is acute GVHD, and $\mathbf{Z} = 0, 1$ depending on whether or not a patient is less than 20 years of age. The data are given in Table 15.1. A logistic regression model for **Y** given **Z** is assumed.

The selection process as described by Pepe et al. (1994) is to select only 1/3 of low risk patients (less than 20 years old and no acute GVHD) into the validation study, while following all other patients. Thus, $\pi(\mathbf{S}, \mathbf{Z}) = 1/3$ if $\mathbf{S} = 0$ and $\mathbf{Z} = 0$, otherwise $\pi(\mathbf{S}, \mathbf{Z}) = 1$. Note that, here, selection

	Validation Data	MLE
$\widehat{\beta}_z$	0.66	1.13
Standard Error	0.37	0.38
p-value	0.078	0.004

Table 15.2 *Analysis of GVHD data set, with the validation data analysis and the maximum likelihood analysis. In this data set, selection depends on both* **Z** *and* **S**, *so that an analysis based only upon the validation data will lead to biased estimates and inference.*

into the validation study depends on both **S** and **Z**, so that an ordinary logistic regression analysis on the completed data ($\Delta = 1$) will be invalid.

We performed the following analyses: (i) use of validation or complete data only, which is not valid in this problem because of the nature of the selection process, but is included for comparison, and (ii) maximum likelihood. The results of the two analyses are listed in Table 15.2. We see that the validation data analysis is badly biased relative to the valid maximum likelihood analysis, with markedly different significant levels.

15.3.2.5 Repeats and Multiple Instruments

We have seen above that there is very little information about the misclassification probabilities if we only observe **S**, while validation data in which **Y** is observed does provide such information. Going from nothing to everything is a large gap!

Suppose that there is no validation study component. In some cases, experiments done by others in which **Y** is observed provide information about the misclassification, along with standard error estimates. Using these estimates provides a means of estimation of the underlying logistic regression model, with standard errors that can be propagated through by drawing bootstrap samples from the previous study and the current study separately.

In other cases, replication of **S** can be used to gain information about the misclassification probabilities. For example, if **Y** and **S** are binary, and if the misclassification probability is the same for both values of **Y**, then two independent replicates of **S** per person suffice to identify the misclassification probability. Otherwise, at least three independent replicates are necessary for identification. Whether technical identifiability results in practical identifiability is not clear, and one has to expect that in the absence of a strong prior distribution on the misclassification rates, the effect of misclassification will be to lower power greatly.

15.4 Likelihood Methods

In this section, we describe the technical details of likelihood methods for response measurement error. As seen in Section 15.3, one has to be careful to separate out theoretical identifiability of parameters from actual identifiability, the latter meaning that there is enough information about the parameters in the observed data to make their estimation practical.

15.4.1 General Likelihood Theory and Surrogates

Let $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}(s|y,z,\gamma)$ denote the density or mass function for **S** given $(\mathbf{Y},\mathbf{Z})$. We will call **S** a *surrogate response* if its distribution depends only on the true response, that is $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}(s|y,z,\gamma) = f_{\mathbf{S}|\mathbf{Y}}(s|y,\gamma)$. All the models we have considered in detail to this point are for surrogate responses.

In the case of a surrogate response, a very pleasant thing occurs. Specifically, *if there is no relationship between the true response* **Y** *and the predictors, then neither is there one between the observed response* **S** *and the predictors*. Thus, if one's only goal is to check whether there is any predictive ability in any of the predictors, and if **S** is a surrogate, then using the observed data provides a valid test. However, like everything having to do with measurement error, a valid test does not mean a powerful test: measurement error in the response lowers power.

This definition of a surrogate response is the natural counterpart to a surrogate predictor, because it implies that all the information in the relationship between **S** and the predictors is explained by the underlying response.

In general, that is for a possibly nonsurrogate response, the likelihood function for the observed response is

$$f_{\mathbf{S}|\mathbf{Z}}(s|z,\mathcal{B},\gamma) = \sum_{y} f_{\mathbf{Y}|\mathbf{Z}}(y|z,\mathcal{B}) f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}(s|y,z,\gamma). \tag{15.4}$$

If **Y** is a continuous random variable, the sum is replaced by an integral.

If **S** is a surrogate, then $f_{\mathbf{S}|\mathbf{Y}}(s|y,\gamma)$ replaces $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}(s|y,z,\gamma)$ in (15.4) showing that if there is no relationship between the true response and the predictors, then neither is there one between the observed response and the predictors. The reason for this is that under the stated conditions, neither term inside the integral depends on the predictors: the first because **Y** is not related to **Z**, and the second because **S** is a surrogate. However, if **S** is *not* a surrogate, then there may be no relationship between the true response and the covariates, but the observed response may be related to the predictors.

It follows that if interest lies in determining whether the predictors contain any information about the response, one can use naive hypothesis tests and ignore response error only if $\mathbf{S}$ is a surrogate. The resulting tests have asymptotically correct level, but decreased power relative to tests derived from true response data. This property of a surrogate is important in clinical trials; see Prentice (1989).

Note that one implication of (15.4) is that a likelihood analysis with mismeasured responses requires a model for the distribution of response error. We have already seen an example of this approach in Section 15.3.

Just as in the predictor-error problem, it is sometimes, but not always, the case that the parameters $(\mathcal{B}, \gamma)$ are identifiable, that is can be estimated from data on $(\mathbf{S}, \mathbf{Z})$ alone. We have seen two examples of this: (a) in regression models with a continuous response and additive unbiased measurement error, the parameters in the model for the mean are identified; and (b) logistic regression when $\mathbf{S}$ is a surrogate. Of course, in the latter case, as seen in Section 15.3.2, the identifiability is merely a technical one, not practical.

15.4.2 Validation Data

We now suppose that there is a validation subsample obtained by measuring $\mathbf{Y}$ on units in the primary sample selected with probability $\pi(\mathbf{S}, \mathbf{Z})$. The presence (absence) of validation data on a primary-sample unit is indicated by $\Delta = 1$ (0). Then, based on a primary sample of size n, the likelihood of the observed data for a general proxy $\mathbf{S}$ is

$$\prod_{i=1}^{n} \Big[\{f(\mathbf{S}_i|\mathbf{Y}_i, \mathbf{Z}_i, \gamma) f(\mathbf{Y}_i|\mathbf{Z}_i, \mathcal{B})\}^{\Delta_i} \times \\ \{f(\mathbf{S}_i|\mathbf{Z}_i, \mathcal{B}, \gamma)\}^{1-\Delta_i} \Big], \tag{15.5}$$

where $f(\mathbf{S}_i|\mathbf{Z}_i, \mathcal{B}, \gamma)$ is computed by (15.4) and we have dropped the subscripts on the density functions for brevity.

The model for the distribution of $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z})$ is a critical component of (15.5). If $\mathbf{S}$ is discrete, then one approach is to model this conditional distribution by a polytomous logistic model. For example, suppose the levels of $\mathbf{S}$ are $(0, 1, \ldots, \mathcal{S})$. A standard logistic model is

$$\text{pr}(\mathbf{S} \geq s|\mathbf{Y}, \mathbf{Z}) = H(\gamma_{0s} + \gamma_1 \mathbf{Y} + \gamma_2^t \mathbf{Z}), \quad s = 1, \ldots, \mathcal{S}.$$

When $\mathbf{S}$ is not discrete, a simple strategy is to categorize it into $\mathcal{S}$ levels, and then use the logistic model above.

As described above, likelihood analysis is, in principle, straightforward. There are two obvious potential drawbacks, namely that one has to worry about the model for the measurement error and then one has to

compute the likelihood. These are little different from what is required for any likelihood problem.

15.5 Use of Complete Data Only

We have downplayed descriptions of the very large literature when there is a gold standard for a covariate $\mathbf{X}$ measured with error. This huge literature, which includes both the missing data likelihood literature and the missing data semiparametric literature, tends to be technical and entire books can, and have been, written about them.

In the case that the response $\mathbf{Y}$ can be observed on a subset of the study data, the literature is much smaller and more manageable. Herewith we make a few remarks on methods that use the validation data only, throwing away any of the data when $\mathbf{Y}$ is not observed. It is not entirely clear why one would do this instead of performing a complete likelihood or Bayesian analysis, except in the case of logistic regression with a surrogate where selection depends only on $\mathbf{S}$, in which case the various validation data analyses are simple variants of logistic regression. Given the availability of logistic regression software, this is certainly a useful simplification.

In what follows, selection into the validation study occurs with probability $\pi(\mathbf{S}, \mathbf{Z})$. Let $\Delta = 1$ denote selection into the validation study.

15.5.1 Likelihood of the Validation Data

The validation data have the likelihood function for a single observation given by

$$f(\mathbf{Y}, \mathbf{S}|\mathbf{Z}, \Delta = 1) = \\ \frac{\pi(\mathbf{S}, \mathbf{Z}) f(\mathbf{S}|\mathbf{Y}, \mathbf{Z}, \gamma) f(\mathbf{Y}|\mathbf{Z}, \mathcal{B})}{\sum_s \sum_y \pi(s, \mathbf{Z}) f(s|y, \mathbf{Z}, \gamma) f(y|\mathbf{Z}, \mathcal{B})}, \tag{15.6}$$

where again if $\mathbf{S}$ or $\mathbf{Y}$ are continuous, the sums are replaced by integrals.

Here are a few implications of (15.6):

- If selection into the validation study is completely at random, or if it simply depends on the predictors but not $\mathbf{S}$, then one can run the standard analysis on the $(\mathbf{Y}, \mathbf{Z})$ data and ignore $\mathbf{S}$ entirely. Checking this is a small math calculation.
- In general, (15.6) cannot be simplified, and in particular, using the standard analysis on the observed $(\mathbf{Y}, \mathbf{Z})$ data leads to bias; see Figure 15.8.
- Logistic regression has a very special place here. If $\mathbf{S}$ is a binary surrogate, and if selection into the validation study depends on $\mathbf{S}$

only, then running a logistic regression ignoring the very existence of $\mathbf{S}$ leads to valid inference about the nonintercept parameters see Tosteson and Ware (1990).

15.5.2 Other Methods

In some problems, it can occur that there are two data sets; a primary one in which $(\mathbf{S}, \mathbf{Z})$ are observed $(\Delta = 0)$, and an *independent* data set in which $(\mathbf{Y}, \mathbf{Z})$ are observed $(\Delta = 1)$. This may occur when $\mathbf{Y}$ is a sensitive endpoint such as income, and $\mathbf{S}$ is reported income. Because of confidentiality concerns, it might be impossible to measure $\mathbf{Y}$ and $\mathbf{S}$ together. In such problems, the likelihood is

$$\prod_{i=1}^{n} \{f(\mathbf{Y}_i|\mathbf{Z}_i, \mathcal{B})\}^{\Delta_i} \{f(\mathbf{S}_i|\mathbf{Z}_i, \mathcal{B}, \gamma)\}^{1-\Delta_i}.$$

15.6 Semiparametric Methods for Validation Data

As we have suggested, likelihood methods can potentially be troublesome because they might be sensitive to the assumed distribution for the mismeasured response. This has led to a small literature on semiparametric methods, which attempt in various guises to model the distribution of $\mathbf{S}$ given $\mathbf{Y}$ and the covariates *nonparametrically.*

15.6.1 Simple Random Sampling

Suppose that selection into the second stage validation study is by simple random sampling, that is, all possible samples of the specified sample size are equally likely. Pepe (1992) constructed a pseudolikelihood method similar in spirit to that of Carroll and Wand (1991) and Pepe and Fleming (1991) for the mismeasured covariate problem with validation data. The basic idea is to use the validation data to form a nonparametric estimator $\widehat{f}_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}$ of $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}$. One then substitutes this estimator into (15.4) to obtain an estimator $\widehat{f}_{\mathbf{S}|\mathbf{Z}}(s|z, \mathcal{B})$ and then maximizes

$$\prod_{i=1}^{n} \{f(\mathbf{Y}_i|\mathbf{Z}_i, \mathcal{B})\}^{\Delta_i} \left\{\widehat{f}(\mathbf{S}_i|\mathbf{Z}_i, \mathcal{B})\right\}^{1-\Delta_i}.$$

This approach requires an estimator of $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}}$. Here are a few comments:

- If all the random variables are discrete, the nonparametric estimator of the probability that $\mathbf{S} = s$ given $(\mathbf{Y}, \mathbf{Z}) = (y, z)$ is the fraction in the validation study which have $\mathbf{S} = s$ among those with $(\mathbf{Y}, \mathbf{Z}) = (y, z)$, although we prefer flexible parametric models in this case.

- Problems that have continuous components of $(\mathbf{S}, \mathbf{Y}, \mathbf{Z})$ are more complicated. For example, suppose that $\mathbf{S}$ is continuous, but the other random variables are discrete. Then the density function of $\mathbf{S}$ in *each* of the cells formed by the various combinations of $(\mathbf{Y}, \mathbf{Z})$ must be estimated. Even in the simplest case that $(\mathbf{Y}, \mathbf{Z})$ are binary, this means estimating four density functions using validation data only. While the asymptotic theory of such a procedure has been investigated (Pepe, 1992), we know of no numerical evidence indicating that the density estimation methods will work adequately in finite samples, nor is there any guidance on the practical problems of bandwidth selection and dimension reduction when two or more components of $(\mathbf{S}, \mathbf{Y}, \mathbf{Z})$ are continuous.
- In practice, if $\mathbf{S}$ is not already naturally categorical, then an alternative strategy is to perform such categorization, fit a flexible logistic model to the distribution of $\mathbf{S}$ given the other variables, and maximize the resulting likelihood (15.5).

15.6.2 Other Types of Sampling

Pseudolikelihood can be modified when selection into the second stage of the study is not by simple random sampling. The estimating equations for the EM-algorithm maximizing (15.5) are

$$\begin{aligned} 0 \quad = \quad & \sum_{i=1}^{n} \Delta_i \{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{B}) + \Psi_2(\mathbf{S}_i, \mathbf{Y}_i, \mathbf{Z}_i, \gamma)\} \\ & + \sum_{i=1}^{n} (1 - \Delta_i) E \{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{B}) \\ & + \Psi_2(\mathbf{S}_i, \mathbf{Y}_i, \mathbf{Z}_i, \gamma)|\mathbf{S}_i, \mathbf{Z}_i\}, \end{aligned}$$

where

$$\begin{aligned} \Psi_1 &= ((\partial/\partial\mathcal{B}) \log(f_{\mathbf{Y}|\mathbf{Z}})^t,\, 0^t)^t, \\ \Psi_2 &= (0^t, (\partial/\partial\gamma) \log(f_{\mathbf{S}|\mathbf{Y},\mathbf{Z}})^t)^t. \end{aligned}$$

The idea is to use the validation data to estimate

$$E\{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{B})|\mathbf{S}_i, \mathbf{Z}_i\}$$

and then solve

$$\begin{aligned} 0 \quad = \quad & \sum_{i=1}^{n} [\Delta_i \Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{B}) + \\ & \qquad (1 - \Delta_i) \widehat{E} \{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{B})|\mathbf{S}_i, \mathbf{Z}_i\}\Big]. \end{aligned}$$

For example, suppose that $(\mathbf{S}, \mathbf{Z})$ are all discrete. Now define I_{ij} to equal one when $(\mathbf{S}_j, \mathbf{Z}_j) = (\mathbf{S}_i, \mathbf{Z}_i)$ and zero otherwise. Then

$$\widehat{E}\{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{B})|\mathbf{S}_i, \mathbf{Z}_i\} = \frac{\sum_{j=1}^n \Delta_j \Psi_1(\mathbf{Y}_j, \mathbf{Z}_j, \mathcal{B}) I_{ij}}{\sum_{j=1}^n \Delta_j I_{ij}}.$$

In other cases, nonparametric regression can be used. In the discrete case, Pepe et al. (1994) derived an estimate of the asymptotic covariance matrix of $\widehat{\mathcal{B}}$ as $A^{-1}(A+B)A^{-t}$, where

$$\begin{aligned} A &= -\sum_{i=1}^n \Delta_i (\partial/\partial \mathcal{B}^{\mathcal{T}}) \Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \widehat{\mathcal{B}}) \\ &\quad - \sum_{i=1}^n (1-\Delta_i) \frac{\sum_{j=1}^n \Delta_j (\partial/\partial \mathcal{B}^{\mathcal{T}}) \Psi_1(\mathbf{Y}_j, \mathbf{Z}_j, \widehat{\mathcal{B}}) I_{ij}}{\sum_{j=1}^n \Delta_j I_{ij}}; \\ B &= \sum_{s,z} \frac{n(s,z) n_2(s,z)}{n_1(s,z)} r(s,z,\widehat{\mathcal{B}}), \end{aligned}$$

$n_1(s,z)$, $n_2(s,z)$, and $n(s,z)$ are the number of validation, nonvalidation and total cases with $(\mathbf{S}, \mathbf{Z}) = (s,z)$, and where $r(s,z,\widehat{\mathcal{B}})$ is the sample covariance matrix of $\Psi_1(\mathbf{Y}, \mathbf{Z}, \widehat{\mathcal{B}})$ computed from observations with $(\Delta, \mathbf{S}, \mathbf{Z}) = (1, s, z)$.

Bibliographic Notes

Lyles, Williamson, Lin, and Heilig (2005) extend McNemar's test for paired binary outcomes to the situation where the outcomes are misclassified.

APPENDIX A

BACKGROUND MATERIAL

A.1 Overview

This Appendix collects some of the technical tools that are required for understanding the theory employed in this monograph. The background material is, of course, available in the literature, but often widely scattered, and one can use this chapter as a brief tour of likelihood, quasi-likelihood and estimating equations.

Section A.2 and A.3 discuss the normal and lognormal, respectively, gamma and inverse-gamma distributions. Section A.4 discusses prediction of an unknown random variable by another random variable and introduces "best prediction," which can be considered a population analog to regression; conversely, regression is the sample analog of best prediction. Section A.5 reviews likelihood methods, which will be familiar to most readers. Section A.6 is a brief introduction to the method of estimating equations, a widely applicable tool that is the basis of all estimators in this book. Section A.8 defines generalized linear models. The bootstrap is explained in Section A.9, but one need only note while reading the text that the bootstrap is a computer-intensive method for performing inference.

A.2 Normal and Lognormal Distributions

We say that X is Normal(μ_x, σ_x^2) if it is normally distributed with mean μ_x and variance σ_x^2. Then the density of X is $\phi\{(x-\mu_x)/\sigma_x\}$ where ϕ is the standard normal pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right), \tag{A.1}$$

and the CDF of X is $\Phi\{(x-\mu_x)/\sigma_x\}$, where $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$ is the standard normal CDF.

We say that the random vector $\mathbf{X} = (X_1, \ldots, X_p)^t$ has a joint multivariate normal distribution with mean (vector) μ_x and covariance matrix Σ_x if $\mathbf{X}$ has density

$$\frac{1}{(2\pi)^{p/2}|\Sigma_x|^{1/2}} \exp\left\{-(1/2)(\mathbf{X}-\mu_x)^t \Sigma_x^{-1}(\mathbf{X}-\mu_x)\right\}.$$

The random variable X is said to have a Lognormal(μ_x, σ_x^2) distribution if $\log(X)$ is Normal(μ_x, σ_x^2). In that case

$$E(X) = \exp(\mu_x + \sigma_x^2/2), \text{ and} \tag{A.2}$$

$$\mathrm{var}(X) = \exp(2\mu_x)\{\exp(2\sigma_x^2) - \exp(\sigma_x^2)\}. \tag{A.3}$$

A.3 Gamma and Inverse-Gamma Distributions

A random variable X has a Gamma(α, β) distribution if its probability density function is

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \;\; x > 0.$$

Here, $\Gamma(\cdot)$ is the gamma function, $\alpha > 0$ is a shape parameter, and $\beta^{-1} > 0$ is a scale parameter. A word of caution is in order—some authors denote the scale parameter by β, in which case their β is the reciprocal of ours. The expectation of this distribution is α/β, while its variance is α/β^2. The chi-square distribution with n degrees of freedom is the Gamma$(n/2, 1/2)$ distribution and arises as a sampling distribution in Gaussian models.

If X is Gamma(α, β), then the distribution of X^{-1} is called Inverse-Gamma(α, β), abbreviated as IG(α, β). Then α is the shape parameter of X and β is its scale (not inverse scale) parameter. The density of X^{-1} is

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp(\beta/x), \;\; x > 0.$$

The mode of the IG(α, β) density is $\beta/(\alpha+1)$, and the expectation is $\beta/(\alpha-1)$ for $\alpha > 1$. For $\alpha \leq 1$, the expectation is infinite.

The inverse Gamma distribution is the conjugate prior for variance parameters in many Gaussian models. As a simple case, suppose that $X_1, \ldots, X_n$ are iid Normal$(0, \sigma^2)$. Then the likelihood is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n X_i^2}{2\sigma^2}\right).$$

Therefore, if the prior on σ^2 is IG(α, β), then the joint density of $\sigma^2, X_1, \ldots, X_n$ is proportional to the IG$(\alpha + n/2, \beta + \sum_{i=1}^n X_i^2/2)$ density. It follows that the posterior distribution of σ^2 is IG$(\alpha + n/2, \beta + \sum X_i^2/2)$.

The parameters α and β in the prior have this interpretation: The prior is equivalent to 2α observations with sum of squares equal to 2β, which, if actually observed, would give a prior variance estimate of β/α. Thus, an Inverse-Gamma(α, β) prior can be viewed as a prior guess at the variance of β/α based on 2α observations. A value of α that is small relative to $n/2$ is "noninformative." Also, the value of β has little influence relative to the data only when it is small relative to $\sum_{i=1}^n X_i^2$.

Because $\sum_{i=1}^n X_i^2$ can be arbitrarily small, there can be no choice of β that is "noninformative" in all situations. For example, $\beta = 0.001$, which seems "small," will completely dominate the likelihood if $\sum_{i=1}^n X_i^2 = 0.0001$. Since the X_i are observed in the present example, one should be aware when β is large relative to $\sum_{i=1}^n X_i^2$. However, in the case of a prior on the variance of unobservable random effects, more care is required. Otherwise, the prior might dominate the likelihood. At the very least, one should reconsider the choice of β unless it is smaller than the posterior mean of this variance.

A.4 Best and Best Linear Prediction and Regression

A.4.1 Linear Prediction

Let X and Y be any two random variables. If the value of Y is unknown but X is known and is correlated with Y, then we can estimate or "predict" Y using X. The best linear predictor of Y is $\gamma_0 + \gamma_x X$, where γ_0 and γ_x are chosen to minimize the mean square error

$$E\{Y - (\gamma_0 + \gamma_x X)\}^2 = [E\{Y - (\gamma_0 + \gamma_x X)\}]^2 + \mathrm{var}(Y - \gamma_x X). \tag{A.4}$$

On the right-hand side of (A.4), the first term is squared bias and the second is variance. The variance does not depend on γ_0, so the optimal value of γ_0 is $\mu_y - \gamma_x \mu_x$, which eliminates the bias. The variance is

$$\sigma_y^2 + \gamma_x^2 \sigma_x^2 - 2\gamma_x \sigma_{xy},$$

where σ_{xy} is the covariance between X and Y, and an easy calculus exercise shows that the variance is minimized by $\gamma_x = \sigma_{xy}/\sigma_x^2$. In summary, the best linear predictor of Y based on X is

$$\widehat{Y} = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(X - \mu_x). \tag{A.5}$$

The prediction error is

$$Y - \widehat{Y} = (Y - \mu_y) - \frac{\sigma_{xy}}{\sigma_x^2}(X - \mu_x), \tag{A.6}$$

where σ_{xy} is the covariance between X and Y. It is an easy calculation to show that the prediction error is uncorrelated with X and that

$$\mathrm{var}(Y - \widehat{Y}) = \sigma_y^2(1 - \rho_{xy}^2). \tag{A.7}$$

There is an intuitive reason for this—if the error were correlated with W, then the error itself could be predicted and therefore we could improve the predictor of Y, but we know that this predictor cannot be improved since it was chosen to be the best possible.

As an illustration, consider the classical error model $W = X + U$, where X and U are uncorrelated and $E(U) = 0$. Then $\sigma_{xw} = \sigma_x^2$, $\sigma_{xw}/\sigma_w^2 = \lambda$

(the attenuation), and $\mu_w = \mu_x$, so the best linear predictor of X given W is

$$\widehat{X} = \mu_x + \lambda(W - \mu_x) = (1 - \lambda)\mu_x + \lambda W \tag{A.8}$$

and

$$X = \mu_x + \lambda(W - \mu_x) + U^*, \tag{A.9}$$

where the prediction error $U^* = X - \widehat{X}$ is uncorrelated with W and has variance $\sigma_x^2(1 - \lambda)$, since $\lambda = \rho_{xw}^2$ because $\sigma_{xw} = \sigma_x^2$.

So far, we have assumed the ideal situation where μ_x, σ_{xy}, and σ_x^2 are known. In practice, they will generally be unknown and replaced by estimates. If we observe an iid sample, (Y_i, X_i), then we use the sample means, variances, and covariances and $\widehat{\gamma}_x = \widehat{\sigma}_{xy}/\widehat{\sigma}_x^2$ and $\widehat{\gamma}_0 = \overline{Y} - \widehat{\gamma}_x\overline{X}$ are the usual ordinary least-squares estimates with

$$\widehat{\gamma}_x = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

If we have more than one X-variable, then the best linear predictor of Y given $\mathbf{X} = (X_1, \ldots, X_p)$ is

$$\widehat{Y} = \mu_y + \Sigma_{yx}\Sigma_x^{-1}(\mathbf{X} - \mu_x), \tag{A.10}$$

where $\Sigma_{yx} = (\sigma_{YX_1}, \ldots, \sigma_{YX_p})$ and Σ_x is the covariance matrix of $\mathbf{X}$. If the means, variances, and covariances are replaced by their analogs from a sample, $(Y_p, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, then (A.10) becomes the ordinary least-squares estimator.

Equation (A.10) remains valid when Y is a vector. Then

$$\Sigma_{yx} = E\left[\{(Y - E(Y)\}\{(X - E(X)\}^t\right]. \tag{A.11}$$

As an illustration of the vector Y case, we will generalize (A.8) to the case where

$$\mathbf{W} = \mathbf{X} + \mathbf{U}$$

with $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{U}$ all vectors. Then $\Sigma_w = \Sigma_x + \Sigma_u$ and

$$\widehat{\mathbf{X}} = \mu_x + \Lambda(\mathbf{W} - \mu_x) = (I - \Lambda)\mu_x + \Lambda\mathbf{W}, \tag{A.12}$$

where I is the identity matrix and

$$\Lambda = \Sigma_x(\Sigma_x + \Sigma_u)^{-1}. \tag{A.13}$$

Here, we are assuming that $(\Sigma_x + \Sigma_u)$ is invertible, which will hold if Σ_x is invertible since Σ_u must be positive semidefinite. Note that Λ is a multivariate generalization of the attenuation λ.

If the distribution of $\mathbf{X}$ depends on $\mathbf{Z}$, then (A.12) is replaced by

$$\widehat{\mathbf{X}} = \mu_x + \begin{pmatrix} \Sigma_x & \Sigma_{xz} \end{pmatrix} \begin{pmatrix} \Sigma_x + \Sigma_u & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_z \end{pmatrix}^{-1} \left\{ \begin{pmatrix} \mathbf{W} \\ \mathbf{Z} \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_z \end{pmatrix} \right\}, \tag{A.14}$$

where $\Sigma_{xz} = E\left[\{\mathbf{X} - \mu_x\}\{\mathbf{Z} - \mu_z\}^t\right]$. When $\mathbf{W}$ is replicated, (A.14) generalizes to (4.4).

A.4.2 Best Linear Prediction without an Intercept

Prediction without an intercept is inferior to prediction with an intercept, except in the rare case where the intercept of the best linear predictor is zero. So you no doubt are wondering why we are bothering you with this topic. The reason is that this material is needed to understand the Fuller–Hwang estimator discussed in Section 4.5.3. You should read this section only if you are reading Section 4.5.3.

If X and Y are scalars and we predict Y by a linear function of X without an intercept, that is, with a predictor of form λX, then a simple calculation shows that the mean squared error (MSE) of prediction is minimized by $\lambda = E(XY)/E(X^2)$.

From a sample of $\{(Y_i, X_i)\}_{i=1}^n$ pairs, $\widehat{Y}_{\text{NI}}$ can be estimated by linear regression without an intercept. Here the subscript "NI" reminds us that the prediction is done with "no intercept." The estimate of λ is

$$\widehat{\lambda} = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2}.$$

As an example, consider the multiplicative error model $W = XU$ with true covariate X and surrogate W, where X and U are independent and $E(U) = 1$. We need to predict X using W, so now X plays the role played by Y in the previous three paragraphs and W plays the role of X. Then $E(W^2) = E(X^2)\ E(U^2)$ and, assuming that $\text{var}(U) \neq 0$, one has that $\lambda = 1/E(U^2) < 1$ since $E(U^2) > \{E(U)\}^2 = 1$. The best linear predictor of X without an intercept is $\widehat{X}_{\text{NI}} = \lambda W$. The predictor $\widehat{X}_{\text{NI}}$ will have at least as large an MSE of prediction as the best linear predictor with an intercept. One problem with prediction without an intercept is that it is biased in the sense that $E(\widehat{X}_{\text{NI}}) = \lambda E(X) \neq E(X)$, assuming that $\text{var}(U) \neq 0$. This is because $\widehat{X}_{\text{NI}}$ attenuates towards zero, whereas from (A.9) we see that $\widehat{X}$ shrinks W towards μ_x.

A.4.3 Nonlinear Prediction

If we do not constrain the predictor to be linear in $\mathbf{X}$, then the predictor minimizing the MSE over *all* functions of $\mathbf{X}$ is the conditional mean of Y given $\mathbf{X}$, that is, the best predictor of Y given $\mathbf{X}$ is $\widehat{Y} = E(Y|\mathbf{X})$.

There are some circumstances in which $E(Y|\mathbf{X})$ is linear in $\mathbf{X}$, in which case the best predictor and the best linear predictor coincide. For example, this happy situation occurs whenever $(Y, \mathbf{X})$ is jointly normally distributed. Another case in which the best predictor is linear in $\mathbf{X}$ occurs when the linearity is a consequence of modeling assumption, for example, if it assumed that $Y = \beta_0 + \beta_x^t \mathbf{X} + \epsilon$, where ϵ is independent of $\mathbf{X}$.

Finding the best predictor $E(Y|\mathbf{X})$ requires knowledge of the joint distribution of $(Y, \mathbf{X})$, not simply means, variance, and covariances. In practice, $E(Y|\mathbf{X})$ is estimated by nonparametric regression. There are many techniques for nonparametric regression, for example, smoothing splines, local polynomial regression such as LOESS, and fixed-knot penalized splines (Ruppert et al., 2003). When $\mathbf{X}$ is multivariate, one typically uses dimension-reduction techniques such as additive models or sliced-inverse regression to avoid the so-called curse of dimensionality.

A.5 Likelihood Methods

A.5.1 Notation

Denote the unknown parameter by Θ. The vector of observations, including response, covariates, surrogates, etc., is denoted by $(\widetilde{\mathbf{Y}}_i, \mathbf{Z}_i)$ for $i = 1, ..., n$, where, as before, $\mathbf{Z}_i$ is the vector of covariates that is observable without error and $\widetilde{\mathbf{Y}}_i$ collects all the other random variables into one vector. The data set $(\widetilde{\mathbf{Y}}_i, \mathbf{Z}_i)$, $i = 1, ..., n$, is the aggregation of all data sets, primary and external, including replication and validation data. Thus, the composition of $\widetilde{\mathbf{Y}}_i$ will depend on i, for example, whether the ith case is a validation case, a replication case, etc. *We emphasize that $\widetilde{\mathbf{Y}}_i$ is different from the response $\mathbf{Y}_i$ used throughout the book, and hence the use of tildes.* The $\widetilde{\mathbf{Y}}_i$ are assumed independent, with the density of $\widetilde{\mathbf{Y}}_i$ depending both on $\mathbf{Z}_i$ and on the type of data set the ith case came from and denoted by $f_i(\widetilde{y}|\Theta)$. We assume that f_i has two continuous derivatives with respect to Θ. The loglikelihood is

$$L(\Theta) \quad = \quad \textstyle\sum_{i=1}^n \log\{f_i(\widetilde{\mathbf{Y}}_i|\Theta)\}.$$

A.5.2 Maximum Likelihood Estimation

In practice, maximum likelihood is probably the most widely used method of estimation. It is reasonably easy to implement, efficient, and the basis of readily available inferential methods, such as standard errors by Fisher information and likelihood ratio tests. Also, many other common estimators are closely related to maximum likelihood estimators, for example, the least squares estimator, which is the maximum likelihood estimator under certain circumstances, and quasilikelihood estimators. In this section, we quickly review some of these topics.

The maximum likelihood estimator (MLE), denoted by $\widehat{\Theta}$, maximizes $L(\Theta)$. Under some regularity conditions, for example in Serfling (1980), the MLE has a simple asymptotic distribution. The "likelihood score" or "score function" is $s_i(y|\Theta) = (\partial/\partial\Theta)\log\{f_i(y|\Theta)\}$. The Fisher information matrix, or expected information, is

$$\begin{aligned} I_n(\Theta) &= -\textstyle\sum_{i=1}^n E\{(\partial/\partial\Theta^t) s_i(\widetilde{\mathbf{Y}}_i|\Theta)\} && \text{(A.15)} \\ &= \textstyle\sum_{i=1}^n E\{s_i(\widetilde{\mathbf{Y}}_i|\Theta) s_i^t(\widetilde{\mathbf{Y}}_i, |\Theta)\}. && \text{(A.16)} \end{aligned}$$

In large samples, the MLE is approximately normally distributed with mean Θ and covariance matrix $I_n^{-1}(\Theta)$, whose entries converge to 0 as $n \to \infty$. There are several methods of estimating $I_n(\Theta)$. The most obvious is $I_n(\widehat{\Theta})$. Efron and Hinkley (1978) presented arguments in favor of using instead the observed Fisher information matrix, defined as

$$\widehat{I}_n = -\textstyle\sum_{i=1}^n \dfrac{\partial}{\partial\Theta^t} s_i(\widetilde{\mathbf{Y}}_i|\widehat{\Theta}), \qquad \text{(A.17)}$$

which is an empirical version of (A.15). The empirical version of (A.16) is

$$\widehat{B}_n = \textstyle\sum_{i=1}^n s_i(\widetilde{\mathbf{Y}}_i|\widehat{\Theta}) s_i^t(\widetilde{\mathbf{Y}}_i|\widehat{\Theta}),$$

which is not used directly to estimate I_n, but is part of the so-called sandwich formula, $\widehat{I}_n^{-1}\widehat{B}_n^{-1}\widehat{I}_n^{-1}$, used to estimate $I_n^{-1}(\Theta)$. As discussed in Section A.6, the sandwich formula has certain "robustness" properties but can be subject to high sampling variability.

A.5.3 Likelihood Ratio Tests

Suppose that the dimension of Θ is $\dim(\Theta) = p$, that φ is a known function of Θ such that $\dim\{\varphi(\Theta)\} = p_1 < p$, and that we wish to test $H_0 : \varphi(\Theta) = 0$ against the general alternative that $\varphi(\Theta) \neq 0$. We suppose that $\text{rank}\{(\partial/\partial\Theta^t)\,\varphi(\Theta)\} = p_1$, so that the constraints imposed by the null hypothesis are linearly independent; otherwise p_1, is not well defined, that is, we can add redundant constraints and increase p_1 without changing H_0, and the following result is invalid.

Let $\widehat{\Theta}_0$ maximize $L(\Theta)$ subject to $\varphi(\Theta) = 0$, and define $LR = \{L(\widehat{\Theta}) - L(\widehat{\Theta}_0)\}$, the log likelihood ratio. Under H_0, $2 \times LR$ converges in distribution to the chi-squared distribution with p_1 degrees of freedom. Thus, an asymptotically valid test rejects the null hypothesis if LR exceeds $\chi^2_{p_1}(\alpha)/2$, the $(1-\alpha)$ quantile of the chi-squared distribution with p_1 degrees of freedom.

A.5.4 Profile Likelihood and Likelihood Ratio Confidence Intervals

Profile likelihood is used to draw inferences about a single component of the parameter vector. Suppose that $\Theta = (\theta_1, \Theta_2)$, where θ_1 is univariate. Let c be a hypothesized value of θ_1. To test $H_0 : \theta_1 = c$ using the theory of Section A.5.3, we use $\varphi(\Theta) = \theta_1 - c$ and find $\widehat{\Theta}_2(c)$ so that $\{c, \widehat{\Theta}_2(c)\}$ maximizes L subject to H_0. The function $L_{\max}(\theta_1) = L\{\theta_1, \widehat{\Theta}_2(\theta_1)\}$ is called the *profile likelihood function* for θ_1—it does not involve Θ_2 since the log likelihood has been maximized over Θ_2. Then, $LR = L(\widehat{\Theta}) - L_{\max}(c)$ where, as before, $\widehat{\Theta}$ is the MLE. One rejects the null hypothesis if LR exceeds $\chi^2_1(\alpha)$.

Inference for θ_1 is typically based on the profile likelihood. In particular, the likelihood ratio confidence region for θ_1 is the set

$$\{\theta_1 : L_{\max}(\theta_1) > L(\widehat{\Theta}) - \chi^2_1(\alpha)/2\}.$$

This region is also the set of all c such that we cannot reject the null hypothesis $H_0 : \theta_1 = c$. The confidence region is typically an interval, but there can be exceptions. An alternative large-sample interval is

$$\widehat{\theta}_1 \pm \Phi^{-1}(1 - \alpha/2)\mathrm{se}(\widehat{\theta}_1), \tag{A.18}$$

where $\mathrm{se}(\widehat{\theta}_1)$ is the standard error of $\widehat{\theta}_1$, say from the Fisher information matrix or from bootstrapping, as in Section A.9. For nonlinear models, the accuracy of (A.18) is questionable, that is, the true coverage probability is likely to be somewhat different than $(1 - \alpha)$, and the likelihood ratio interval is preferred.

A.5.5 Efficient Score Tests

The efficient score test, or simply the "score test," is due to Rao (1947). Under the null hypothesis, the efficient score test is asymptotically equivalent to the likelihood ratio test, for example, the difference between the two test statistics converges to 0 in probability. The advantage of the efficient score test is that the MLE needs to be computed only under the null hypothesis, not under the alternative, as for the likelihood ratio test. This can be very convenient when testing the null hypothesis of no effects for covariates measured with error, since these covariates, and hence measurement error, can be ignored when fitting under H_0.

To define the score test, start by partitioning Θ as $(\Theta_1^t, \Theta_2^t)^t$, where $\dim(\Theta_1) = p_1$, $1 \le p_1 \le p$. We will test the null hypothesis that $H_0 : \Theta_1 = 0$. Many hypotheses can be put into this form, possibly after reparameterization. Let $S(\Theta) = \sum_{i=1}^n s_i(\widetilde{\mathbf{Y}}_i|\Theta)$ and partition S into S_1 and S_2 with dimensions p_1 and $(p - p_1)$, respectively, that is,

$$S(\Theta) = \begin{pmatrix} S_1(\Theta) \\ S_2(\Theta) \end{pmatrix} = \begin{pmatrix} (\partial/\partial\Theta_1)\sum_{i=1}^n \log\{f_i(y|\Theta)\} \\ (\partial/\partial\Theta_2)\sum_{i=1}^n \log\{f_i(y|\Theta)\} \end{pmatrix}.$$

Note that, in general, $S_1(\Theta)$ depends on both Θ_1 and Θ_2, and similarly for $S_2(\Theta)$. Let $\widehat{\Theta}_0 = (0^t, \widehat{\Theta}_{0,2}^t)^t$ be the MLE of Θ under H_0. Notice that $S_2(\widehat{\Theta}_0) = 0$ since $\widehat{\Theta}_{0,2}$ maximizes the likelihood over Θ_2 when $\Theta_1 = 0$.

The basic idea behind the efficient score test is that under H_0 we expect $S_1(\widehat{\Theta}_0)$ to be close to 0, since the expectation of $S(\Theta)$ is 0 and $\widehat{\Theta}_0$ is consistent for Θ.

Let I_n^{11} be the upper left corner of $(I_n)^{-1}$ evaluated at $\widehat{\Theta}_0$. The efficient score test statistic measures the departure of $S_1(\widehat{\Theta}_0)$ from 0 and is defined as

$$R_n = S_1(\widehat{\Theta}_0)^t I_n^{11} S_1(\widehat{\Theta}_0) = S(\widehat{\Theta}_0) I_n^{-1} S(\widehat{\Theta}_0).$$

The equality holds because $S_2(\widehat{\Theta}_0) = 0$.

Under H_0, R_n asymptotically has a chi-squared distribution with p_1 degrees of freedom, so we reject H_0 is R_n exceeds $(1 - \alpha)$ chi-squared quantile, $\chi^2_{p_1}(\alpha)$. See Cox and Hinkley (1974, Section 9.3) for a proof of the asymptotic distribution.

A.6 Unbiased Estimating Equations

All of the estimators described in this book, including the MLE, can be characterized as solutions to *unbiased estimating equations.* Understanding the relationship between estimators and estimating equations is useful because it permits easy and routine calculation of estimated standard errors. The theory of estimating equations arose from two distinct lines of research, in Godambe's (1960) study of efficiency and Huber's (1964, 1967) work on robust statistics. Huber's (1967) seminal paper used estimating equations to understand the behavior of the MLE under model misspecification, but his work also applies to estimators that are not the MLE under any model. Over time, estimating equations became established as a highly effective, unified approach for studying wide classes of estimators; see, for example, Carroll and Ruppert (1988) who use estimating equation theory to analyze a variety of transformation and weighting methods in regression.

This section reviews the basic ideas of estimating equations; See Huber (1967), Ruppert (1985), Carroll and Ruppert (1988), McLeish and Small (1988), Desmond (1989), or Godambe (1991) for more extensive discussion.

A.6.1 Introduction and Basic Large Sample Theory

As in Section A.5, the unknown parameter is Θ, and the vector of observations, including response, covariates, surrogates, etc., is denoted by $(\widetilde{\mathbf{Y}}_i, \mathbf{Z}_i)$ for $i = 1, ..., n$. For each i, let Ψ_i be a function of $(\widetilde{\mathbf{Y}}_i, \Theta)$ taking values in p-dimensional space ($p = \dim(\Theta)$). Typically, Ψ_i depends on i through $\mathbf{Z}_i$ and the type of data set the ith case belongs to, for example, whether that case is validation data, etc. An estimating equation for Θ

has the form

$$0 = n^{-1}\textstyle\sum_{i=1}^n \Psi_i(\widetilde{\mathbf{Y}}_i, \Theta). \tag{A.19}$$

The solution, $\widehat{\Theta}$, to (A.19) as Θ ranges across the set of possible parameter values is called an *M-estimator* of Θ, a term due to Huber (1964). In practice, one obtains an estimator by some principle, for example, maximum likelihood, least squares, generalized least squares, etc. Then, one shows that the estimator satisfies an equation of form (A.19) and Ψ_i is identified. The point is that one doesn't choose the Ψ_i's directly; but rather, they are defined through the choice of an estimator.

In (A.19), the function Ψ_i is called an *estimating function* and depends on i through $\mathbf{Z}_i$. The estimating function (and hence the estimating equation) is said to be *conditionally unbiased* if it has mean zero when evaluated at the true parameter, that is,

$$0 = E\left\{\Psi_i(\widetilde{\mathbf{Y}}_i, \Theta)\right\}, \text{ for } i = 1, ..., n. \tag{A.20}$$

As elsewhere in this book, expectations and covariances are always conditional upon $\{\mathbf{Z}_i\}_1^n$.

If the estimating equations are unbiased, then under certain regularity conditions $\widehat{\Theta}$ is a consistent estimator of Θ. See Huber (1967) for the regularity conditions and proof in the iid case. The basic idea is that for each value of Θ the right-hand side of (A.19) converges to its expectation by the law of large numbers, and the true Θ is a zero of the expectation of (A.19). One of the regularity conditions is that the true Θ is the only zero, so that $\widehat{\Theta}$ will converge to Θ under some additional conditions.

Moreover, if $\widehat{\Theta}$ is consistent, then by a Taylor series approximation

$$0 \approx n^{-1}\textstyle\sum_{i=1}^n \Psi_i(\widetilde{\mathbf{Y}}_i, \Theta) + \left\{n^{-1}\textstyle\sum_{i=1}^n \dfrac{\partial}{\partial\Theta^t}\Psi_i(\widetilde{\mathbf{Y}}_i, \Theta)\right\}(\widehat{\Theta} - \Theta),$$

where Θ now is the true parameter value. Applying the law of large numbers to the term in curly brackets, we have

$$\widehat{\Theta} - \Theta \approx -A_n(\Theta)^{-1} n^{-1}\textstyle\sum_{i=1}^n \Psi_i(\widetilde{\mathbf{Y}}_i, \Theta), \tag{A.21}$$

where $A_n(\Theta)$ is given by (A.23) below. Define $A_n^{-t}(\Theta) = \left\{A_n^{-1}(\Theta)\right\}^t$. Then $\widehat{\Theta}$ is asymptotically normally distributed with mean Θ and covariance matrix $n^{-1}A_n^{-1}(\Theta)B_n(\Theta)A_n^{-t}(\Theta)$, where

$$B_n(\Theta) = n^{-1}\textstyle\sum_{i=1}^n \text{cov}\left\{\Psi_i(\widetilde{\mathbf{Y}}_i, \Theta)\right\}; \tag{A.22}$$

$$A_n(\Theta) = n^{-1}\textstyle\sum_{i=1}^n E\left\{\dfrac{\partial}{\partial\Theta^t}\Psi_i(\widetilde{\mathbf{Y}}_i, \Theta)\right\}. \tag{A.23}$$

See Huber (1967) for a proof. There are two ways to estimate this covariance matrix. The first uses *empirical expectation* and is often called the *sandwich estimator* or a *robust covariance estimator* (a term we do not like—see below); in the former terminology, B_n is sandwiched between the inverse of A_n. The sandwich estimator uses

$$\widehat{A}_n = n^{-1}\textstyle\sum_{i=1}^n \dfrac{\partial}{\partial\Theta^t}\Psi_i(\widetilde{\mathbf{Y}}_i, \widehat{\Theta}); \tag{A.24}$$

$$\widehat{B}_n = n^{-1}\textstyle\sum_{i=1}^n \Psi_i(\widetilde{\mathbf{Y}}_i, \widehat{\Theta})\Psi_i^t(\widetilde{\mathbf{Y}}_i, \widehat{\Theta}). \tag{A.25}$$

Note that $\widehat{B}_n$ is a sample covariance matrix of $\{\Psi_i(\widetilde{\mathbf{Y}}_i, \widehat{\Theta})\}_{i=1}^n$, since $\widehat{\Theta}$ solves (A.19).

The second method, called the *model-based expectation method*, uses an underlying model to evaluate (A.22) and (A.23) exactly, and then substitutes the estimated value $\widehat{\Theta}$ for Θ, that is, uses $A_n^{-1}(\widehat{\Theta})\, B_n(\widehat{\Theta})A_n^{-t}(\widehat{\Theta})$.

If Ψ_i is the likelihood score, that is, $\Psi_i = s_i$, where s_i is defined in Section A.5.2, then $\widehat{\Theta}$ is the MLE. In this case, both $B_n(\Theta)$ and $A_n(\Theta)$ equal the Fisher information matrix, $I_n(\Theta)$. However, $\widehat{A}_n$ and $\widehat{B}_n$ are generally different, so the sandwich method differs from using the observed Fisher information.

As a general rule, the sandwich method provides a consistent estimate of the covariance matrix of $\widehat{\Theta}$, without the need to make any distribution assumptions. In this sense it is *robust*. However, in comparison with the model-based expectation method, when a distributional model is reasonable the sandwich estimator is typically inefficient, which can unnecessarily inflate the length of confidence intervals (Kauermann and Carroll, 2001). This inefficiency is why we do not like to call the sandwich method "robust." Robustness usually means insensitivity to assumptions at the price of a *small* loss of efficiency, whereas the sandwich formula can lose a great deal of efficiency.

A.6.2 Sandwich Formula Example: Linear Regression without Measurement Error

As an example, consider ordinary multiple regression without measurement errors so that $\mathbf{Y}_i = \beta_0 + \beta_z^t\mathbf{Z}_i + \epsilon_i$, where the ϵ's are independent, mean-zero random variables. Let $\mathbf{Z}_i^* = (1,\ \mathbf{Z}_i^t)^t$ and $\Theta = (\beta_0, \beta_z^t)^t$. Then the ordinary least squares estimator is an M-estimator with $\Psi_i(\mathbf{Y}_i, \Theta) = (\mathbf{Y}_i - \beta_0 - \beta_z^t\mathbf{Z}_i)\mathbf{Z}_i^*$. Also,

$$\begin{aligned}\frac{\partial}{\partial\Theta^t}\Psi_i(\mathbf{Y}_i, \Theta) &= -\mathbf{Z}_i^*(\mathbf{Z}_i^*)^t,\\ A_n &= -n^{-1}\textstyle\sum_{i=1}^n \mathbf{Z}_i^*(\mathbf{Z}_i^*)^t,\end{aligned} \tag{A.26}$$

and if one assumes that the variance of ϵ_i is a constant σ^2 for all i, then

$$B_n = -\sigma^2 A_n. \tag{A.27}$$

Notice that A_n and B_n do not depend on Θ, so they are known exactly except for the factor σ^2 in B_n. The model-based expectation method gives covariance matrix $-\sigma^2 A_n^{-1}$, the well known variance of the least squares estimator. Generally, σ^2 is estimated by the residual mean square.

The sandwich formula uses $\widehat{A}_n = A_n$ and

$$\widehat{B}_n = n^{-1}\textstyle\sum_{i=1}^n(\mathbf{Y}_i - \widehat{\beta}_0 - \widehat{\beta}_z^t\mathbf{Z}_i)^2\mathbf{Z}_i^*(\mathbf{Z}_i^*)^t. \tag{A.28}$$

We have not made distributional assumptions about ϵ_i, but we have assumed homoscedasticity, that is, that $\text{var}(\epsilon_i) \equiv \sigma^2$. To illustrate the "robustness" of the sandwich formula, consider the heteroscedatic model where the variance of ϵ_i is σ_i^2 depending on $\mathbf{Z}_i$. Then B_n is no longer given by (A.27) but rather by

$$B_n = n^{-1}\textstyle\sum_{i=1}^n\sigma_i^2\mathbf{Z}_i^*(\mathbf{Z}_i^*)^t,$$

which is consistently estimated by (A.28). Thus, the sandwich formula is heteroscedasticity consistent. In contrast, the model-based estimator of B_n, which is $\widehat{\sigma}^2 A_n$ with A_n given by (A.26), is inconsistent for B_n. This makes model-based estimation of the covariance matrix of $\widehat{\Theta}$ inconsistent.

The inefficiency of the sandwich estimator can also be seen in this example. Suppose that there is a high leverage point, that is, an observation with an outlying value of $\mathbf{Z}_i$. Then, as seen in (A.28), the value of $\widehat{B}_n$ is highly dependent upon the squared residual of this observation. This makes $\widehat{B}_n$ highly variable and indicates the additional problem that $\widehat{B}_n$ is very sensitive to outliers.

A.6.3 Sandwich Method and Likelihood-Type Inference

Less well known are likelihood ratio-type extensions of sandwich standard errors; see Huber (1967), Schrader and Hettmansperger (1980), Kent (1982), Ronchetti (1982), and Li and McCullagh (1994). This theory is essentially an extension of the theory of estimating equations, where the estimating equation is assumed to correspond to a criterion function, that is, solving the estimating equation minimizes the criterion function.

In the general theory, we consider inferences about a parameter vector Θ, and we assume that the estimate $\widehat{\Theta}$ maximizes an estimating criterion, $\ell(\Theta)$, which is effectively the working log likelihood, although, it need not be the logarithm of an actual density function. Following Li and McCullagh (1994), we refer to $\exp(\ell) = \exp(\sum \ell_i)$ as the quasilikelihood function. (Here, ℓ_i is the log quasilikelihood for the ith case and ℓ is the log quasilikelihood for the entire data set.) Define the score function, a type of estimating function, as

$$\mathbf{U}_i(\Theta) = \frac{\partial}{\partial\Theta}\ell_i(\Theta|\widetilde{\mathbf{Y}}_i),$$

the score covariance,

$$\mathcal{J}_n = \textstyle\sum_{i=1}^n \text{E}\{\mathbf{U}_i(\Theta)\,\mathbf{U}_i(\Theta)^t\}, \tag{A.29}$$

and the negative expected hessian,

$$\mathcal{H}_n = -\textstyle\sum_{i=1}^n \text{E}\left\{\frac{\partial}{\partial\Theta^t}\mathbf{U}_i(\Theta)\right\}. \tag{A.30}$$

If ℓ were the true log likelihood, then we would have $\mathcal{H}_n = \mathcal{J}_n$, but this equality usually fails for quasilikelihood. As in the theory of estimating equations, the parameter Θ is determined by the equation $E\{\mathbf{U}_i(\Theta)\} = 0$ for all i (conditionally unbiased), or possibly through the weaker constraint that $\sum_{i=1}^n E\{\mathbf{U}_i(\Theta)\} = 0$ (unbiased).

We partition $\Theta = (\gamma^t, \eta^t)^t$, where γ is the p-dimensional parameter vector of interest and η is the vector of nuisance parameters. Partition $\mathcal{H}$, omitting the subscript n for ease of notation, similarly as

$$\mathcal{H} = \begin{pmatrix} \mathcal{H}_{\gamma\gamma} & \mathcal{H}_{\gamma\eta} \\ \mathcal{H}_{\eta\gamma} & \mathcal{H}_{\eta\eta} \end{pmatrix},$$

and define $\mathcal{H}_{\gamma\gamma\cdot\eta} = \mathcal{H}_{\gamma\gamma} - \mathcal{H}_{\gamma\eta}\mathcal{H}_{\eta\eta}^{-1}\mathcal{H}_{\eta\gamma}$.

Let $\widehat{\Theta}_0 = (\gamma_0^t, \widehat{\eta}_0^t)^t$ denote the maximum quasilikelihood estimate subject to $\gamma = \gamma_0$. We need the large sample distribution of the log quasilikelihood ratio,

$$\mathcal{L}(\gamma_0) = 2\{\ell(\widehat{\Theta}) - \ell(\widehat{\Theta}_0)\}.$$

The following result is well known under various regularity conditions. For the basic idea of the proof, see Kent (1982).

Theorem: *If $\gamma = \gamma_0$, then, as the number of independent observations increases, $\mathcal{L}(\gamma_0)$ converges in distribution to $\sum_{k=1}^p \lambda_k W_k$, where $W_1, ..., W_p$ are independently distributed as χ_1^2, and $\lambda_1, ..., \lambda_p$ are the eigenvalues of $\mathcal{H}_{\gamma\gamma\cdot\eta}(\mathcal{H}^{-1}\mathcal{J}\mathcal{H}^{-1})_{\gamma\gamma}$.*

To use this result in practice, either to perform a quasilikelihood ratio test of $H_0 : \gamma = \gamma_0$ or to compute a quasilikelihood confidence set for γ_0, we need to estimate the matrices $\mathcal{H}$ and $\mathcal{J}$. If all data are independent, an obvious approach is to replace the theoretical expectations in (A.29) and (A.30) with the analogous empirical averages.

We also need to compute quantiles of the distribution of $\sum_k \widehat{\lambda}_k W_k$. Observe that if $p = 1$, the appropriate distribution is simply a scaled χ_1^2 distribution. If $p > 1$, then algorithms given by Marazzi (1980) and Griffiths and Hill (1985) may be used. A quick and simple way to do the computation is to simulate from the distribution of $\sum_k \widehat{\lambda}_k W_k$, since chi-squared random variables are easy to generate.

A.6.4 Unbiased, but Conditionally Biased, Estimating Equations

It is possible to relax (A.20) to

$$0 = \textstyle\sum_{i=1}^n E\Big\{\Psi_i(\widetilde{\mathbf{Y}}_i, \widehat{\Theta})\Big\},$$

and then the estimating function and estimating equation are *not* conditionally unbiased, but are still said to be *unbiased*. The theory of conditionally unbiased estimating equations carries over without change to estimating equations that are merely unbiased.

A.6.5 Biased Estimating Equations

The estimation methods described in Chapters 4, 5, and 6 are approximately consistent, in the sense that they consistently estimate a value that closely approximates the true parameter. These estimators are formed by estimating equations such as (A.19), but the estimating functions are *not* unbiased for the true parameter Θ. Usually there exists Θ_*, which is close to Θ and which solves

$$0 = \textstyle\sum_{i=1}^n E\left\{\Psi_i(\widetilde{\mathbf{Y}}_i, \Theta_*)\right\}. \tag{A.31}$$

In such cases, $\widehat{\Theta}$ is *still* asymptotically normally distributed, but with mean Θ_* instead of mean Θ. In fact, the theory of Section A.6.4 is applicable since the equations are unbiased for Θ_*. If

$$0 = E\left\{\Psi_i(\widetilde{\mathbf{Y}}_i, \Theta_*)\right\}, \text{ for } i = 1, ..., n,$$

then the the estimating functions are conditionally unbiased for Θ_* and the sandwich method yields asymptotically correct standard error estimators.

A.6.6 Stacking Estimating Equations: Using Prior Estimates of Some Parameters

To estimate the regression parameter, $\mathcal{B}$, in a measurement error model, one often uses the estimates of the measurement error parameters, α, obtained from another data set. How does uncertainty about the measurement error parameters affect the accuracy of the estimated regression parameter? In this subsection, we develop the theory to answer this question. The fact that such complicated estimating schemes can be easily analyzed by the theory of estimating equations further illustrates the power of this theory.

We work generally in that α and $\mathcal{B}$ can be any parameter vectors in a statistical model, and we assume that both $\widehat{\alpha}$ and $\widehat{\mathcal{B}}$ are M-estimators. Suppose that that $\widehat{\alpha}$ solves the estimating equation

$$0 = \textstyle\sum_{i=1}^n \phi_i(\widetilde{\mathbf{Y}}_i, \alpha), \tag{A.32}$$

and $\widehat{\mathcal{B}}$ solves

$$0 = \textstyle\sum_{i=1}^n \Psi_i(\widetilde{\mathbf{Y}}_i, \mathcal{B}, \widehat{\alpha}), \tag{A.33}$$

with $\widehat{\alpha}$ in (A.33) fixed at the solution to (A.32). The estimating functions in (A.32) and (A.33) are assumed to be conditionally unbiased. Since $(\widehat{\alpha}, \widehat{\mathcal{B}})$ solves (A.32) and (A.33) simultaneously, the asymptotic distribution of $(\widehat{\alpha}, \widehat{\mathcal{B}})$ can be found by stacking (A.32) and (A.33) into a single estimating equation:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = n^{-1}\textstyle\sum_{i=1}^n \begin{pmatrix} \phi_i(\widetilde{\mathbf{Y}}_i, \alpha) \\ \Psi_i(\widetilde{\mathbf{Y}}_i, \mathcal{B}, \alpha) \end{pmatrix}. \tag{A.34}$$

One then applies the usual theory to (A.34). Partition $A_n = A_n(\Theta)$, $B_n = B_n(\Theta)$, and $A_n^{-1} B_n A_n^{-t}$ according to the dimensions of α and $\mathcal{B}$. Then the asymptotic variance of $\widehat{\mathcal{B}}$ is n^{-1} times the lower right submatrix of $A_n^{-1} B_n A_n^{-t}$. After some algebra, one gets

$$\begin{aligned} \text{var}(\widehat{\mathcal{B}}) \approx n^{-1} A_{n,22}^{-1} \Big\{ & B_{n,22} - A_{n,21} A_{n,11}^{-1} B_{n,12} \\ & - B_{n,12}^t A_{n,11}^{-t} A_{n,21}^t + A_{n,21} A_{n,11}^{-1} B_{n,11} A_{n,11}^{-t} A_{n,21}^t \Big\} A_{n,22}^{-t}, \end{aligned}$$

where

$$\begin{aligned} A_{n,11} &= \textstyle\sum_{i=1}^n E\Big\{\frac{\partial}{\partial \alpha^t} \phi_i(\widetilde{\mathbf{Y}}_i, \alpha)\Big\}, \\ A_{n,21} &= \textstyle\sum_{i=1}^n E\Big\{\frac{\partial}{\partial \alpha^t} \Psi_i(\widetilde{\mathbf{Y}}_i, \mathcal{B}, \alpha)\Big\}, \\ A_{n,22} &= \textstyle\sum_{i=1}^n E\Big\{\frac{\partial}{\partial \mathcal{B}^t} \Psi_i(\widetilde{\mathbf{Y}}_i, \mathcal{B}, \alpha)\Big\}, \\ B_{n,11} &= \textstyle\sum_{i=1}^n \phi_i(\widetilde{\mathbf{Y}}_i, \alpha) \phi_i^t(\widetilde{\mathbf{Y}}_i, \alpha), \\ B_{n,12} &= \textstyle\sum_{i=1}^n \phi_i(\widetilde{\mathbf{Y}}_i, \alpha) \Psi_i^t(\widetilde{\mathbf{Y}}_i, \alpha, \mathcal{B}), \quad \text{and} \\ B_{n,22} &= \textstyle\sum_{i=1}^n \Psi_i(\widetilde{\mathbf{Y}}_i, \alpha, \mathcal{B}) \Psi_i^t(\widetilde{\mathbf{Y}}_i, \alpha, \mathcal{B}). \end{aligned}$$

As usual, the components of A_n and B_n can be estimated by model-based expectations or by the sandwich method.

A.7 Quasilikelihood and Variance Function Models (QVF)

A.7.1 General Ideas

In the case of no measurement error, Carroll and Ruppert (1988) described estimation based upon the mean and variance functions of the observed data, that is, the conditional mean and variance of $\mathbf{Y}$ as functions of $(\mathbf{Z}, \mathbf{X})$. We will call these QVF methods, for quasilikelihood and variance functions. The models include the important class of generalized linear models (McCullagh and Nelder, 1989; Section A.8 of this monograph), and in particular linear, logistic, Poisson, and gamma regression. QVF estimation is an important special case of estimating equations.

The typical regression model is a specification of the relationship between the mean of a response $\mathbf{Y}$ and the predictors $(\mathbf{Z}, \mathbf{X})$:

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}), \tag{A.35}$$

where $m_{\mathbf{Y}}(\cdot)$ is the *mean function* and $\mathcal{B}$ is the *regression parameter*. Generally, specification of the model is incomplete without an accompanying model for the variances,

$$\operatorname{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta), \tag{A.36}$$

where $g(\cdot)$ is called the *variance function* and θ is called the *variance function parameter*. We find it convenient in (A.36) to separate the variance parameters into the scale factor σ^2 and θ, which determines the possible heteroscedasticity.

The combination of (A.35) and (A.36) includes many important special cases, among them:

- Homoscedastic linear and nonlinear regression, with $g(z, x, \mathcal{B}, \theta) \equiv 1$. For linear regression, $m_{\mathbf{Y}}(z, x, \mathcal{B}) = \beta_0 + \beta_x^t x + \beta_z^t z$.
- Generalized linear models, including Poisson and gamma regression, with

  $$g(z, x, \mathcal{B}, \theta) = m_{\mathbf{Y}}{}^{\theta}(z, x, \mathcal{B})$$

  for some parameter θ. For example, $\theta = 1/2$ for Poisson regression, while $\theta = 1$ for gamma and lognormal models.
- Logistic regression, where $m_{\mathbf{Y}}(z, x, \mathcal{B}) = H(\beta_0 + \beta_x^t x + \beta_z^t z)$, $H(v) = 1/\{1 + \exp(-v)\}$, and since $\mathbf{Y}$ is Bernoulli distributed, $g^2 = m_{\mathbf{Y}}(1 - m_{\mathbf{Y}})$, $\sigma^2 = 1$, and there is no parameter θ.

Model (A.35)–(A.36) includes examples from fields including epidemiology, econometrics, fisheries research, quality control, pharmacokinetics, assay development, etc. See Carroll and Ruppert (1988, Chapters 2-4) for more details.

A.7.2 Estimation and Inference for QVF Models

Specification of only the mean and variance models (A.35)–(A.36) allows one to construct estimates of the parameters $(\mathcal{B}, \theta)$. No further detailed distributional assumptions are necessary. Given θ, $\mathcal{B}$ can be estimated by generalized (weighted) least squares (GLS), a term often now referred to as quasilikelihood estimation. The *conditionally unbiased estimating function* for estimating $\mathcal{B}$ by GLS is

$$\frac{\mathbf{Y} - m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B})}{\sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta)} m_{\mathbf{Y}\mathcal{B}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}), \tag{A.37}$$

where

$$f_{\mathcal{B}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = \frac{\partial}{\partial \mathcal{B}} f(\mathbf{Z}, \mathbf{X}, \mathcal{B})$$

is the vector of partial derivatives of the mean function. The *conditionally unbiased estimating equation* for $\mathcal{B}$ is the sum of (A.37) over the observed data.

To understand why (A.37) is the GLS estimating function, note that the nonlinear least squares (LS) estimator, which minimizes

$$\textstyle\sum_{i=1}^n \{\mathbf{Y} - m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B})\}^2,$$

solves

$$\textstyle\sum_{i=1}^n \{\mathbf{Y} - m_{\mathbf{Y}}(\mathbf{Z}, \mathbf{X}, \mathcal{B})\} m_{\mathbf{Y}\mathcal{B}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = 0. \tag{A.38}$$

The LS estimator is inefficient and can be improved by weighting the summands in (A.38) by reciprocal variances; the result is (A.37).

There are many methods for estimating θ. These may be based on true replicates if they exist, or on functions of squared residuals. These methods are reviewed in Chapters 3 and 6 of Carroll and Ruppert (1988); see also Davidian and Carroll (1987) and Rudemo et al. (1989). Let $(\cdot)$ stand for the argument $(\mathbf{Z}, \mathbf{X}, \mathcal{B})$. If we define

$$\mathbf{R}(\mathbf{Y}, \cdot, \theta, \sigma) = \{\mathbf{Y} - m_{\mathbf{Y}}(\cdot)\} / \{\sigma g(\cdot, \theta)\}, \tag{A.39}$$

then one such (approximately) conditionally unbiased score function for θ (and σ) given $\mathcal{B}$ is

$$\left\{\mathbf{R}^2(\mathbf{Y}, \cdot, \theta, \sigma) - \frac{n - \dim(\mathcal{B})}{n}\right\} \frac{\partial}{\partial\,(\sigma, \theta)^t} \log\{\sigma g(\cdot, \theta)\}, \tag{A.40}$$

where $\dim(\mathcal{B})$ is the number of components of the vector $\mathcal{B}$. The (approximately) *conditionally unbiased estimating equation* for θ and σ is the sum of (A.40) over the observed data. The resulting M-estimator is closely related to the REML estimator used in variance components modeling; see Searle, Casella, and McCulloch (1992).

As described by Carroll and Ruppert (1988), (A.37)–(A.40) are weighted least squares estimating equations, and nonlinear regression algorithms can be used to estimate the parameters.

There are two specific types of covariance estimates, depending on whether or not one believes that the variance model has been approximately correctly specified. We concentrate here on inference for the regression parameter $\mathcal{B}$, referring the reader to Chapter 3 of Carroll and Ruppert (1988) for variance parameter inference. Based on a sample of size n, $\widehat{\mathcal{B}}$ is generally asymptotically normally distributed with mean $\mathcal{B}$ and covariance matrix $n^{-1}A_n^{-1}B_nA_n^{-1}$, where if $(\cdot)$ stands for $(\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})$,

$$\begin{aligned} A_n &= n^{-1}\textstyle\sum_{i=1}^n \{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}\{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}^t \{\sigma^2 g^2(\cdot,\theta)\}^{-1}; \\ B_n &= n^{-1}\textstyle\sum_{i=1}^n \{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}\{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}^t \dfrac{E\{\mathbf{Y}_i - m_{\mathbf{Y}}(\cdot)\}^2}{\sigma^4 g^4(\cdot,\theta)}. \end{aligned}$$

The matrix B_n in this expression is the same as (A.22) in the general theory of unbiased estimating equations. The matrix A_n is the same as (A.23), but it is simplified somewhat by using the fact that $E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = f(\mathbf{Z},\mathbf{X},\mathcal{B})$.

If the variance model is correct, then $E\{\mathbf{Y}_i - m_{\mathbf{Y}}(\mathbf{Z}_i,\mathbf{X}_i,\mathcal{B})\}^2 = \sigma^2 g^2(\mathbf{Z}_i,\ \mathbf{X}_i,\mathcal{B},\theta)$, $A_n = B_n$ and an asymptotically correct covariance matrix is $n^{-1}\widehat{A}_n^{-1}$, where $(\cdot)$ stands for $(\mathbf{Z}_i, \mathbf{X}_i, \widehat{\mathcal{B}})$ and

$$\widehat{A}_n = n^{-1}\textstyle\sum_{i=1}^n \{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}\{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}^t \left\{\widehat{\sigma}^2 g^2(\cdot,\widehat{\theta})\right\}^{-1}.$$

If one has severe doubts about the variance model, one can use the sandwich method to estimate $E\{\mathbf{Y}_i - m_{\mathbf{Y}}(\cdot)\}^2$, leading to the covariance matrix estimate $\widehat{A}_n^{-1}\widehat{B}_n\widehat{A}_n^{-1}$, where

$$\widehat{B}_n = n^{-1}\textstyle\sum_{i=1}^n \{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}\{m_{\mathbf{Y}\mathcal{B}}(\cdot)\}^t \dfrac{\{\mathbf{Y}_i - m_{\mathbf{Y}}(\cdot)\}^2}{\widehat{\sigma}^4 g^4(\cdot,\widehat{\theta})}.$$

In some situations, the method of Section A.6.3 can be used in place of the sandwich method.

With a flexible variance model that seems to fit the data fairly well, we prefer the covariance matrix estimate $n^{-1}\widehat{A}_n^{-1}$, because it can be much less variable than the sandwich estimator. Drum and McCullagh (1993) basically come to the same conclusion, stating that "unless there is good reason to believe that the assumed variance function is substantially incorrect, the model-based estimator seems to be preferable in applied work." Moreover, if the assumed variance function is clearly inadequate, most statisticians would find a better variance model and then use $n^{-1}\widehat{A}_n^{-1}$ with the better-fitting model.

In addition to formal fitting methods, simple graphical displays exist to evaluate the models (A.35)–(A.36). Ordinary and weighted residual plots with smoothing can be used to understand departures from the assumed mean function, while absolute residual plots can be used to detect deviations from the assumed variance function. These graphical techniques are discussed in Chapter 2, section 7 of Carroll and Ruppert (1988).

A.8 Generalized Linear Models

Exponential families have density or mass function

$$f(y|\xi) = \exp\left\{\frac{y\xi - \mathcal{C}(\xi)}{\phi} + c(y,\phi)\right\}. \tag{A.41}$$

With superscripted (j) referring to the jth derivative, the mean and variance of $\mathbf{Y}$ are $\mu = \mathcal{C}'(\xi)$ and $\phi\mathcal{C}''(\xi)$, respectively. See, for example, McCullagh and Nelder (1989).

If ξ is a function of a linear combination of predictors, say $\xi = \mathbf{X}_i\,(\eta)$ where $\eta = (\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z})$, then we have a generalized linear model. Generalized linear models include many of the common regression models, for example, normal, logistic, Poisson, and gamma. Consideration of specific models is discussed in detail in Chapter 7. Generalized linear models are mean and variance models in the observed data, and can be fit using QVF methods.

If we define $L = (\mathcal{C}' \circ \mathbf{X}_i)^{-1}$, then $L(\mu) = \eta$; L is called the *link* function since it links the mean of the response and the linear predictor, η. If $\mathbf{X}_i$ is the identity function, then we say that the model is canonical; this implies that $L = (\mathcal{C}')^{-1}$, which is called the *canonical link function*. The link function L, or equivalently $\mathbf{X}_i$, should be chosen so that the model fits the data as well as possible. However, if the canonical link function fits reasonably well, then it is typically used, because doing so simplifies the analysis.

A.9 Bootstrap Methods

A.9.1 Introduction

The bootstrap is a widely used tool for analyzing the sampling variability of complex statistical methods. The basic idea is quite simple. One creates simulated data sets, called bootstrap data sets, whose distribution is equal to an estimate of the probability distribution of the actual data. Any statistical method that is applied to the actual data can also be applied to the bootstrap data sets. Thus, the empirical distribution of an estimator or test statistic across the bootstrap data sets can be used to estimate the actual sampling distribution of that statistic.

For example, suppose that $\widehat{\Theta}$ is obtained by applying some estimator to the actual data, and $\widehat{\Theta}^{(m)}$ is obtained by applying the same estimator to the m^{th} bootstrap data set, $m = 1, ..., M$, where M is the number of bootstrap data sets that we generate, and let $\bar{\Theta}$ be the average of $\widehat{\Theta}', ..., \widehat{\Theta}^{(m)}$. Then, the covariance matrix of $\widehat{\Theta}$ can be estimated by

$$\widehat{\mathrm{var}}(\widehat{\Theta}) = (M-1)^{-1} \sum_{m=1}^{M} \left(\widehat{\Theta}^{(m)} - \bar{\Theta}\right)\left(\widehat{\Theta}^{(m)} - \bar{\Theta}\right)^{t}. \tag{A.42}$$

Despite this underlying simplicity, implementation of the bootstrap can be a complex, albeit fascinating subject. There are many ways to estimate the probability distribution of the data, and it is not always obvious which way is most appropriate. Bootstrap standard errors are easily found from (A.42), and these can be plugged into (A.18) to get "normal theory" confidence intervals. However, these simple confidence intervals are not particularly accurate, and several improved bootstrap intervals have been developed. Comparing bootstrap standard errors and confidence intervals with traditional methods and comparing the various bootstrap intervals with each other requires the powerful methodology of Edgeworth expansions. Efron and Tibshirani (1993) give an excellent, comprehensive account of bootstrapping theory and applications. For more mathematical theory, including Edgeworth expansions, see Hall (1992). Here we give enough background so that the reader can understand how the bootstrap is applied to obtain standard errors in the examples.

A.9.2 Nonlinear Regression without Measurement Error

To illustrate the basic principles of bootstrapping, we start with nonlinear regression without measurement error. Suppose that $\mathbf{Y}_i = m_{\mathbf{Y}}(\mathbf{Z}_i, \mathcal{B}) + \epsilon_i$, where the $\mathbf{Z}_i$ are, as usual, covariates measured without error, and the ϵ_i's are independent with the density of ϵ_i possibly depending on $\mathbf{Z}_i$. There are at least three distinct methods for creating the bootstrap data sets. Efron and Tibshirani (1993) call the first two methods *resampling pairs* and *resampling residuals*. The third method is a form of the *parametric bootstrap*.

A.9.2.1 Resampling Pairs

Resampling pairs means forming a bootstrap data set by sampling at random *with replacement* from $\{(\mathbf{Y}_i, \mathbf{Z}_i)\}_i^n$. The advantage of this method is that it requires minimal assumptions. If ϵ_i has a distribution depending on $\mathbf{Z}_i$ in the real data, then this dependence is captured by the resampling, since the $(\mathbf{Y}_i, \mathbf{Z}_i)$ pairs are never broken during the resampling. Therefore, standard errors and confidence intervals produced by this type of bootstrapping will be asymptotically valid in the presence of heteroscedasticity or other forms on nonhomogeneity. Besides this type of robustness, another advantage of resampling pairs is that it is easy to extend to more complex situations, such as measurement error models.

The disadvantage of resampling pairs is that the bootstrap data sets will have different sets of $\mathbf{Z}_i$'s than the original data. For example, if there is a high leverage point in the original data, it may appear several times or not at all in a given bootstrap data set. Therefore, this form of the bootstrapping estimates unconditional sampling distributions, not sampling distributions conditional on the $\mathbf{Z}_i$'s. Some statisticians will object to this, asking, "Even if the $\mathbf{Z}_i$'s are random, why should I care that I might have gotten different $\mathbf{Z}_i$'s than I did? I know the values of the $\mathbf{Z}_i$'s that I got, and I want to condition upon them." We feel that this objection is valid. However, as Efron and Tibshirani (1993) point out, often conditional and unconditional standard errors are nearly equal.

A.9.2.2 Resampling Residuals

The purpose behind resampling residuals is to condition upon the $\mathbf{Z}_i$'s. The ith residual is $e_i = \mathbf{Y}_i - m_{\mathbf{Y}}(\mathbf{Z}_i, \widehat{\mathcal{B}})$, where $\widehat{\mathcal{B}}$ is, say, the nonlinear least squares estimate. To create the m^{th} bootstrap data set we first center the residuals by subtracting their sample mean, $\bar{e}$, and then draw $\{e_i^{(m)}\}_{i=1}^n$ randomly, with replacement, from $\{(e_i - \bar{e})\}_i^n$. Then we let $Y_i^{(m)} = m_{\mathbf{Y}}(\mathbf{Z}_i, \widehat{\mathcal{B}}) + e_i^{(m)}$. The m^{th} bootstrap data set is $\{(Y_i^{(m)}, \mathbf{Z}_i)\}_{i=1}^n$. Notice that the bootstrap data sets have the same set of $\mathbf{Z}_i$'s as the original data, so that bootstrap sampling distributions are conditional on the $\mathbf{Z}_i$'s. By design, the distribution of the ith "error" in a bootstrap data set is independent of $\mathbf{Z}_i$. Therefore, resampling residuals is only appropriate when the ϵ_i's in the actual data are identically distributed, and is particularly sensitive to the homoscedasticity assumption.

A.9.2.3 The Parametric Bootstrap

The parametric bootstrap can be used when we assume a parametric model for the ϵ_i's. Let f be a known mean-zero density, say the standard normal density, ϕ. Assume that the density of ϵ_i is in the scale family $m_{\mathbf{Y}}(\cdot/\sigma)/\sigma$, $\sigma > 0$, and let $\widehat{\sigma}$ be a consistent estimator of σ, say the residual root-mean square if f is equal to ϕ. Then, as when resampling residuals, the bootstrap data sets are $\{(\mathbf{Y}_i^{(m)}, \mathbf{Z}_i)\}_i^n$, where $\mathbf{Y}_i = m_{\mathbf{Y}}(\mathbf{Z}_i, \widehat{\mathcal{B}}) + e_i^{(m)}$, but now the $\epsilon_i^{(m)}$s are, conditional on the observed data, iid from $f(\cdot/\widehat{\sigma})/\widehat{\sigma}$. Like resampling residuals, the parametric bootstrap estimates sampling distributions that are conditional on the $\mathbf{Z}_i$'s and requires that the ϵ_i's be independent of the $\mathbf{Z}_i$'s. In addition, like other parametric statistical methods, the parametric bootstrap is more efficient when the parametric assumptions are met, but possibly biased otherwise.

A.9.3 Bootstrapping Heteroscedastic Regression Models

Consider the QVF model

$$\mathbf{Y}_i = m_{\mathbf{Y}}(\mathbf{Z}_i, \mathcal{B}) + \sigma g(\mathbf{Z}_i, \mathcal{B}, \theta)\epsilon_i,$$

where the ϵ_i's are iid. The assumption of iid errors holds when $\mathbf{Y}_i$ given $\mathbf{Z}_i$ is normal, but this assumption precludes logistic, Poisson, and gamma regression, for example. This model can be fit by the methods of section A.7.2. To estimate the sampling distribution of the QVF estimators, bootstrap data sets can be formed by resampling from the set of pairs $\{(\mathbf{Y}_i, \mathbf{Z}_i)\}_i^n$, as discussed for nonlinear regression models in Section A.9.2.

Resampling residual requires some reasonably obvious changes from Section A.9.2. First, define the ith residual to be

$$e_i = \frac{\mathbf{Y}_i - m_{\mathbf{Y}}(\mathbf{Z}_i, \widehat{\mathcal{B}})}{\widehat{\sigma} g(\mathbf{Z}_i, \widehat{\mathcal{B}}, \widehat{\theta})} - \bar{e},$$

where $\bar{e}$ is defined so that the e_i's sum to 0. To form m^{th} bootstrap data set, let $\{e_i^{(m)}\}_{i=1}^n$ be sampled with replacement from the residuals and then let

$$\mathbf{Y}_i^{(m)} = m_{\mathbf{Y}}(\mathbf{Z}_i, \widehat{\mathcal{B}}) + \widehat{\sigma} g(\mathbf{Z}_i, \widehat{\mathcal{B}}, \widehat{\theta}) e_i^{(m)}.$$

Note that $e_i^{(m)}$ is not the residual from the ith of the original observations, but is equally likely to be any of the n residuals from the original observations. See Carroll and Ruppert (1991) for further discussion of bootstrapping heteroscedastic regression models, with application to prediction and tolerance intervals for the response.

A.9.4 Bootstrapping Logistic Regression Models

Consider the logistic regression model without measurement error,

$$\text{pr}(\mathbf{Y}_i = 1 | \mathbf{Z}_i) = H(\beta_0 + \beta_z^T \mathbf{Z}_i),$$

where, as elsewhere in this book, $H(v) = \{1 + \exp(-v)\}^{-1}$. The general purpose technique of resampling pairs works here, of course. Resampling residuals is not applicable, since the residuals will have skewness depending on $\mathbf{Z}_i$ so are not homogeneous even after weighting as in Section A.9.3. The parametric bootstrap, however, is easy to implement. To form the m^{th} data set, fix the $\mathbf{Z}_i$'s equal to their values in the real data and let $\mathbf{Y}_i^{(m)}$ be Bernoulli with

$$\text{pr}(\mathbf{Y}_i^{(m)} = 1 | \mathbf{Z}_i) = H(\widehat{\beta}_0 + \widehat{\beta}_z^t \mathbf{Z}_i).$$

A.9.5 Bootstrapping Measurement Error Models

In a measurement error problem, a typical data vector consists of $\mathbf{Z}_i$ and a subset of the following data: the response $\mathbf{Y}_i$, the true covariates $\mathbf{X}_i$, $\{\mathbf{w}_{i,j} : j = 1, ..., k_i\}$ which are replicate surrogates for $\mathbf{X}_i$, and a second surrogate $\mathbf{T}_i$. We divide the total collection of data into homogeneous data sets, which have the same variables measured on each observation and are from a common source, for example, primary data, internal replication data, external replication data, and internal validation data.

The method of "resampling pairs" ignores the various data subsets, and can often be successful (Efron, 1994). Taking into account the data subsets is better called "resampling vectors," and consists of resampling, with replacement, independently from each of the homogeneous data sets. This ensures that each bootstrap data set has the same amount of validation data, data with two replicates of $\mathbf{w}$, data with three replications, etc., as the actual data set. Although in principle we wish to condition on the $\mathbf{Z}_i$'s and resampling vectors does not do this, resampling vectors is a useful expedient and allows us to bootstrap any collection of data sets with minimal assumptions. In the examples in this monograph, we have reported the "resampling pairs" bootstrap analyses, but because of the large sample sizes, the reported results do not differ substantially from the "resampling vectors" bootstrap.

Resampling residuals is applicable to validation data when there are two regression models: one for $\mathbf{Y}_i$ given $(\mathbf{Z}_i, \mathbf{X}_i)$ and another for $\mathbf{w}_i$ given $(\mathbf{Z}_i, \mathbf{X}_i)$. One fits both models and resamples residuals from the first to create the bootstrap $\mathbf{Y}_i^{(m)}$'s and from the second to create the $\mathbf{w}_i^{(m)}$'s. This method generates sampling distributions that are conditional on the observed $(\mathbf{Z}_i, \mathbf{X}_i)$'s.

The parametric bootstrap can be used when the response, given the observed covariates, has a distribution in a known parametric family. For example, suppose one has a logistic regression model with internal validation data. One can fix the $(\mathbf{Z}_i, \mathbf{X}_i, \mathbf{w}_i)$ vectors of the validation data and create bootstrap responses as in Section A.9.4, using $(\mathbf{Z}_i, \mathbf{X}_i)$ in place of $\mathbf{Z}_i$. Because $\mathbf{w}_i$ is a surrogate, it is not used to create the bootstrap responses of validation data. For the nonvalidation data, one fixes the $(\mathbf{Z}_i, \mathbf{w}_i)$ vectors. Using regression calibration as described in Chapter 4, one fits an approximate logistic model for $\mathbf{Y}_i$ given $(\mathbf{Z}_i, \mathbf{w}_i)$ and again creates bootstrap responses distributed according to the fitted model. The bootstrap sampling distributions generated in this way are conditional on all observed covariates.

A.9.6 Bootstrap Confidence Intervals

As in Section A.5.4, let $\Theta^t = (\theta_1, \Theta_2^t)$, where θ_1 is univariate, and suppose that we want a confidence interval for θ_1. The simplest bootstrap confidence interval is "normal based." The bootstrap covariance matrix in (A.42) is used for a standard error

$$\mathrm{se}(\widehat{\theta}_1) = \sqrt{\widehat{\mathrm{var}}(\widehat{\Theta})_{11}}.$$

This standard error is then plugged into (A.18), giving

$$\widehat{\theta}_1 \pm \Phi^{-1}(1-\alpha/2)\mathrm{se}(\widehat{\theta}_1). \tag{A.43}$$

The so-called percentile methods replace the normal approximation in (A.43) by percentiles of the empirical distribution of $\{(\widehat{\theta}_1^{(m)} - \widehat{\theta}_1)\}_1^M$. The best of these percentile methods are the so-called BC$_a$ and ABC intervals, and they are generally more accurate than (A.43) in the sense of having a true coverage probability closer to the nominal $(1-\alpha)$; see Efron and Tibshirani (1993) for a full description of these intervals.

Hall (1992) has stressed the advantages of bootstrapping an asymptotically pivotal quantity, that is, a quantity whose asymptotic distribution is independent of unknown parameters. The percentile-t methods used the "studentized" quantity

$$t = \frac{\widehat{\theta}_1 - \theta_1}{\mathrm{se}(\widehat{\theta}_1)}, \tag{A.44}$$

which is an asymptotic pivot with a large-sample standard normal distribution for all values of θ. Let $\mathrm{se}^{(m)}(\widehat{\theta}_1)$ be the standard error of $\widehat{\theta}_1$ computed from the m^{th} bootstrap data set and let

$$t^{(m)} = \frac{\widehat{\theta}_1^{(m)} - \widehat{\theta}_1}{\mathrm{se}^{(m)}(\widehat{\theta}_1)}.$$

Typically, $\mathrm{se}^{(m)}(\widehat{\theta}_1)$ will come from an expression for the asymptotic variance matrix of $\widehat{\Theta}$ (for example, the inverse of the observed Fisher information matrix given by (A.17)) rather than bootstrapping, since the latter would require two levels of bootstrapping: an outer level for $\{t^{(m)}\}_1^M$ and, for each m, an inner level for calculating the denominator of $t^{(m)}$. This would be very computationally expensive, especially for the nonlinear estimators in this monograph. Let $t_{1-\alpha}$ be the $(1-\alpha)$ quantile of $\{|t^{(m)}|\}_1^M$. Then the symmetric percentile-t confidence interval is

$$\widehat{\theta}_1 \pm \mathrm{se}(\widehat{\theta}_1)\, t_{1-\alpha}. \tag{A.45}$$

Note that $\mathrm{se}(\widehat{\theta}_1)$ is calculated from the original data in the same way that $\mathrm{se}^{(m)}(\widehat{\theta}_1)$ is calculated from the m^{th} bootstrap data set.

APPENDIX B

TECHNICAL DETAILS

This Appendix is a collection of technical details that will interest some, but certainly not all, readers.

B.1 Appendix to Chapter 1: Power in Berkson and Classical Error Models

In Chapter 1, we emphasized why it is important not to mix up Berkson and classical error models when calculating power. In this section, we discuss what is common between these two error models with regard to power.

In the normal linear model with $(\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{U})$ jointly normal, if $\mathbf{W}$ is a surrogate (nondifferential error model), then power is a function of the measurement error model only via the squared correlation between $\mathbf{W}$ and $\mathbf{X}$, $\rho_{xw}^2 = \sigma_x^2/\sigma_w^2$ (classical) or $\rho_{xw}^2 = \sigma_w^2/\sigma_x^2$ (Berkson), and thus is the same for classical and Berkson error models with equal correlations. So whether the error model is Berkson or classical, the loss of power is the same provided ρ_{xw}^2 is the same in the two models. This is also true of noncalibrated measurements, that is, ones with biases and multiplicative constants (for example, $\mathbf{W} = \alpha_0 + \alpha_x\mathbf{X} + \mathbf{U}$). So when comparing the effects of measurement error on loss of power across different types of measurement error models, the squared correlation is the most relevant quantity on which to focus the discussion. Looking at variances is not very meaningful because var$(\mathbf{U})$ is not comparable across error models, and it depends on whether or not we have calibrated measurements, but correlations are comparable across error types. See Buzas, Tosteson, and Stefanski (2004) for further discussion.

The explanation of the effects of measurement error on power in terms of correlations has some useful implications for designing studies. The most useful of which is that one should try to make an educated guess (or estimate) of the squared correlation between $\mathbf{W}$ and $\mathbf{X}$, as this is the most relevant quantity.

This discussion also shows why assuming a Berkson model when a classical model holds can lead to an inflated estimate of power. Suppose we have an estimate of σ_u^2 and estimate σ_w^2 from the observed $\mathbf{W}_i$. Define $f(x) = (x - \sigma_u^2)/x = 1 - \sigma_u^2/x$, $x > 0$, and note that f is strictly increasing. If one assumes Berkson error, then ρ_{xy}^2 is estimated by $f(\sigma_w^2 + \sigma_u^2)$. If classical errors are assumed, then ρ_{xy}^2 is estimated by $f(\sigma_w^2)$. Since f is strictly increasing, for fixed estimates of σ_u^2 and σ_w^2, the estimate of ρ_{xy}^2 is larger for the Berkson error model than for the classical error model. The point here is that although the power would be the same for the two models if they had the same values of ρ_{xy}^2, for fixed values of σ_u^2 and σ_w^2, ρ_{xy}^2 is larger when Berkson errors are assumed.

B.2 Appendix to Chapter 3: Linear Regression and Attenuation

Here we establish (3.10) and (3.11) under the assumption of multivariate normality. Taking expectations of both sides of (3.9) conditional on $(\mathbf{X}, \mathbf{Z})$ leads to the identity

$$E(\mathbf{Y} \mid \mathbf{W}, \mathbf{Z}) = \beta_0 + \beta_x E(\mathbf{X} \mid \mathbf{W}, \mathbf{Z}) + \beta_z^t \mathbf{Z}. \tag{B.1}$$

Under joint normality the regression of $\mathbf{X}$ on $(\mathbf{W}, \mathbf{Z})$ is linear. To facilitate the derivation, we parameterize this as

$$E(\mathbf{X} \mid \mathbf{W}, \mathbf{Z}) = \gamma_0 + \gamma_w \left\{\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})\right\} + \gamma_z^t \left\{\mathbf{Z} - E(\mathbf{Z})\right\}. \tag{B.2}$$

Because of the orthogonalization in (B.2) it is immediate that

$$\begin{aligned} \gamma_w &= \frac{E\left(E\left[\mathbf{X}\left\{\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})\right\} \mid \mathbf{Z}\right]\right)}{E\left(E\left[\left\{\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})\right\}^2 \mid \mathbf{Z}\right]\right)} \\ &= \frac{E\left\{E(\mathbf{XW} \mid \mathbf{Z}) - E(\mathbf{X} \mid \mathbf{Z})E(\mathbf{W} \mid \mathbf{Z})\right\}}{\sigma_{w|z}^2}, \end{aligned} \tag{B.3}$$

where $\sigma_{w|z}^2 = \text{var}(\mathbf{W} \mid \mathbf{Z})$.

Now because $\mathbf{U}$ is independent of $\mathbf{Z}$, $E(\mathbf{W} \mid \mathbf{Z}) = E(\mathbf{X} \mid \mathbf{Z})$, $E(\mathbf{XW} \mid \mathbf{Z}) = E(\mathbf{X}^2 \mid \mathbf{Z})$, and the numerator in (B.3) is just $\sigma_{x|z}^2$. Independence of $\mathbf{U}$ and $\mathbf{Z}$ also implies that $\sigma_{w|z}^2 = \sigma_{x|z}^2 + \sigma_u^2$. It follows that

$$\gamma_w = \frac{\sigma_{x|z}^2}{\sigma_{w|z}^2} = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}, \tag{B.4}$$

as claimed.

Suppose now that $E(\mathbf{X} \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z^t \mathbf{Z}$. As noted previously, $E(\mathbf{W} \mid \mathbf{Z}) = E(\mathbf{X} \mid \mathbf{Z})$, and thus $E(\mathbf{W} \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z^t \mathbf{Z}$ also.

Again, because of the orthogonalization in (B.2), it is immediate that $\gamma_z = \Gamma_z$.

If we now replace $E(\mathbf{W} \mid \mathbf{Z})$ with $\Gamma_0 + \Gamma_z^t \mathbf{Z}$ in (B.2) and substitute the right-hand side of (B.2) into (B.1), and then collect coefficients of $\mathbf{Z}$ using the definition of (B.4), we find that the coefficient of $\mathbf{Z}$ in (B.1) is

$$\beta_{z*}^t = \beta_z^t + \beta_x(1 - \lambda_1)\Gamma_z^t. \tag{B.5}$$

B.3 Appendix to Chapter 4: Regression Calibration

B.3.1 Standard Errors and Replication

As promised in Section 4.6, here we provide formulae for asymptotic standard errors for generalized linear models, wherein

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) &= f(\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}); \\ \text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) &= \sigma^2 g^2(\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}). \end{aligned}$$

Let $f'(\cdot)$ be the derivative of the function $f(\cdot)$, and let $\mathcal{B} = (\beta_0, \beta_x^t, \beta_z^t)^t$.

We will use here the best linear approximations of Section 4.4.2. Let n be the size of the main data set, and $N - n$ the size of any independent data set giving information about the measurement error variance Σ_{uu}. Let $\Delta = 1$ mean that the main data set is used, and $\Delta = 0$ otherwise. Remember that there are k_i replicates for the i^{th} individual and that $\nu = \sum_{i=1}^n k_i - \sum_{i=1}^n k_i^2 / \sum_{i=1}^n k_i$.

Make the definitions $\alpha = (n-1)/\nu$, $\widehat{\Sigma}_{wz} = \widehat{\Sigma}_{xz}$, $\widehat{\Sigma}_{zw} = \widehat{\Sigma}_{wz}^t$, $\widehat{\Sigma}_{ww} = \widehat{\Sigma}_{xx} + \alpha\widehat{\Sigma}_{uu}$, $r_{wi} = (\overline{\mathbf{W}}_{i\cdot} - \mu_w)$, $r_{zi} = (\mathbf{Z}_i - \mu_z)$, and

$$\begin{aligned} \widehat{\mu}_w &= \textstyle\sum_{i=1}^N \Delta_i k_i \overline{\mathbf{W}}_{i\cdot} / \sum_{i=1}^N \Delta_i k_i; \; \widehat{\mu}_z = n^{-1} \sum_{i=1}^N \Delta_i \mathbf{Z}_i; \\ \Psi_{1i*} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & (nk_i/\nu) r_{wi} r_{wi}^t & (nk_i/\nu) r_{wi} r_{zi}^t \\ 0 & (nk_i/\nu) r_{zi} r_{wi}^t & \{n/(n-1)\} r_{zi} r_{zi}^t \end{bmatrix}; \\ \Psi_{1i} &= \Psi_{1i*} - V_i; \\ V_i &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & b_{i1} & b_{i2} \\ 0 & b_{i2}^t & b_{i3} \end{bmatrix}; \\ b_{i1} &= \Sigma_{xx} \frac{nk_i}{\nu} \left\{ 1 - 2k_i / \sum_j \Delta_j k_j + \sum_j \Delta_j k_j^2 / (\sum_j \Delta_j k_j)^2 \right\} \\ &\qquad + \Sigma_{uu}(n/\nu)(1 - k_i / \sum_j \Delta_j k_j); \\ b_{i2} &= \Sigma_{xz}(n/\nu)(k_i - k_i^2 / \sum_j \Delta_j k_j); \quad b_{i3} = \Sigma_{zz}. \end{aligned} \tag{B.6}$$

In what follows, except where explicitly noted, we assume that the data have been centered, so that $\widehat{\mu}_w = 0$ and $\widehat{\mu}_z = 0$. This is accomplished by subtracting the original values of the quantities (B.6) from the $\mathbf{W}$'s and $\mathbf{Z}$'s, and has an effect only on the intercept. Reestimating the intercept after "uncentering" is described at the end of this section.

The analysis requires an estimate of Σ_{uu}. For this we only assume that for some random variables Ψ_{2i} and Ψ_{3i}, if

$$\widehat{S} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \widehat{\Sigma}_{uu} & 0 \\ 0 & 0 & 0 \end{pmatrix}; \qquad S = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Sigma_{uu} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

then

$$\begin{aligned} \widehat{S} - S \quad &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \widehat{\Sigma}_{uu} - \Sigma_{uu} & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &\approx n^{-1}\textstyle\sum_{i=1}^{N} \left\{\Delta_i \Psi_{2i} + (1-\Delta_i)\Psi_{3i}\right\}. \end{aligned} \tag{B.7}$$

For example, if the estimator comes from an independent data set of size $N-n$, then $\Psi_{2i} = 0$ and

$$\Psi_{3i} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \psi_{3i} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ where}$$

$$\psi_{3i} = \frac{\sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t - (k_i - 1)\Sigma_{uu}}{n^{-1}\sum_{l=1}^{N}(1-\Delta_l)(k_l - 1)}.$$

If the estimate of Σ_{uu} comes from internal data, then $\Psi_{3i} = 0$ and

$$\Psi_{2i} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \psi_{2i} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ where}$$

$$\psi_{2i} = \frac{\sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t - (k_i - 1)\Sigma_{uu}}{n^{-1}\sum_{l=1}^{N}\Delta_l(k_l - 1)}.$$

Now make the further definitions

$$\widehat{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \widehat{\Sigma}_{ww} & \widehat{\Sigma}_{wz} \\ 0 & \widehat{\Sigma}_{zw} & \widehat{\Sigma}_{zz} \end{bmatrix};$$

$$\widehat{c}_i = \left\{\widehat{D} - (\alpha - k_i^{-1})\widehat{S}\right\}^{-1}.$$

Let D and S be the limiting values of $\widehat{D}$ and $\widehat{S}$. Let I be the identity matrix of the same dimension as $\mathcal{B}$. Define $R_i = (1, \overline{\mathbf{W}}_{i\cdot}^t, \mathbf{Z}_i^t)^t$ and $\widehat{Q}_i = (\widehat{D} - \alpha\widehat{S})\widehat{c}_i R_i$. Using the fact that the data are centered, it is an easy but crucial calculation to show that $\widehat{Q}_i = (1, \widehat{E}(\mathbf{X}_i^t | \mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot}), \mathbf{Z}_i^t)^t$, that is, it

reproduces the regression calibration estimates. Now make the following series of definitions:

$$\begin{aligned}
\widehat{s}_i &= \left\{f'(\widehat{Q}_i^t\widehat{\mathcal{B}})/g(\widehat{Q}_i^t\widehat{\mathcal{B}})\right\}^2; \\
\widehat{A}_{1n} &= n^{-1}\textstyle\sum_{i=1}^{N}\Delta_i\widehat{Q}_i\widehat{Q}_i^t\widehat{s}_i; \\
r_i &= \left\{Y_i - f(Q_i^t\mathcal{B})\right\} f'(Q_i^t\mathcal{B})Q_i/g^2(Q_i^t\mathcal{B}); \\
d_{in1} &= n^{-1}\sum_{j=1}^{N}\Delta_j s_j Q_j R_j^t c_j \Psi_{1i}\left\{I - c_j(D - \alpha S)\right\}\mathcal{B}; \\
d_{in2} &= n^{-1}\sum_{j=1}^{N}\Delta_j s_j Q_j R_j^t c_j \Psi_{2i}\left\{(\alpha - k_j^{-1})(D - \alpha S)c_j - \alpha I\right\}\mathcal{B}; \\
d_{in3} &= n^{-1}\sum_{j=1}^{N}\Delta_j s_j Q_j R_j^t c_j \Psi_{3i}\left\{(\alpha - k_j^{-1})(D - \alpha S)c_j - \alpha I\right\}\mathcal{B}; \\
e_{in} &= \Delta_i(r_i - d_{in1} - d_{in2}) - (1-\Delta_i)d_{in3}.
\end{aligned}$$

Here and in what follows, s_i, Q_i, c_i, A_{1n}, etc. are obtained by removing the estimates in each of their terms. Similarly, $\widehat{r}_i$, $\widehat{d}_{in1}$, $\widehat{d}_{in2}$, $\widehat{e}_{in}$, etc. are obtained by replacing population quantities by their estimates.

We are going to show that

$$\widehat{\mathcal{B}} - \mathcal{B} \approx A_{1n}^{-1} n^{-1}\textstyle\sum_{i=1}^{N} e_{in}, \tag{B.8}$$

and hence a consistent asymptotic covariance matrix estimate obtained by using the sandwich method is

$$n^{-1}\widehat{A}_{1n}^{-1}\widehat{A}_{2n}\widehat{A}_{1n}^{-1}, \quad \text{where} \tag{B.9}$$

$$\begin{aligned} \widehat{A}_{2n} = n^{-1}\textstyle\sum_{i=1}^{N}\Big\{&\Delta_i\left(\widehat{r}_i - \widehat{d}_{in1} - \widehat{d}_{in2}\right)\left(\widehat{r}_i - \widehat{d}_{in1} - \widehat{d}_{in2}\right)^t \\ &+ (1-\Delta_i)\widehat{d}_{in3}\widehat{d}_{in3}^{\,t}\Big\}. \end{aligned} \tag{B.10}$$

The information-type asymptotic covariance matrix uses

$$\widehat{A}_{2n,i} = \widehat{A}_{2n} + \widehat{A}_{1n} - n^{-1}\textstyle\sum_{i=1}^{N}\Delta_i\widehat{r}_i\widehat{r}_i^t. \tag{B.11}$$

It is worth noting that deriving (B.9) and (B.11) takes considerable effort, and that programming it is not trivial. The bootstrap avoids both steps, at the cost of extra computer time.

To verify (B.8), note by the definition of the quasilikelihood estimator and by a Taylor series, we have the expansion

$$\begin{aligned} 0 &= n^{-1/2}\textstyle\sum_{i=1}^{N}\Delta_i\left\{Y_i - f(\widehat{Q}_i^t\widehat{\mathcal{B}})\right\}f'(\widehat{Q}_i^t\widehat{\mathcal{B}})\widehat{Q}_i/g^2(\widehat{Q}_i^t\widehat{\mathcal{B}}) \\ &\approx n^{-1/2}\textstyle\sum_{i=1}^{N}\Delta_i\left\{r_i - s_iQ_i\left(\widehat{Q}_i^t\widehat{\mathcal{B}} - Q_i^t\mathcal{B}\right)\right\} \end{aligned}$$

$$
\begin{aligned}
&\approx n^{-1/2}{\textstyle\sum_{i=1}^{N}}\Delta_i\left\{r_i - s_iQ_i\left(\widehat{Q}_i - Q_i\right)^t\mathcal{B}\right\} \qquad \text{(B.12)}\\
&\qquad -A_{1n}n^{1/2}\left(\widehat{\mathcal{B}} - \mathcal{B}\right).
\end{aligned}
$$

However, by a standard linear expansion of matrices,

$$
\begin{aligned}
\widehat{Q}_i - Q_i &= \left\{(\widehat{D} - \alpha\widehat{S})\widehat{c}_i - (D - \alpha S)c_i\right\}R_i\\
&\approx \left\{(\widehat{D} - D) - \alpha(\widehat{S} - S)\right\}c_iR_i\\
&\quad -(D - \alpha S)c_i\left\{(\widehat{D} - D) - (\alpha - k_i^{-1})(\widehat{S} - S)\right\}c_iR_i\\
&= \{I - (D - \alpha S)c_i\}(\widehat{D} - D)c_iR_i\\
&\qquad + \left\{(\alpha - k_i^{-1})(D - \alpha S)c_i - \alpha I\right\}(\widehat{S} - S)c_iR_i.
\end{aligned}
$$

However, $n^{1/2}(\widehat{D} - D) \approx n^{-1/2}\sum_{i=1}^{N}\Delta_i\Psi_{1i}$, and substituting this together with (B.7) means that

$$
\begin{aligned}
&n^{-1/2}{\textstyle\sum_{i=1}^{N}}\Delta_i\left\{r_i - s_iQ_i\left(\widehat{Q}_i - Q_i\right)^t\mathcal{B}\right\}\\
&\quad\approx n^{-1/2}{\textstyle\sum_{i=1}^{N}}\Delta_ir_i\\
&\quad -n^{-1/2}{\textstyle\sum_{i=1}^{N}}\Delta_is_iQ_iR_i^tc_in^{-1}\sum_{j=1}^{N}\Delta_j\Psi_{1j}\{I - c_i(D - \alpha S)\}\mathcal{B}\\
&\quad -n^{-1/2}{\textstyle\sum_{i=1}^{N}}\Delta_is_iQ_iR_i^tc_in^{-1}\sum_{j=1}^{N}\{\Delta_j\Psi_{2j} + (1 - \Delta_j)\Psi_{3j}\}\\
&\qquad\qquad \times\left\{(\alpha - k_i^{-1})(D - \alpha S)c_i - \alpha I\right\}\mathcal{B}.
\end{aligned}
$$

If we interchange the roles of i and j in the last expressions and inset into (B.12), we obtain (B.8).

While the standard error formulae have assumed centering, one can still make inference about the original intercept that would have been obtained had one not centered. Letting the original means of the $\mathbf{Z}_i$'s and $\overline{\mathbf{W}}_{i\cdot}$'s be $\widehat{\mu}_{z,o}$ and $\widehat{\mu}_{w,o}$, the original intercept is estimated by $\widehat{\beta}_0 + \widehat{\beta}_x^t\widehat{\mu}_{w,o} + \widehat{\beta}_z^t\widehat{\mu}_{z,o}$. If one conditions on the observed values of $\widehat{\mu}_{z,o}$ and $\widehat{\mu}_{w,o}$, then this revised intercept is the linear combination $a^t\widehat{\mathcal{B}} = (1, \widehat{\mu}_{z,o}^t, \widehat{\mu}_{z,o}^t)\widehat{\mathcal{B}}$, and its variance is estimated by $n^{-1}a^t\widehat{A}_{1n}^{-1}\widehat{A}_{2n}\widehat{A}_{1n}^{-1}a$.

If Σ_{uu} is known, or if one is willing to ignore the variation in its estimate $\widehat{\Sigma}_{uu}$, set $d_{in2} = d_{in3} = 0$. This may be relevant if $\widehat{\Sigma}_{uu}$ comes from a large, careful independent study, for which only summary statistics are available (a common occurrence).

In other cases, $\mathbf{W}$ is a scalar variable, Σ_{uu} cannot be treated as known and one must rely on an independent experiment that reports only an estimate of it. If that experiment reports an asymptotic variance $\widehat{\xi}/n$ based on a sample of size $N - n$, then Ψ_{3i} is a scalar and simplifications result which enable a valid asymptotic analysis. Define

$$
d_{n4} = n^{-1}{\textstyle\sum_{j=1}^{N}}\Delta_j\widehat{s}_j\widehat{Q}_jR_j^t\widehat{c}_j\left\{(\alpha - k_j^{-1})(\widehat{D} - \alpha\widehat{S})\widehat{c}_j - \alpha I\right\}\widehat{\mathcal{B}}.
$$

Then, in (B.10) replace $n^{-1}\sum_i(1 - \Delta_i)\widehat{d}_{in3}\widehat{d}_{in3}^{\,t}$ with $d_{n4}d_{n4}^tn\widehat{\xi}/(N - n)$.

B.3.2 Quadratic Regression: Details of the Expanded Calibration Model

Here we show that, as stated in Section 4.9.2, in quadratic regression, if $\mathbf{X}$ given $\mathbf{W}$ is symmetrically distributed and homoscedastic, the expanded model (4.17) accurately summarizes the variance function. Let $\kappa = E\left\{(\mathbf{X} - m)^4|\mathbf{W}\right\}$, which is constant because of the homoscedasticity.* Then, if $r = \mathbf{X} - m$, the variance function is given by

$$
\begin{aligned}
\text{var}(\mathbf{Y}|\mathbf{W}) &= \sigma^2 + \beta_{x,1}^2\text{var}(\mathbf{X}|\mathbf{W}) + \beta_{x,2}^2\text{var}(\mathbf{X}^2|\mathbf{W})\\
&\qquad +2\beta_{x,1}\beta_{x,2}\text{cov}\left\{(\mathbf{X}, \mathbf{X}^2)|\mathbf{W}\right\}\\
&= \sigma^2 + \beta_{x,1}^2\sigma^2 + \beta_{x,2}^2E\left\{\mathbf{X}^4 - (m^2 + \sigma^2)^2|\mathbf{W}\right\}\\
&\qquad +2\beta_{x,1}\beta_{x,2}E\left[r\left\{r^2 + 2mr - \sigma^2\right\}|\mathbf{W}\right]\\
&= \sigma^2 + \beta_{x,1}^2\sigma^2 + \beta_{x,2}^2(\kappa + 4m^2\sigma^2 - \sigma^4) + 4\beta_{x,1}\beta_{x,2}m\sigma^2\\
&= \sigma_*^2 + \sigma^2(\beta_{x,1} + 2\beta_{x,2}m)^2,
\end{aligned}
$$

where $\sigma_*^2 = \sigma^2 + \beta_{x,2}^2\kappa - \sigma^4$. The approximation (4.17) is of exactly the same form. The only difference is that it replaces the correct σ_*^2 with σ^2, but this replacement is unimportant since both are constant.

B.3.3 Heuristics and Accuracy of the Approximations

The essential step in regression calibration is the replacement of $\mathbf{X}$ with $E(\mathbf{X}|\mathbf{W}, \mathbf{Z}) = m(\mathbf{Z}, \mathbf{W}, \gamma)$ in (4.10) and (4.11), leading to the model (4.12)–(4.13). This model can be justified by a "small-σ" argument, that is, by assuming that the measurement error is small. The basic idea is that under small measurement error, $\mathbf{X}$ will be close to its expectation. However, even with small measurement error, $\mathbf{X}$ may not be close to $\mathbf{W}$, so naively replacing $\mathbf{X}$ with $\mathbf{W}$ may lead to large bias, hence the need for calibration. For simplicity, assume that $\mathbf{X}$ is univariate. Let $\mathbf{X} = E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) + \mathbf{V}$, where $E(\mathbf{V}|\mathbf{Z}, \mathbf{W}) = 0$ and $\text{var}(\mathbf{V}|\mathbf{Z}, \mathbf{W}) = \sigma_{X|Z,W}^2$. Let $m(\cdot) = m(\mathbf{Z}, \mathbf{W}, \gamma)$. Let f_x and f_{xx} be the first and second partial derivatives of $f(z, x, \mathcal{B})$ with respect to x. Assuming that $\sigma_{X|Z,W}^2$ is small,

[*] More precisely, in addition to a constant variance, we are also assuming a constant kurtosis, as is true if, for example, $\mathbf{X}$ given $\mathbf{W}$ is normally distributed.

and hence that **V** is small with high probability, we have the Taylor approximation:

$$\begin{aligned}
E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &= E\Big\{E(\mathbf{Y}|\mathbf{Z},\mathbf{W},\mathbf{X})\Big|\mathbf{Z},\mathbf{W}\Big\} \\
&\approx E\Big\{f(\mathbf{Z},m(\cdot),\mathcal{B}) + f_x(\mathbf{Z},m(\cdot),\mathcal{B})\,\mathbf{V} \\
&\quad +(1/2)f_{xx}(\mathbf{Z},m(\cdot),\mathcal{B})\,\mathbf{V}^2\Big|\mathbf{Z},\mathbf{W}\Big\} \\
&= f\left\{\mathbf{Z},m(\cdot),\mathcal{B}\right\} + (1/2)f_{xx}\left\{\mathbf{Z},m(\cdot),\mathcal{B}\right\}\sigma^2_{X|Z,W}.
\end{aligned}$$

Model (4.12) results from dropping the term involving $\sigma^2_{X|Z,W}$, which can be justified by the small-σ assumption. This term is retained in the expanded regression calibration model developed in Section 4.7.

To derive (4.13), note that

$$\begin{aligned}
\text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \quad = \quad & \text{var}\Big\{E(\mathbf{Y}|\mathbf{Z},\mathbf{W},\mathbf{X})\Big|\mathbf{Z},\mathbf{W}\Big\} \\
& +E\Big\{\text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W},\mathbf{X})\Big|\mathbf{Z},\mathbf{W}\Big\}.
\end{aligned} \tag{B.13}$$

The first term on the right-hand side of (B.13) is

$$\begin{aligned}
\text{var}\{f(\mathbf{Z},\mathbf{X},\mathcal{B})|\mathbf{Z},\mathbf{W}\} &\approx \text{var}\{f_x(\mathbf{Z},m(\cdot),\mathcal{B})\mathbf{V}|\mathbf{Z},\mathbf{W}\} \\
&= f_x^2\left\{\mathbf{Z},m(\cdot),\mathcal{B}\right\}\sigma^2_{X|Z,W},
\end{aligned}$$

which represents variability in **Y** due to measurement error and is set equal to 0 in the regression calibration approximation, but is used in the expanded regression calibration approximation of Section 4.7. Let s_x and s_{xx} be the first and second partial derivatives of $g^2(z,x,\mathcal{B},\theta)$ with respect to x. The second term on the right-hand side of (B.13) is

$$\begin{aligned}
E\{\sigma^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)|\mathbf{Z},\mathbf{W}\} &\approx \sigma^2 g^2(\mathbf{Z},m(\cdot),\mathcal{B},\theta) \\
&+\frac{1}{2}s_{xx}(Z,m(\cdot),\mathcal{B},\theta)\sigma^2_{X|Z,W}.
\end{aligned}$$

Setting the term involving $\sigma^2_{X|Z,W}$ equal to 0 gives the regression calibration approximation, while both terms are used in expanded regression calibration.

B.4 Appendix to Chapter 5: SIMEX

The ease with which estimates can be obtained via SIMEX, even for very complicated and nonstandard models, is offset somewhat by the complexity of the resulting estimates, making the calculation of standard errors difficult or at least nonstandard. Except for the computational burden of nested resampling schemes, SIMEX is a natural candidate for the use of the bootstrap or a standard implementation of Tukey's jackknife to calculate standard errors.

We now describe two methods of estimating the covariance matrix of the asymptotic distribution of $\widehat{\Theta}_{\text{simex}}$ that avoid nested resampling. We do so in the context of homoscedastic measurement error. The first method uses the pseudo estimates, $\widehat{\Theta}_b(\zeta)$, generated during the SIMEX simulation step in a procedure akin to Tukey's jackknife variance estimate. Its applicability is limited to situations in which σ_u^2 is known or estimated well enough to justify such an assumption. The second method exploits the fact that $\widehat{\Theta}_{\text{simex}}$ is asymptotically equivalent to an M-estimator and makes use of standard formulae from Appendix A. This method requires additional programming but has the flexibility to accommodate situations in which σ_u^2 is estimated and the variation in $\widehat{\sigma}_u^2$ is not negligible.

B.4.1 Simulation Extrapolation Variance Estimation

Stefanski and Cook (1995) establish a close relationship between SIMEX inference and jackknife inference. In particular, they identified a method of variance estimation applicable when σ_u^2 is known that closely parallels Tukey's jackknife variance estimation. We now describe the implementation of their method of estimating $\text{var}(\widehat{\Theta}_{\text{simex}})$.

It is convenient to introduce a function $\mathcal{T}$ to denote the estimator under study. For example, $\mathcal{T}\left\{(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{X}_i)_1^n\right\}$ is the estimator of Θ when **X** is observable, and $\mathcal{T}\left\{(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i)_1^n\right\}$ is the naive estimator.

For theoretical purposes we let

$$\widehat{\Theta}_b(\zeta) = \mathcal{T}\left\{(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i+\sqrt{\zeta}\mathbf{U}_{b,i})_1^n\right\},$$

we redefine

$$\widehat{\Theta}(\zeta) = E\left\{\widehat{\Theta}_b(\zeta)\mid(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i)_1^n\right\}. \tag{B.14}$$

The expectation in (B.14) is with respect to the distribution of $(\mathbf{U}_{b,i})_{i=1}^n$ only, since we condition on the observed data. It can be obtained as the limit as $B\to\infty$ of the average $\{\widehat{\Theta}_1(\zeta)+\cdots+\widehat{\Theta}_B(\zeta)\}/B$. In effect, $\widehat{\Theta}(\zeta)$ is the estimator obtained when computing power is unlimited.

We now introduce a second function, $\mathcal{T}_{\text{var}}$ to denote an associated variance estimator, that is,

$$\mathcal{T}_{\text{var}}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\} = \widehat{\text{var}}(\widehat{\Theta}_{\text{true}}) = \widehat{\text{var}}[\mathcal{T}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\}],$$

where $\widehat{\Theta}_{\text{true}}$ denotes the "estimator" calculated from the "true" data $(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n$.

We allow $\mathcal{T}$ to be p-dimensional, in which case $\mathcal{T}_{\text{var}}$ is $(p\times p)$-matrix valued, and variance refers to the variance–covariance matrix. For example, $\mathcal{T}_{\text{var}}$ could be the inverse of the information matrix when $\widehat{\Theta}_{\text{true}}$

is a maximum likelihood estimator. Alternatively, $\mathcal{T}_{\text{var}}$ could be a sandwich estimator for either a maximum likelihood estimator or a general M-estimator (Appendix A).

We use τ^2 to denote the parameter $\text{var}(\widehat{\Theta}_{\text{true}})$, $\widehat{\tau}^2_{\text{true}}$ to denote the true variance estimator $\mathcal{T}_{\text{var}}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\}$, and $\widehat{\tau}^2_{\text{naive}}$ to denote the naive variance estimator $\mathcal{T}_{\text{var}}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_i)_1^n\}$.

Stefanski and Cook (1995) show that

$$E\{\widehat{\Theta}_{\text{simex}} \mid (\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\} \approx \widehat{\Theta}_{\text{true}}, \tag{B.15}$$

where the approximation is due to both a large-sample approximation and to use of an approximate extrapolant function. We will make use of such approximations without further explanation; see Stefanski and Cook (1995) for additional explanation.

It follows from Equation (B.15) that

$$\text{var}(\widehat{\Theta}_{\text{simex}}) \approx \text{var}(\widehat{\Theta}_{\text{true}}) + \text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}}). \tag{B.16}$$

Equation (B.16) decomposes the variance of $\widehat{\Theta}_{\text{simex}}$ into a component due to sampling variability, $\text{var}(\widehat{\Theta}_{\text{true}}) = \tau^2$, and a component due to measurement error variability, $\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}})$.

SIMEX estimation can be used to estimate the first component τ^2. That is,

$$\widehat{\tau}^2_b(\zeta) = \mathcal{T}_{\text{var}}[\{\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_{b,i}(\zeta)\}_1^n]$$

is calculated for $b = 1, \ldots, B$, and upon averaging and letting $B \to \infty$, results in $\widehat{\tau}^2(\zeta)$. The components of $\widehat{\tau}^2(\zeta)$ are then plotted as functions of ζ, extrapolant models are fit to the components of $\{\widehat{\tau}^2(\zeta_m), \zeta_m\}_1^M$ and the modeled values at $\zeta = -1$ are estimates of the corresponding components of τ^2.

The basic building blocks required to estimate the second component of the variance, $\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}})$, are the differences

$$\Delta_b(\zeta) = \widehat{\Theta}_b(\zeta) - \widehat{\Theta}(\zeta), \quad b = 1, \ldots, B. \tag{B.17}$$

Define

$$s^2_\Delta(\zeta) = (B-1)^{-1}{\textstyle\sum_{b=1}^B}\Delta_b(\zeta)\Delta_b^t(\zeta), \tag{B.18}$$

that is, the sample variance matrix of $\{\widehat{\Theta}_b(\zeta)\}_{b=1}^B$. Its significance stems from the fact that

$$\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}}) = -\lim_{\zeta \to -1} \text{var}\{\widehat{\Theta}_b(\zeta) - \widehat{\Theta}(\zeta)\}; \tag{B.19}$$

see Stefanski and Cook (1995). The minus sign on the right-hand side of (B.19) is not an misprint; $\text{var}\{\widehat{\Theta}_b(\zeta) - \widehat{\Theta}(\zeta)\}$ is positive for $\zeta > 0$ and zero for $\zeta = 0$, so the extrapolant is negative for $\zeta < 0$.

The variance matrix $s^2_\Delta(\zeta)$ is an unbiased estimator of the conditional variance $\text{var}\{\widehat{\Theta}_b(\zeta) - \widehat{\Theta}(\zeta) \mid (\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n\}$ for all $B > 1$ and converges in probability to its conditional expectation as $B \to \infty$. Since $E\{\widehat{\Theta}_b(\zeta) - \widehat{\Theta}(\zeta) \mid (\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n\} = 0$, it follows that unconditionally $E\{s^2_\Delta(\zeta)\} = \text{var}\{\widehat{\Theta}_b(\zeta) - \widehat{\Theta}(\zeta)\}$.

Thus, the component of variance we want to estimate is given by

$$\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}}) = -\lim_{\zeta \to -1} E\{s^2_\Delta(\zeta)\}.$$

This can be (approximately) estimated by fitting models to the components of $s^2_\Delta(\zeta)$ as functions of $\zeta > 0$ and extrapolating the component models back to $\zeta = -1$. We use $\widehat{s}^2_\Delta(-1)$ to denote the estimated variance matrix obtained by this procedure.

In light of (B.16), the definition of τ^2, and (B.19) the difference, $\widehat{\tau}^2_{\text{simex}} - \widehat{s}^2_\Delta(-1)$, is an estimator of $\text{var}\{\widehat{\Theta}_{\text{simex}}\}$. In practice, separate extrapolant functions are not fit to the components of both $\widehat{\tau}^2(\zeta)$ and $s^2_\Delta(\zeta)$; rather, the components of the difference, $\widehat{\tau}^2(\zeta) - s^2_\Delta(\zeta)$, are modeled and extrapolated to $\zeta = -1$.

In summary, for SIMEX estimation with known σ^2_u, the simulation step results in $\widehat{\Theta}(\zeta)$, $\widehat{\tau}^2(\zeta)$ and $s^2_\Delta(\zeta)$ for $\zeta \in \Lambda$. The model extrapolation of $\widehat{\Theta}(\zeta)$ to $\zeta = -1$, $\widehat{\Theta}_{\text{simex}}$, provides an estimator of Θ, and the model extrapolation of (the components of) the difference, $\widehat{\tau}^2(\zeta) - s^2_\Delta(\zeta)$ to $\zeta = -1$ provides an estimator of $\text{var}(\widehat{\Theta}_{\text{simex}})$. It should be emphasized that the entire procedure is approximate in the sense that it is generally valid only in large samples with small measurement error.

There is no guarantee that the estimated covariance matrix so obtained is positive definite. This is similar to the nonpositivity problems that arise in estimating components-of-variance. We have not encountered problems of this nature, although there is no guarantee that they will not occur. If it transpires that the estimated variance of a linear combination, say $\gamma^t\widehat{\Theta}$, is negative, a possible course of action is to plot, model, and extrapolate directly the points $[\gamma^t\{\widehat{\tau}^2(\zeta_m) - s^2_\Delta(\zeta_m\}\gamma,\ \zeta_m]_1^M$, though there is also no guarantee that its extrapolation will be nonnegative.

B.4.2 Estimating Equation Approach to Variance Estimation

This section is based on the results in Carroll, Küchenhoff, Lombard, and Stefanski (1996). Assuming iid sampling, these authors showed that the estimate of Θ is asymptotically normally distributed and proposed an estimator of its asymptotic covariance matrix. We highlight the main points of the asymptotic analysis in order to motivate the proposed variance estimator.

We describe the application of SIMEX in the setting of M-estimation, that is, using unbiased estimating equations (Appendix A), assuming that in the absence of measurement errors, M-estimation produces consistent estimators.

The estimator obtained in the absence of measurement error is denoted $\widehat{\Theta}_{\text{true}}$ and solves the system of equations

$$0 = n^{-1}\textstyle\sum_{i=1}^{n}\Psi(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \widehat{\Theta}_{\text{true}}). \tag{B.20}$$

This is just a version of (A.19) and is hence applicable to variance function and generalized linear models. In multiple linear regression, $\Psi(\cdot)$ represents the normal equations for a single observation, namely,

$$\Psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta) = (\mathbf{Y} - \beta_0 - \beta_z^t\mathbf{Z} - \beta_x\mathbf{X})(1, \mathbf{Z}^t, \mathbf{X}^t)^t.$$

In multiple logistic regression, with $H(\cdot)$ being the logistic distribution function,

$$\Psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta) = \left\{\mathbf{Y} - H\left(\beta_0 + \beta_z^t\mathbf{Z} + \beta_x\mathbf{X}\right)\right\}(1, \mathbf{Z}^t, \mathbf{X}^t)^t.$$

The solution to (B.20) cannot be calculated, since it depends on the unobserved true predictors. The estimator obtained by ignoring measurement error is denoted by $\widehat{\Theta}_{\text{naive}}$ and solves the system of equations

$$0 = n^{-1}\textstyle\sum_{i=1}^{n}\Psi(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}_{\text{naive}}).$$

For fixed b and ζ and large n, a standard linearization (Appendix A) shows that

$$\begin{aligned} n^{1/2}\left\{\widehat{\Theta}_b(\zeta) - \Theta(\zeta)\right\} &\approx -\mathcal{A}^{-1}\{\sigma_u^2, \zeta, \Theta(\zeta)\} \\ &\quad \times n^{-1/2}\textstyle\sum_{i=1}^{n}\Psi\{\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_{b,i}(\zeta), \Theta(\zeta)\}, \end{aligned} \tag{B.21}$$

where $\mathcal{A}\{\sigma_u^2, \zeta, \Theta(\zeta)\} = E\left[\Psi_\Theta\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W}_{b,i}(\zeta), \Theta(\zeta)\}\right]$, and

$$\Psi_\Theta\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W}_{b,i}(\zeta), \Theta\} = (\partial/\partial\Theta^t)\Psi\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W}_{b,i}(\zeta), \Theta\}.$$

Averaging (B.21) over b results in the asymptotic approximation

$$\begin{aligned} n^{1/2}\left\{\widehat{\Theta}(\zeta) - \Theta(\zeta)\right\} &\approx -\mathcal{A}^{-1}(\cdot) \\ &\quad \times n^{-1/2}\textstyle\sum_{i=1}^{n}\chi_{B,i}\{\sigma_u^2,\ \zeta,\ \Theta(\zeta)\}, \end{aligned} \tag{B.22}$$

where $\chi_{B,i}\{\sigma_u^2, \zeta, \Theta(\zeta)\} = B^{-1}\sum_{b=1}^{B}\Psi\{\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_{b,i}(\zeta), \Theta(\zeta)\}$, and $\mathcal{A}^{-1}(\cdot)$ $= \mathcal{A}^{-1}\{\sigma_u^2, \zeta, \Theta(\zeta)\}$. The summands $\chi_{B,i}(\cdot)$ in (B.22) are independent and identically distributed with mean zero.

Let $\Lambda = \{\zeta_1, \ldots, \zeta_M\}$ denote the grid of values used in the extrapolation step. Let $\widehat{\Theta}_*(\Lambda)$ denote $\{\widehat{\Theta}^t(\zeta_1), \ldots, \widehat{\Theta}^t(\zeta_M)\}^t$, which we also denote vec$\{\widehat{\Theta}(\zeta), \zeta \in \Lambda\}$. The corresponding vector of estimands is denoted by $\Theta_*(\Lambda)$. Define

$$\begin{aligned} \Psi_{B,i(1)}\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\} &= \text{vec}[\chi_{B,i}\{\sigma_u^2, \zeta, \Theta(\zeta)\}, \zeta \in \Lambda] \\ \mathcal{A}_{11}\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\} &= \text{diag}[\mathcal{A}\{\sigma_u^2,\ \zeta,\ \Theta(\zeta)\}, \zeta \in \Lambda]. \end{aligned}$$

Then, using (B.22), the joint limit distribution of $n^{1/2}\{\widehat{\Theta}_*(\Lambda) - \Theta_*(\Lambda)\}$ is seen to be multivariate normally distributed with mean zero and covariance Σ, where

$$\Sigma = \mathcal{A}_{11}^{-1}(\cdot)\mathcal{C}_{11}\left\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\right\}\left\{\mathcal{A}_{11}^{-1}(\cdot)\right\}^t \tag{B.23}$$

$$\mathcal{C}_{11}\left\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\right\} = \text{Cov}\left[\Psi_{B,1(1)}\left\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\right\}\right]. \tag{B.24}$$

Define $\mathcal{G}^*(\Lambda, \Gamma^*) = \text{vec}\left[\{\mathcal{G}(\zeta_m, \Gamma_j)\}_{m=1,\ldots,M,\ j=1,\ldots,p}\right]$, where $\Gamma^* = (\Gamma_1^t, \ldots, \Gamma_p^t)^t$ and Γ_j is the parameter vector estimated in the extrapolation step for the j^{th} component of $\widehat{\Theta}(\zeta)$, $j = 1, \ldots, p$.

Define $R(\Gamma^*) = \widehat{\Theta}_*(\Lambda) - \mathcal{G}^*(\Lambda, \Gamma^*)$. The extrapolation steps results in $\widehat{\Gamma}^*$, obtained by minimizing $R^t(\Gamma^*)R(\Gamma^*)$. The estimating equation for $\widehat{\Gamma}^*$ has the form $0 = s(\Gamma^*)R(\Gamma^*)$ where $s^t(\Gamma^*) = \{\partial/\partial(\Gamma^*)^t\}R(\Gamma^*)$. With $D(\Gamma^*) = s(\Gamma^*)s^t(\Gamma^*)$, standard asymptotic results show that

$$n^{-1/2}(\widehat{\Gamma}^* - \Gamma^*) \approx \text{N}\{0,\ \Sigma(\Gamma^*)\},$$

where $\Sigma(\Gamma^*) = D^{-1}(\Gamma^*)s(\Gamma^*)\Sigma s^t(\Gamma^*)D^{-1}(\Gamma^*)$ and Σ is given by (B.23). Now $\widehat{\Theta}_{\text{simex}} = \mathcal{G}^*(-1, \widehat{\Gamma}^*)$ and thus by the Δ method, the $\sqrt{n}$-normalized SIMEX estimator is asymptotically normal with asymptotic variance,

$$\mathcal{G}_{\Gamma^*}^*(-1, \Gamma^*)\Sigma(\Gamma^*)\{\mathcal{G}_{\Gamma^*}^*(-1, \Gamma^*)\}^t$$

where $\mathcal{G}_{\Gamma^*}^*(\zeta, \Gamma^*) = \{\partial/\partial(\Gamma^*)^t\}\mathcal{G}^*(\zeta, \Gamma^*)$.

Note that the matrix $C_{11}(\cdot)$ in (B.24) is consistently estimated by $\widehat{\mathcal{C}}_{11}(\cdot)$, the sample covariance matrix of $[\Psi_{B,i(1)}\{\sigma_u^2, \Lambda, \widehat{\Theta}_*(\Lambda)\}]_1^n$. Also, $A_{11}(\cdot)$ is consistently estimated by $\widehat{\mathcal{A}}_{11}(\cdot) = \text{diag}\{\widehat{\mathcal{A}}_m(\cdot)\}$ for $m = 1, \ldots, M$, where

$$\widehat{\mathcal{A}}_m(\cdot) = (nB)^{-1}\textstyle\sum_{i=1}^{n}\sum_{b=1}^{B}\Psi_\Theta\{\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_{b,i}(\zeta_m),\ \widehat{\Theta}(\zeta_m)\}.$$

The indicated variance estimator is

$$n^{-1}\mathcal{G}_{\Gamma^*}^*(-1, \widehat{\Gamma}^*)\widehat{\Sigma}(\widehat{\Gamma}^*)\{\mathcal{G}_{\Gamma^*}^*(-1, \widehat{\Gamma}^*)\}^t, \tag{B.25}$$

where $\widehat{D}(\widehat{\Gamma}^*) = s(\widehat{\Gamma}^*)s^t(\widehat{\Gamma}^*)$ and

$$\begin{aligned} \widehat{\Sigma}(\widehat{\Gamma}^*) &= \widehat{D}^{-1}(\widehat{\Gamma}^*)s(\widehat{\Gamma}^*)\widehat{\Sigma}s^t(\widehat{\Gamma}^*)\widehat{D}^{-1}(\widehat{\Gamma}^*); \\ \widehat{\Sigma} &= \widehat{\mathcal{A}}_{11}^{-1}(\cdot)\widehat{\mathcal{C}}_{11}^{-1}(\cdot)\left\{\widehat{\mathcal{A}}_{11}^{-1}(\cdot)\right\}^t. \end{aligned}$$

When σ_u^2 is estimated, the estimating equation approach is modified by the inclusion of additional estimating equations employed in the estimation of $\widehat{\sigma_u}^2$. We illustrate the case in which each $\mathbf{W}_i$ is the mean of two replicate measurements, $\mathbf{W}_{ij}$, $j = 1, 2$ where

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{i,j}, \qquad j = 1, 2, \quad i = 1, \ldots, n.$$

With replicates, $\mathbf{W}_i$ is replaced by $\mathbf{W}_i^* = \overline{\mathbf{W}}_{i,\cdot}$ and σ_u^2 by $\sigma_{u,*}^2 = \sigma_u^2/2$.

Let

$$\Psi_{(i)2}(\sigma_{u,*}^2, \mu) = \left\{ \begin{array}{c} (\mathbf{D}_i - \mu)^2 - \sigma_{u,*}^2 \\ \mathbf{D}_i - \mu \end{array} \right\},$$

where $\mathbf{D}_i = (\mathbf{W}_{i1} - \mathbf{W}_{i2})/2$. Then solving $\sum \Psi_{i(2)}(\sigma_{u,*}^2, \mu) = 0$, results in the estimators $\widehat{\mu} = \overline{\mathbf{D}}$ and $\widehat{\sigma}_{u,*}^2 = (n-1)s_d^2/n$, where s_d^2 is the sample variance of $(\mathbf{D}_i)_1^n$ and consistently estimates $\sigma_{u,*}^2$.

By combining $\Psi_{B,i(1)}$ and $\Psi_{i(2)}$ into a single estimating equation and applying standard theory, the covariance matrix of the joint distribution of $\widehat{\Theta}_*(\Lambda)$, $\widehat{\sigma}_{u,*}^2$, and $\widehat{\mu}$ is

$$n^{-1} \left\{ \begin{array}{cc} \mathcal{A}_{11}(\cdot) & \mathcal{A}_{12}(\cdot) \\ 0 & \mathcal{A}_{22}(\cdot) \end{array} \right\}^{-1} \left\{ \begin{array}{cc} \mathcal{C}_{11}(\cdot) & \mathcal{C}_{12}(\cdot) \\ \mathcal{C}_{12}^t(\cdot) & \mathcal{C}_{22}(\cdot) \end{array} \right\} \left\{ \begin{array}{cc} \mathcal{A}_{11}(\cdot) & \mathcal{A}_{12}(\cdot) \\ 0 & \mathcal{A}_{22}(\cdot) \end{array} \right\}^{-t} \tag{B.26}$$

where

$$\left\{ \begin{array}{cc} \mathcal{C}_{11}(\cdot) & \mathcal{C}_{12}(\cdot) \\ \mathcal{C}_{12}^t(\cdot) & \mathcal{C}_{22}(\cdot) \end{array} \right\} = \mathcal{C}_*(\cdot) = \text{cov} \left[\begin{array}{c} \Psi_{B,1(1)} \left\{ \sigma_{u,*}^2, \Lambda, \Theta_*(\Lambda) \right\} \\ \Psi_{1(2)} \left(\sigma_{u,*}^2, \mu \right) \end{array} \right],$$

$$\begin{aligned} &\mathcal{A}_{12}\{\sigma_{u,*}^2, \Lambda, \Theta_*(\Lambda)\} \\ &\quad = n^{-1} {\textstyle\sum_{i=1}^n} E[\frac{\partial}{\partial(\sigma_{u,*}^2, \mu)} \Psi_{B,i(1)} \left\{ \sigma_{u,*}^2, \Lambda, \Theta_*(\Lambda) \right\}], \end{aligned}$$

and

$$\begin{aligned} \mathcal{A}_{22} \left(\sigma_{u,*}^2, \mu \right) &= n^{-1} {\textstyle\sum_{i=1}^n} E \left\{ \frac{\partial}{\partial(\sigma_{u,*}^2, \mu)} \Psi_{i(2)} \left(\sigma_{u,*}^2, \mu \right) \right\} \\ &= -n^{-1} {\textstyle\sum_{i=1}^n} E \left\{ \begin{array}{cc} 1 & 2(\mathbf{D}_i - \mu) \\ 0 & 1 \end{array} \right\} = - \left\{ \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right\}. \end{aligned}$$

Estimating these quantities via the sandwich method is straightforward. For $\mathcal{A}_{12}(\cdot)$ remove the expectation symbol and replace $\left\{ \sigma_{u,*}^2, \Theta_*(\Lambda), \mu \right\}$ by the estimates $\left\{ \widehat{\sigma}_{u,*}^2, \widehat{\Theta}_*(\Lambda), \widehat{\mu} \right\}$. The covariance matrix $\mathcal{C}_*(\cdot)$ can be estimated by the sample covariance matrix of the vectors

$$\left[\begin{array}{c} \Psi_{B,i(1)} \left\{ \widehat{\sigma}_{u,*}^2, \Lambda, \widehat{\Theta}_*(\Lambda) \right\} \\ \Psi_{i(2)} \left(\widehat{\sigma}_{u,*}^2, \widehat{\mu} \right) \end{array} \right].$$

These estimates are substituted into (B.26), thereby obtaining an estimate of the joint covariance matrix of $\widehat{\Theta}_*(\Lambda)$, $\widehat{\sigma}_{u,*}^2$, and $\widehat{\mu}$. The submatrix corresponding to the components of $\widehat{\Theta}_*(\Lambda)$ is now employed in (B.25) in place of $\widehat{\Sigma}$.

B.5 Appendix to Chapter 6: Instrumental Variables

B.5.1 Derivation of the Estimators

In this section, we derive the estimators presented in Section 6.3. We start with the following assumptions:

$$\begin{aligned} E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}, \mathbf{W}) = \beta_{X|\underline{1}ZTW}^t + \beta_{X|1\underline{Z}TW}^t \mathbf{Z} + \\ \beta_{X|1Z\underline{T}W}^t \mathbf{T} + \beta_{X|1ZT\underline{W}}^t \mathbf{W}; \qquad \text{(B.27)} \\ E(\mathbf{X} - \mathbf{W} \mid \mathbf{Z}, \mathbf{X}, \mathbf{T}) = 0; \qquad \text{(B.28)} \\ E(\mathbf{Y} \mid \mathbf{Z},\ \mathbf{T},\ \mathbf{W}) = E\big\{E(\mathbf{Y} \mid \mathbf{Z},\ \mathbf{X}) \mid \mathbf{Z},\ \mathbf{T},\ \mathbf{W}\big\}. \qquad \text{(B.29)} \end{aligned}$$

We have discussed each of these previously. Note that (B.27) and (B.28) imply that $E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}) = E(\mathbf{W} \mid \mathbf{Z}, \mathbf{T})$ and also that $\beta_{W|\underline{1}ZT} = \beta_{X|\underline{1}ZT}$, $\beta_{W|1\underline{Z}T} = \beta_{X|1\underline{Z}T}$, and $\beta_{W|1Z\underline{T}} = \beta_{X|1Z\underline{T}}$.

B.5.1.1 First Regression Calibration Instrumental Variable Algorithm

The first algorithms are simple to describe once (6.17) is justified, which we do now. Making use of the fact that $\mathbf{T}$ is a surrogate, (B.29) and the standard regression calibration approximation results in the approximate model

$$\begin{aligned} E(\mathbf{Y}|\widetilde{\mathbf{T}}) &= f\{\beta_{Y|\underline{\widetilde{X}}}^t E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}})\} = f(\beta_{Y|\underline{\widetilde{X}}}^t \beta_{\widetilde{X}|\underline{\widetilde{T}}}^t \widetilde{\mathbf{T}}), && \text{(B.30)} \\ \text{var}(\mathbf{Y}|\widetilde{\mathbf{T}}) &= \sigma^2 g^2 \{\beta_{Y|\underline{\widetilde{X}}}^t E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}}), \theta\} && \text{(B.31)} \\ &= \sigma^2 g^2 (\beta_{Y|\underline{\widetilde{X}}}^t \beta_{\widetilde{X}|\underline{\widetilde{T}}}^t \widetilde{\mathbf{T}}, \theta). \end{aligned}$$

It follows from (B.30)–(B.31) that the coefficient of $\widetilde{\mathbf{T}}$ in the generalized linear regression of $\mathbf{Y}$ on $\widetilde{\mathbf{T}}$ is $\beta_{Y|\underline{\widetilde{T}}}^t = \beta_{Y|\underline{\widetilde{X}}}^t \beta_{\widetilde{X}|\underline{\widetilde{T}}}^t$. By (B.28) $\beta_{\widetilde{W}|\underline{\widetilde{T}}} = \beta_{\widetilde{X}|\underline{\widetilde{T}}}$, and (6.17) follows.

B.5.1.2 Second Regression Calibration Instrumental Variable Algorithm

The derivation of the second algorithm is somewhat involved, but the estimator is relatively easy to compute. Remember that the strategy is to exploit the fact that both $\mathbf{W}$ and $\mathbf{T}$ are surrogates.

Making use of the fact that both $\mathbf{T}$ and $\mathbf{W}$ are surrogates, application of the standard regression calibration approximation produces the approximate model

$$\begin{aligned} E(\mathbf{Y}|\widetilde{\mathbf{T}}, \widetilde{\mathbf{W}}) &= f\{\beta_{Y|\underline{\widetilde{X}}}^t E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}}, \widetilde{\mathbf{W}})\}, && \text{(B.32)} \\ \text{var}(\mathbf{Y}|\widetilde{\mathbf{T}}, \widetilde{\mathbf{W}}) &= \sigma^2 g^2 \{\beta_{Y|\underline{\widetilde{X}}}^t E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}}, \widetilde{\mathbf{W}}), \theta\}. && \text{(B.33)} \end{aligned}$$

Under the linear regression assumption (B.27), there exist coefficient matrices $\beta^t_{\widetilde{X}|\underline{\widetilde{T}}\widetilde{W}}$ and $\beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}}$, such that

$$E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}}, \widetilde{\mathbf{W}}) = \beta^t_{\widetilde{X}|\underline{\widetilde{T}}\widetilde{W}}\widetilde{\mathbf{T}} + \beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}}\widetilde{\mathbf{W}}. \tag{B.34}$$

Taking conditional expectations of both sides of (B.34) with respect to $\widetilde{\mathbf{T}}$ and using the fact that $E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}}) = \beta^t_{\widetilde{X}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}$ results in the identity

$$\beta^t_{\widetilde{X}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}} = \beta^t_{\widetilde{X}|\underline{\widetilde{T}}\widetilde{W}}\widetilde{\mathbf{T}} + \beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}}\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}.$$

Equating coefficients of $\widetilde{\mathbf{T}}$ and using the fact that $\beta_{\widetilde{W}|\underline{\widetilde{T}}} = \beta_{\widetilde{X}|\underline{\widetilde{T}}}$, we find that

$$\beta^t_{\widetilde{X}|\underline{\widetilde{T}}} = \beta^t_{\widetilde{X}|\underline{\widetilde{T}}\widetilde{W}} + \beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}}\beta^t_{\widetilde{X}|\underline{\widetilde{T}}}. \tag{B.35}$$

Solving (B.35) for $\beta^t_{\widetilde{X}|\underline{\widetilde{T}}\widetilde{W}}$ and then substitution into (B.34) shows that

$$E(\widetilde{\mathbf{X}} \mid \widetilde{\mathbf{T}}, \widetilde{\mathbf{W}}) = (I - \beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}})\beta^t_{\widetilde{X}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}} + \beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}}\widetilde{\mathbf{W}}. \tag{B.36}$$

By convention, $\beta^t_{Y|\underline{\widetilde{T}\widetilde{W}}}$ is the regression coefficient of $(\widetilde{\mathbf{T}}^t,\ \widetilde{\mathbf{W}}^t)^t$ in the generalized linear regression of $\mathbf{Y}$ on $\widetilde{\mathbf{T}}$ and $\widetilde{\mathbf{W}}$. The indicated model is overparameterized ,and thus the components of $\beta^t_{Y|\underline{\widetilde{T}\widetilde{W}}}$ are not uniquely determined. Although other specifications are possible, we define the components of $\beta_{Y|\underline{\widetilde{T}\widetilde{W}}}$ uniquely as

$$\begin{aligned}
\beta_{Y|\underline{\widetilde{T}}\widetilde{W}} &= \beta_{Y|\underline{1ZT}W}, \\
\beta_{Y|\widetilde{T}\underline{\widetilde{W}}} &= (0_{1\times d},\ \beta^t_{Y|1ZT\underline{W}})^t,
\end{aligned}$$

where $d = 1 + \dim(\mathbf{Z})$. Let H_1 and H_2 be the matrices that define $\beta_{Y|\underline{\widetilde{T}}\widetilde{W}}$ and $\beta_{Y|\widetilde{T}\underline{\widetilde{W}}}$ in terms of $\beta_{Y|\underline{1ZTW}}$, so that $\beta_{Y|\underline{\widetilde{T}}\widetilde{W}} = H_1\beta_{Y|\underline{1ZTW}}$ and $\beta_{Y|\widetilde{T}\underline{\widetilde{W}}} = H_2\beta_{Y|\underline{1ZTW}}$. Also note that because $\widetilde{\mathbf{T}} = (1, \mathbf{Z}^t, \mathbf{T}^t)^t$, our notation allows us to write $\beta^t_{Y|\underline{1ZTW}} = \beta^t_{Y|\underline{\widetilde{T}W}}$.

Substitution of (B.36) into (B.32) and equating coefficients of $\widetilde{\mathbf{T}}$ and $\widetilde{\mathbf{W}}$ results in the two equations:

$$\begin{aligned}
\beta^t_{Y|\underline{\widetilde{T}}\widetilde{W}} &= \beta^t_{Y|\underline{\widetilde{X}}}(I - \beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}})\beta^t_{\widetilde{X}|\underline{\widetilde{T}}}, && \text{(B.37)}\\
\beta^t_{Y|\widetilde{T}\underline{\widetilde{W}}} &= \beta^t_{Y|\underline{\widetilde{X}}}\beta^t_{\widetilde{X}|\widetilde{T}\underline{\widetilde{W}}}. && \text{(B.38)}
\end{aligned}$$

Postmultiplying (B.38) by $\beta^t_{\widetilde{X}|\underline{\widetilde{T}}}$ and adding the resulting equation to (B.37) results in the single equation,

$$\beta_{Y|\underline{\widetilde{T}}\widetilde{W}} + \beta_{\widetilde{X}|\underline{\widetilde{T}}}\beta_{Y|\widetilde{T}\underline{\widetilde{W}}} = \beta_{\widetilde{X}|\underline{\widetilde{T}}}\beta_{Y|\underline{\widetilde{X}}},$$

which, upon using the definitions of H_1 and H_2 and the identity $\beta_{\widetilde{X}|\underline{\widetilde{T}}} =$ $\beta_{\widetilde{W}|\underline{\widetilde{T}}}$, is shown to be equivalent to

$$H_1\beta_{Y|\underline{\widetilde{T}W}} + \beta_{\widetilde{W}|\underline{\widetilde{T}}}H_2\beta_{Y|\underline{\widetilde{T}W}} = \beta_{\widetilde{W}|\underline{\widetilde{T}}}\beta_{Y|\underline{\widetilde{X}}}.$$

Let $\widehat{\beta}_{Y|\underline{\widetilde{T}W}}$ be the estimated regression parameter from the generalized linear regression of $\mathbf{Y}$ on $(\mathbf{1}, \mathbf{Z}, \mathbf{T}, \mathbf{W})$, and let $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}$ be as before. Under the identifiability assumption that for a given matrix M_2, $(\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}M_2\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})$ is asymptotically nonsingular, it follows that the estimator (6.19), namely,

$$\widehat{\beta}^{IV2,(M_2)}_{Y|1\underline{X}} \quad = \quad \widehat{\beta}^{-(M_2)}_{\widetilde{W}|\underline{\widetilde{T}}}(H_1\widehat{\beta}_{Y|\underline{\widetilde{T}W}} + \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}H_2\widehat{\beta}_{Y|\underline{\widetilde{T}W}}),$$

is approximately consistent for $\beta_{Y|\underline{\widetilde{X}}}$.

When $\mathbf{T}$ and $\mathbf{W}$ are the same dimension, $\widehat{\beta}^{IV2,(M_2)}_{Y|1\underline{X}}$ does not depend on M_2. In Section B.5.2.1, we derive an estimate $\widehat{M_2}$ that minimizes the asymptotic variance of $\widehat{\beta}^{IV2,(M_2)}_{Y|1\underline{X}}$ for the case $\dim(\mathbf{T}) > \dim(\mathbf{W})$.

B.5.2 Asymptotic Distribution Approximations

We first derive the asymptotic distributions assuming that M_1 and M_2 are fixed and that M-estimation is used in the generalized linear and linear regression modeling steps. We then show how to estimate M_1 and M_2 for efficient asymptotic inference.

Let ψ denote the score function for the generalized linear model under consideration (6.17)–(6.17). This score function has as many as three components, the first corresponding to the unknown regression parameter, the second and third to the parameters in the variance function. All of the components are functions of the unknown parameters, the response variable and the vector of covariate/predictor variables. For example, with logistic regression there are no variance parameters and $\psi(y, x, \beta) = \{y - H(\beta^t x)\}\, x$ where $H(t) = 1/\{1 + \exp(-t)\}$.

Let

$$\psi_{1i} = \psi\left\{\mathbf{Y}_i, \widetilde{\mathbf{T}}_i, \beta_{Y|\underline{\widetilde{T}}}, \sigma_1^2, \theta_1\right\}$$

denote the i^{th} score function employed in fitting the approximate model (B.30)–(B.31) to $(\mathbf{Y}_i, \widetilde{\mathbf{T}}_i)_1^n$. Let

$$\psi_{2i} = \psi\left\{\mathbf{Y}_i, (\widetilde{\mathbf{T}}_i^t, \mathbf{W}_i^t)^t, \beta_{Y|\underline{\widetilde{T}W}}, \sigma_2^2, \theta_2\right\}$$

denote the i^{th} score function employed in fitting the approximate model (B.32)–(B.33) to $\left\{\mathbf{Y}_i, (\widetilde{\mathbf{T}}_i^t, \mathbf{W}_i^t)^t\right\}_1^n$. Note that each fit of the generalized linear model produces estimates of the variance parameters as well as the regression coefficients. These are denoted with subscripts as above, for example, σ_1^2, θ_1, etc.

Let ψ_{3i} denote the i^{th} score function used to estimate $\text{vec}(\beta_{\widetilde{W}|\underline{\widetilde{T}}})$, for example, for least squares estimation

$$\psi_{3i} = \left(\widetilde{\mathbf{W}}_i - \beta^t_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}_i\right) \otimes \widetilde{\mathbf{T}}_i,$$

and let

$$\psi_{4i} = \psi\left\{\mathbf{Y}_i, (\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}_i), \beta^{IV1,RC}_{Y|\underline{\widetilde{X}}}, \sigma^2_3, \theta_3\right\}.$$

Finally, define ψ_{5i} and ψ_{6i} as

$$\psi_{5i} = \left(\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}M_1\beta_{\widetilde{W}|\underline{\widetilde{T}}}\right)\beta^{IV1,(M_1)}_{Y|\underline{\widetilde{X}}} - \beta^t_{\widetilde{W}|\underline{\widetilde{T}}}M_1\beta_{Y|\underline{\widetilde{T}}},$$

and

$$\begin{aligned}\psi_{6i} \quad = \quad & \left(\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}M_2\beta_{\widetilde{W}|\underline{\widetilde{T}}}\right)\beta^{IV2,(M_2)}_{Y|\underline{\widetilde{X}}} \\ & -\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}M_2(H_1\beta_{Y|\underline{\widetilde{T}W}} + \beta_{\widetilde{W}|\underline{\widetilde{T}}}H_2\beta_{Y|\underline{\widetilde{T}W}}).\end{aligned}$$

Note that neither ψ_{5i} nor ψ_{6i} depends on i.

Define the composite parameter

$$\begin{aligned}\Theta \quad = \quad & \left\{\beta^t_{Y|\underline{\widetilde{T}}}, \sigma^2_1, \theta^t_1, \beta^t_{Y|\underline{\widetilde{T}W}}, \sigma^2_2, \theta^t_2, \text{vec}^t\left(\beta^t_{\widetilde{W}|\underline{\widetilde{T}}}\right),\right. \qquad \text{(B.39)} \\ & \left.\left(\beta^{IV1,RC}_{Y|\underline{\widetilde{X}}}\right)^t, \sigma^2_3, \theta^t_3, \left(\beta^{IV1,(M_1)}_{Y|\underline{\widetilde{X}}}\right)^t, \left(\beta^{IV2,(M_2)}_{Y|\underline{\widetilde{X}}}\right)^t\right\}^t,\end{aligned}$$

and the i^{th} composite score function

$$\psi_i(\Theta) = \left(\psi^t_{1i}, \psi^t_{2i}, \psi^t_{3i}, \psi^t_{4i}, \psi^t_{5i}, \psi^t_{6i}\right)^t. \qquad \text{(B.40)}$$

It follows that $\widehat{\Theta}$ solves

$$\textstyle\sum_{i=1}^n \psi_i(\widehat{\Theta}) = 0_{\text{dim}(\Theta)\times 1},$$

showing that $\widehat{\Theta}$ is an M-estimator. Thus, under fairly general conditions $\widehat{\Theta}$ is approximately normally distributed in large samples and the theory of Chapter A applies.

An estimate of the asymptotic covariance matrix of $\widehat{\Theta}$ is given by the sandwich formula $\widehat{A}^{-1}_n\widehat{B}_n(\widehat{A}^{-1}_n)^t$, where $\widehat{A}_n = \sum_{i=1}^n \psi_{i\Theta}(\widehat{\Theta})$ with $\psi_{i\Theta}(\Theta) = \partial\psi_i(\Theta)/\partial\Theta^t$, and $\widehat{B}_n = \sum_{i=1}^n \psi_i(\widehat{\Theta})\psi^t_i(\widehat{\Theta})$. Note that because we are fitting approximate (or misspecified) models, information-based standard errors, that is, standard errors obtained by replacing $\widehat{A}_n$ and $\widehat{B}_n$ with model-based estimates exploiting the information identity, are generally not appropriate.

Let $\widehat{\Omega} = \widehat{A}^{-1}_n\widehat{B}_n(\widehat{A}^{-1}_n)^t$ and let $\widehat{\Omega}_{i,j},\ i,j = 1,\ldots,12$ denote the $(i,j)^{\text{th}}$ submatrix of $\widehat{\Omega}$ corresponding to the natural partitioning induced by the components of Θ in (B.39). It follows that $\widehat{\Omega}_{8,8}$, $\widehat{\Omega}_{11,11}$, and $\widehat{\Omega}_{12,12}$

are estimates of the variance matrices of the asymptotic distributions of $\widehat{\beta}^{IV1,RC}_{Y|1\underline{X}}$, $\widehat{\beta}^{IV1,(M_1)}_{Y|1\underline{X}}$, and $\widehat{\beta}^{IV2,(M_2)}_{Y|1\underline{X}}$, respectively.

B.5.2.1 Two-Stage Estimation

When $\mathbf{T}$ and $\mathbf{W}$ have the same dimension, the estimators (6.18) and (6.19) do not depend on M_1 and M_2. However, when there are more instruments than predictors measured with error it is possible to identify and consistently estimate matrices M_1 and M_2 which minimize the asymptotic variance matrix of the corresponding estimators. We give the results first and then sketch their derivations.

For an asymptotically efficient estimator (6.18), replace M_1 with

$$\widehat{M}_{1,\text{opt}} = \widehat{\Omega}_{1,1} - \widehat{\Omega}_{1,7}\widehat{C}^t - \widehat{C}\widehat{\Omega}_{7,1} + \widehat{C}\widehat{\Omega}_{7,7}\widehat{C}^t)^{-1},$$

where $\widehat{C} = I_{d_{\widetilde{T}}} \otimes \left(\widehat{\beta}^{IV1,(I)}_{Y|1\underline{X}}\right)^t$, $I_{d_{\widetilde{T}}}$ is the identity matrix of dimension $d_{\widetilde{T}} = \text{dim}(\widetilde{T})$, and $\widehat{\beta}^{IV1,(I)}_{Y|1\underline{X}}$ is the estimator obtained by setting M_1 equal to $I_{d_{\widetilde{T}}}$.

For an asymptotically efficient estimator (6.19), replace M_2 with

$$\begin{aligned}\widehat{M}_{2,\text{opt}} = & \left\{(H_1 + \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})\widehat{\Omega}_{4,4}(H_1 + \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})^t + \right. \\ & \left.(H_1 + \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})\widehat{\Omega}_{4,7}\widehat{D}^t + \widehat{D}\widehat{\Omega}_{7,4}(H_1 + \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})^t + \widehat{D}\widehat{\Omega}_{7,7}\widehat{D}^t\right\}^{-1},\end{aligned}$$

where $\widehat{D} = I_{d_{\widetilde{T}}} \otimes (H_2\widehat{\beta}_{Y|\underline{\widetilde{T}W}})^t - I_{d_{\widetilde{T}}} \otimes \widehat{\beta}^{IV2,(I)}_{Y|1\underline{X}})^t$ and $\widehat{\beta}^{IV2,(I)}_{Y|1\underline{X}}$ is the estimator obtained by setting M_2 equal to $I_{d_{\widetilde{T}}}$.

We now describe the main steps in the demonstrations of the asymptotic efficiency of $\widehat{M}_{1,\text{opt}}$ and $\widehat{M}_{2,\text{opt}}$.

The argument for $\widehat{M}_{1,\text{opt}}$ and the estimator (6.18) is simpler and is given first. We start with a heuristic derivation of the efficient estimator.

Consider the basic identity in (6.17), $\beta_{Y|\underline{\widetilde{T}}} = \beta_{\widetilde{W}|\underline{\widetilde{T}}}\beta_{Y|\underline{\widetilde{X}}}$. Replacing $\beta_{Y|\underline{\widetilde{T}}}$ with $\widehat{\beta}_{Y|\underline{\widetilde{T}}} - (\widehat{\beta}_{Y|\underline{\widetilde{T}}} - \beta_{Y|\underline{\widetilde{T}}})$ and $\beta_{\widetilde{W}|\underline{\widetilde{T}}}$ with $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}} - (\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}} - \beta_{\widetilde{W}|\underline{\widetilde{T}}})$ and rearranging terms shows that this equation is equivalent to

$$\widehat{\beta}_{Y|\underline{\widetilde{T}}} = \widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}\beta_{Y|\underline{\widetilde{X}}} + (\widehat{\beta}_{Y|\underline{\widetilde{T}}} - \beta_{Y|\underline{\widetilde{T}}}) - (\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}} - \beta_{\widetilde{W}|\underline{\widetilde{T}}})\beta_{Y|\underline{\widetilde{X}}}.$$

This equation has the structure of a linear model with response vector $\widehat{\beta}_{Y|\underline{\widetilde{T}}}$, design matrix $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}$, regression parameter $\beta_{Y|\underline{\widetilde{X}}}$, and equation error $(\widehat{\beta}_{Y|\underline{\widetilde{T}}} - \beta_{Y|\underline{\widetilde{T}}}) - (\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}} - \beta_{\widetilde{W}|\underline{\widetilde{T}}})\beta_{Y|\underline{\widetilde{X}}}$. Let Σ denote the covariance matrix of this equation error. The best linear unbiased estimator of $\beta_{Y|\underline{\widetilde{X}}}$ in this pseudolinear model is

$$(\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}\Sigma^{-1}\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}})^{-1}\widehat{\beta}^t_{\widetilde{W}|\underline{\widetilde{T}}}\Sigma^{-1}\widehat{\beta}_{Y|\underline{\widetilde{T}}},$$

which is exactly (6.18) with $M_1 = \Sigma^{-1}$. Note that the estimator $\widehat{M}_{1,\text{opt}}$ is a consistent estimator of Σ^{-1}.

Showing that the heuristic derivation is correct and that there is no penalty for using an estimated covariance matrix is somewhat more involved, but entails nothing more than linearization via Taylor series approximations and Δ-method arguments.

Let $\widehat{M}_1$ be a consistent estimator of the matrix M_1. Expanding the estimating equation for $\widehat{\beta}^{IV1,(\widehat{M}_1)}_{Y|1\underline{X}}$ around the true parameters results in the approximation

$$\sqrt{n}\left\{\widehat{\beta}^{IV1,(\widehat{M}_1)}_{Y|1\underline{X}} - \beta_{Y|\underline{\widetilde{X}}}\right\} \approx \beta^{-(M_1)}_{\widetilde{W}|\underline{\widetilde{T}}}\left(\epsilon_2 - C\epsilon_3\right),$$

where

$$\begin{aligned}
\epsilon_2 &= \sqrt{n}\left(\widehat{\beta}_{Y|\underline{\widetilde{T}}} - \beta_{Y|\underline{\widetilde{T}}}\right), \\
\epsilon_3 &= \sqrt{n}\left\{\text{vec}\left(\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}\right) - \text{vec}\left(\beta_{\widetilde{W}|\underline{\widetilde{T}}}\right)\right\}, \\
C &= I_{d_{\widetilde{T}}} \otimes \beta^t_{Y|\underline{\widetilde{X}}}.
\end{aligned}$$

This Taylor series approximation is noteworthy for the fact that it is the same for M_1 known as it is for M_1 estimated. Consequently, there is no penalty asymptotically for estimating M_1.

Thus, with AVAR denoting asymptotic variance, we have that

$$\text{AVAR}\left\{\sqrt{n}\,\widehat{\beta}^{IV1,(\widehat{M}_1)}_{Y|1\underline{X}}\right\} = \beta^{-(M_1)}_{\widetilde{W}|\underline{\widetilde{T}}}\left\{\text{AVAR}\left(\epsilon_2 - C\epsilon_3\right)\right\}\left(\beta^{-(M_1)}_{\widetilde{W}|\underline{\widetilde{T}}}\right)^t.$$

That this asymptotic variance is minimized when

$$M_1 = \left\{\text{AVAR}\left(\epsilon_2 - C\epsilon_3\right)\right\}^{-1}$$

is a consequence of the optimality of weighted-least squares linear regression.

Let $\widehat{M}_2$ be a consistent estimator of the matrix M_2. Expanding the estimating equation for $\widehat{\beta}^{IV2,(\widehat{M}_2)}_{Y|1\underline{X}}$ around the true parameters results in the approximation

$$\sqrt{n}\left\{\widehat{\beta}^{IV2,(\widehat{M}_2)}_{Y|1\underline{X}} - \beta_{Y|\underline{\widetilde{X}}}\right\} \approx \beta^{-(M_2)}_{\widetilde{W}|\underline{\widetilde{T}}}\left\{\left(H_1 + \beta_{\widetilde{W}|\underline{\widetilde{T}}}H_2\right)\epsilon_1 + D\epsilon_3\right\},$$

where

$$\begin{aligned}
\epsilon_1 &= \sqrt{n}\left(\widehat{\beta}_{Y|\underline{\widetilde{T}W}} - \beta_{Y|\underline{\widetilde{T}W}}\right), \\
D &= I_{d_{\widetilde{T}}} \otimes \left(H_2\beta_{Y|\underline{\widetilde{T}W}}\right)^t - I_{d_{\widetilde{T}}} \otimes \beta^t_{Y|\underline{\widetilde{X}}}.
\end{aligned}$$

As before, estimating M_2 does not affect the asymptotic distribution of

the parameter estimates. From the approximation we find that

$$\begin{aligned}
&\text{AVAR}\left(\sqrt{n}\,\widehat{\beta}^{IV2,(\widehat{M}_2)}_{Y|1\underline{X}}\right) = \\
&\qquad \beta^{-(M_2)}_{\widetilde{W}|\underline{\widetilde{T}}}\left[\text{AVAR}\left\{\left(H_1 + \beta_{\widetilde{W}|\underline{\widetilde{T}}}H_2\right)\epsilon_1 + D\epsilon_3\right\}\right]\left(\beta^{-(M_2)}_{\widetilde{W}|\underline{\widetilde{T}}}\right)^t,
\end{aligned}$$

which is minimized when

$$M_2 = \left[\text{AVAR}\left\{\left(H_1 + \beta_{\widetilde{W}|\underline{\widetilde{T}}}H_2\right)\epsilon_1 + D\epsilon_3\right\}\right]^{-1}.$$

B.5.2.2 Computing Estimates and Standard Errors

The two-stage estimates are only slightly more difficult to compute than the first-stage estimates. Here, we describe an algorithm that results in both estimates.

Note that for fixed matrices M_1 and M_2 all of the components of $\widehat{\Theta}$ in (B.39) either are calculated directly as linear regression or generalized linear regression estimates, or are simple transformations of such estimates. So for fixed M_1 and M_2, obtaining $\widehat{\Theta}$ is straightforward.

Asymptotic variance estimation is most easily accomplished by first programming the two functions

$$\begin{aligned}
G_1(\Theta) &= \textstyle\sum_{i=1}^n \psi_i(\Theta), \\
G_2(\Theta) &= \textstyle\sum_{i=1}^n \psi_i(\Theta)\psi_i(\Theta)^t,
\end{aligned}$$

where $\psi_i(\Theta)$ is the i^{th} composite score function from (B.40). Although we do not actually solve $G_1(\Theta) = 0$ to find $\widehat{\Theta}$, it should be true that $G_1(\widehat{\Theta}) = 0$. This provides a check on the programming of G_1.

Numerical differentiation of G_1 at $\Theta = \widehat{\Theta}$ results in the matrix $\widehat{A}_n$. Alternatively, analytical derivatives of $\psi_i(\Theta)$ can be used, but these are complicated and tedious to program. Evaluation of G_2 at $\Theta = \widehat{\Theta}$ is the matrix $\widehat{B}_n$. The covariance matrix of $\widehat{\Theta}$ is then found as $\widehat{\Omega} = \widehat{A}_n^{-1}\widehat{B}_n(\widehat{A}_n^{-1})^t$.

The algorithm described above is first used with M_1 and M_2 set to the identity matrix of dimension $\dim(\widetilde{\mathbf{T}})$, resulting in the first-stage estimates and estimated asymptotic covariance matrix. Next, M_1 and M_2 are set to $\widehat{M}_{1,\text{opt}}$ and $\widehat{M}_{2,\text{opt}}$, respectively, as described in Section B.5.2.1. A second implementation of the algorithm results in the second-stage estimates and estimated asymptotic covariance matrix.

B.6 Appendix to Chapter 7: Score Function Methods

B.6.1 Technical Complements to Conditional Score Theory

We first justify (7.18). The joint density of $\mathbf{Y}$ and $\mathbf{W}$ is the product of (7.14) and the normal density, and hence is proportional to

$$\begin{aligned} &\propto \exp\left\{\frac{y\eta - \mathcal{D}(\eta)}{\phi} + c(y,\phi) - (1/2)(w-x)^t\Sigma_{uu}^{-1}(w-x)\right\} \\ &\propto \exp\left\{y(\beta_0 + \beta_z^t z)/\phi + c(y,\phi) - \right. \\ &\qquad (1/2)w^t\Sigma_{uu}^{-1}w + x^t\Sigma_{uu}^{-1}(w + y\Sigma_{uu}\beta_x/\phi\left.\right\}, \end{aligned}$$

where by $\sim$ we mean terms that do not depend on y or w. Now set $\delta = w + y\Sigma_{uu}\beta_x/\phi$ and make a change of variables (which has Jacobian 1). The joint density of $(\mathbf{Y},\Delta)$ given $(\mathbf{Z},\mathbf{X})$ is thus seen to be proportional to

$$\begin{aligned} &\propto \exp\left\{y(\beta_0 + \beta_x^t\delta + \beta_z^t z)/\phi + c(y,\phi) - (1/2)(y/\phi)^2\beta_x^t\Sigma_{uu}\beta_x\right\} \\ &= \exp\left\{y\eta_*/\phi + c_*(y,\phi,\beta_x^t\Sigma_{uu}\beta_x)\right\}. \qquad \text{(B.41)} \end{aligned}$$

The conditional density of $\mathbf{Y}$ given $(\mathbf{Z},\mathbf{X},\Delta)$ is (B.41) divided by its integral with respect to y, which is necessarily in the form (7.18) as claimed, with

$$\begin{aligned} &\mathcal{D}_*(\eta_*,\phi,\beta_x^t\Sigma_{uu}\beta_x) = \\ &\quad \phi\log[\int \exp\left\{y\eta_*/\phi + c_*(y,\phi,\beta_x^t\Sigma_{uu}\beta_x)\right\}\ d\mu(y)], \qquad \text{(B.42)} \end{aligned}$$

where as before the notation means that (B.42) is a sum if $\mathbf{Y}$ is discrete and an integral otherwise.

B.6.2 Technical Complements to Distribution Theory for Estimated Σ_{uu}

Next we justify the estimated standard errors for $\widehat{\Theta}$ when there is partial replication. Recall that with normally distributed observations, the sample mean and the sample covariance matrix are independent. Hence, $\widehat{\Sigma}_{uu}$ and $\widehat{\gamma} = \text{vech}(\widehat{\Sigma}_{uu})$ are independent of all the terms $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \overline{\mathbf{U}}_{i\cdot})$ and also independent of $(\mathbf{Y}_i, \mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot})$. By a Taylor series expansion,

$$\begin{aligned} &A_n(\cdot)\left(\widehat{\Theta} - \Theta\right) \approx \\ &\quad \textstyle\sum_{i=1}^n \left\{\Psi_C(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu})\right\} + D_n(\Theta,\Sigma_{uu})\left(\widehat{\gamma} - \gamma\right). \end{aligned}$$

Because the two terms in the last sum are independent, the total covariance is the sum of the two covariances, namely, $B_n(\cdot) = D_n(\cdot)C_n(\cdot)D_n^t(\cdot)$, as claimed.

B.7 Appendix to Chapter 8: Likelihood and Quasilikelihood

B.7.1 Monte Carlo Computation of Integrals

If one can easily generate observations from the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ (error model) or given $(\mathbf{Z},\mathbf{W})$ (calibration model), an appealing and easily programmed Monte Carlo approximation due to McFadden (1989) can be used to compute likelihoods. The error model likelihood (8.7) can be approximated as follows. Generate on a computer a sample $(\mathbf{X}_1^s, \cdots, \mathbf{X}_N^s)$ of size N from the density $f(x|z,\widetilde{\alpha}_2)$ of $\mathbf{X}$ given $\mathbf{Z} = z$. Then for large enough N,

$$\begin{aligned} &f_{Y,W|Z}(y,w|Z,\mathcal{B},\widetilde{\alpha}_1,\widetilde{\alpha}_2) \qquad \text{(B.43)} \\ &\quad \approx N^{-1}\textstyle\sum_{i=1}^N f_{Y|Z,X}(y|z,\mathbf{X}_i^s,\mathcal{B}) f_{W|Z,X}(w|z,\mathbf{X}_i^s,\widetilde{\alpha}_1). \end{aligned}$$

The dependence of (B.43) on $\widetilde{\alpha}_2$ comes from the fact that the distribution of $\mathbf{X}$ given $\mathbf{Z}$ depends on $\widetilde{\alpha}_2$.

We approximate

$$\begin{aligned} &f_{Y|Z,W}(y|z,w,\mathcal{B},\widetilde{\gamma}) \\ &\quad = \int f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{X|Z,W}(x|z,w,\widetilde{\gamma})d\mu(x) \qquad \text{(B.44)} \end{aligned}$$

by generating a sample $(\mathbf{X}_1^s, \cdots, \mathbf{X}_N^s)$ of size N from the distribution $f_{X|Z,W}(x|z,w,\widetilde{\gamma})$ of $\mathbf{X}$ given $(\mathbf{Z} = z, \mathbf{W} = w)$. Then for large enough N,

$$f_{Y|Z,W}(y|z,w,\mathcal{B},\widetilde{\gamma}) \approx N^{-1}\textstyle\sum_{i=1}^N f_{Y|Z,X}(y|z,\mathbf{X}_i^s,\mathcal{B}).$$

This "brute force" Monte Carlo integration method is computing intensive. There are two reasons for this. First, one has to generate random observations for each value of $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)$, which may be a formidable task if the sample size is large. Second, and somewhat less important, maximum likelihood is an iterative algorithm, and one must generate simulated $\mathbf{X}$'s at each iteration. Brown and Mariano (1993) suggested that N must be fairly large compared to $n^{1/2}$ in order to eliminate the effects of Monte Carlo variance. They also suggested a modification that will be less computing intensive.

There is a practical matter with using things such as (B.43). Specifically, if (B.43) is computed at each stage of an iterative process with different random numbers, then optimization routines will tend to get confused unless N is very, very large. If, for example, $\mathbf{X}$ given $\mathbf{Z}$ is normally distributed, a better approach is to generate a fixed but large number N of standard normals once, and then simply modify these fixed numbers at each iteration to have the appropriate mean and variance.

B.7.2 Linear, Probit, and Logistic Regression

B.7.2.1 Linear Regression

In some cases, the required likelihoods can be computed exactly or very nearly so. Suppose that $\mathbf{W}$ and $\mathbf{T}$ are each normally distributed unbiased replicates of $\mathbf{X}$, being independent given $\mathbf{X}$, and each having covariance matrix Σ_{uu}. Suppose also that $\mathbf{X}$ itself is normally distributed with mean $\gamma^t\mathbf{Z}$ and covariance matrix Σ_{xx}. As elsewhere, all distributions are conditioned on $\mathbf{Z}$. In this case, in order to allow for an intercept, the first element of $\mathbf{Z}$ equals 1.0.

In normal linear regression where the response has mean $\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}$ and variance σ^2, the joint distribution of $(\mathbf{Y}, \mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ is multivariate normal with means $\beta_0 + \beta_x^t\gamma^t\mathbf{Z} + \beta_z^t\mathbf{Z}$, $\gamma^t\mathbf{Z}$ and $\gamma^t\mathbf{Z}$, respectively, and covariance matrix

$$\Sigma_{y,w,t} = \begin{bmatrix} \sigma^2 + \beta_x^t\Sigma_{xx}\beta_x & \beta_x^t\Sigma_{xx} & \beta_x^t\Sigma_{xx} \\ \Sigma_{xx}\beta_x & \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xx} \\ \Sigma_{xx}\beta_x & \Sigma_{xx} & \Sigma_{xx} + \Sigma_{uu} \end{bmatrix}.$$

B.7.2.2 Distribution of $\mathbf{X}$ *Given the Observed Data*

For probit and logistic regression, we compute the joint density using the formulas $f_{Y,W|Z} = f_{Y|Z,W}f_{W|Z}$ and $f_{Y,W,T|Z} = f_{Y|Z,W,T}f_{W,T|Z}$. This requires a few preliminary calculations.

First, consider $\mathbf{W}$ alone. Our model says that $\mathbf{W}$ given $\mathbf{Z}$ is normally distributed with mean $\alpha_{21}^t\mathbf{Z}$ and covariance matrix $\Sigma_{xx} + \Sigma_{uu}$. Define $\Lambda_w = \Sigma_{xx}(\Sigma_{xx} + \Sigma_{uu})^{-1}$, $m(\mathbf{Z}, \mathbf{W}) = (I - \Lambda_w)\gamma^t\mathbf{Z} + \Lambda_w\mathbf{W}$ and $\Sigma_{x|z,w} = (I - \Lambda_w)\Sigma_{xx}$. From linear regression theory, for example, see Section A.4, $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is normally distributed with mean $m(\mathbf{Z}, \mathbf{W})$ and covariance matrix $\Sigma_{x|z,w}$.

Next, consider $\mathbf{W}$ and $\mathbf{T}$ together. Our model says that given $\mathbf{Z}$ they are jointly normally distributed with common mean $\gamma_1^t\mathbf{Z}$, common individual covariances $(\Sigma_{xx} + \Sigma_{uu})$, and cross-covariance matrix Σ_{xx}. If we define

$$\Lambda_{w,t} = (\Sigma_{xx}, \Sigma_{xx}) \begin{bmatrix} \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xx} \\ \Sigma_{xx} & \Sigma_{xx} + \Sigma_{uu} \end{bmatrix}^{-1} = (\Sigma_{xx}, \Sigma_{xx})\Gamma_{w,t}^{-1},$$

then $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$ is normally distributed with mean and covariance matrix given by

$$\begin{aligned} m(\mathbf{Z}, \mathbf{W}, \mathbf{T}) &= \gamma^t\mathbf{Z} \\ &\quad + \Lambda_{w,t}\left\{(\mathbf{W} - \gamma^t\mathbf{Z})^t, (\mathbf{T} - \gamma^t\mathbf{Z})^t\right\}^t \\ \Sigma_{x|z,w,t} &= \Sigma_{xx} - \Lambda_{w,t}(\Sigma_{xx}, \Sigma_{xx})^t, \end{aligned}$$

respectively.

B.7.2.3 Probit and Logistic Regression

Now we return to probit and logistic regression. In probit regression, exact statements are possible. We have indicated that given either $(\mathbf{Z}, \mathbf{W})$ or $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$, $\mathbf{X}$ is normally distributed with mean $m(\cdot)$ and covariance matrix $\Sigma_{x|\cdot}$, where $\Sigma_{x|\cdot}$is either $\Sigma_{x|z,w}$ or $\Sigma_{x|z,w,t}$, and similarly for $m(\cdot)$. From the calculations in Section B.3, it follows that

$$\mathrm{pr}(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{W}, \mathbf{T}) = \Phi\left[\frac{\beta_0 + \beta_x^t m(\cdot) + \beta_z^t\mathbf{Z}}{(1 + \beta_x^t\Sigma_{x|\cdot}\beta_x)^{1/2}}\right].$$

For logistic regression (Section 4.8.2), a good approximation is

$$\mathrm{pr}(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{W}, \mathbf{T}) \approx H\left[\frac{\beta_0 + \beta_x^t m(\cdot) + \beta_z^t\mathbf{Z}}{(1 + \beta_x^t\Sigma_{x|\cdot}\beta_x/1.7^2)^{1/2}}\right]; \tag{B.45}$$

see also Monahan and Stefanski (1992).

Write $\Theta = (\mathcal{B}, \Sigma_{uu}, \alpha_{21}, \Sigma_{xx})$ and $r(\mathbf{W}) = r(\mathbf{W}, \alpha_{21}) = (\mathbf{W} - \alpha_{21}^t\mathbf{Z})$. Using (B.45), except for a constant in logistic regression, the logarithm of the approximate likelihood for $(\mathbf{Y}, \mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ is

$$\begin{aligned} \ell(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{T}, \Theta) = &-(1/2)\log\{\det(\Gamma_{w,t})\} \\ &+ \mathbf{Y}\log\{H(\cdot)\} + (1 - \mathbf{Y})\log\{1 - H(\cdot)\} \\ &-(1/2)\left\{r^t(\mathbf{W}), r^t(\mathbf{T})\right\}\Gamma_{w,t}^{-1}\left\{r^t(\mathbf{W}), r^t(\mathbf{T})\right\}^t. \end{aligned} \tag{B.46}$$

A similar result applies if only $\mathbf{W}$ is measured, namely,

$$\begin{aligned} \ell(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta) = &-(1/2)\log\{\det(\Sigma_{xx} + \Sigma_{uu})\} \\ &+ \mathbf{Y}\log\{H(\cdot)\} + (1 - \mathbf{Y})\log\{1 - H(\cdot)\} \\ &-(1/2)r^t(\mathbf{W}, \gamma_1)\left(\Sigma_{uu} + \Sigma_{xx}\right)^{-1}r(\mathbf{W}, \gamma_1). \end{aligned}$$

B.8 Appendix to Chapter 9: Bayesian Methods

B.8.1 Code for Section 9.8.1

```
model
{#BEGIN MODEL
for (i in 1:Nobservations)
    {#BEGIN FOR i in 1:Nobservations

    #Outcome model
    Y[i]~dnorm(meanY[i],taueps)
    meanY[i]<-beta[1]+beta[2]*X[i]+beta[3]*Z[i]

    #Replication model
    for (j in 1:Nreplications) {W[i,j]~dnorm(X[i],tauu)}
```

```
        #Exposure model
        X[i]~dnorm(meanX[i],taux)
        meanX[i]<-alpha[1]+alpha[2]*Z[i]
        }#END FOR i in 1:Nobservations

    #Noninformative priors on the model parameters
    tauu~dgamma(3,1)
    taueps~dgamma(3,1)
    taux~dgamma(3,1)

    #Priors for alpha and beta
    for (i in 1:nalphas){alpha[i]~dnorm(0,1.0E-6)}
    for (i in 1:nbetas){beta[i]~dnorm(0,1.0E-6)}

    #Deterministic transformations: standard deviations
    sigmaeps<-1/sqrt(taueps)
    sigmau<-1/sqrt(tauu)
    sigmax<-1/sqrt(taux)

    #Deterministic transformation: attenuation
    lambda<-tauu/(tauu+taux)
    }#END MODEL
```

B.8.2 Code for Section 9.11

```
    model
    {#BEGIN MODEL
    for (i in 1:Nobservations)
        {#BEGIN for i in 1:Nobservations

        #Outcome model (repeated observations of FFQ)
        logFFQ1[i]~dnorm(meanlogFFQ[i],taueps)
        logFFQ2[i]~dnorm(meanlogFFQ[i],taueps)

        #Model for mean of log FFQ
        meanlogFFQ[i]~dnorm(meanmeanlogFFQ[i],taur)

        #Define the fixed effects part of the mean FFQ
        meanmeanlogFFQ[i]<-beta[1]+beta[2]*X[i]

        #Biomarker model for log protein
        logprotein1[i]~dnorm(X[i],tauu)
```

```
        logprotein2[i]~dnorm(X[i],tauu)

        X[i]~dnorm(meanX[i],taux)
        meanX[i]<-alpha[1]+alpha[2]*AGE[i]+alpha[3]*BMI[i]
        }#END for i in 1:Nobservations

    #Define lambda (a noninformative prior is assigned)
    tauu<-lambda*taux/(1-lambda)

    #Noninformative priors on the model parameters
    lambda~dunif(0,1)
    taueps~dgamma(3,0.1)
    taux~dgamma(3,0.1)
    taur~dgamma(3,0.1)

    #Define the signal attenuation
    attenuation<-beta[2]/(pow(beta[2],2)+taux/taur+taux/taueps)

    #Priors for the fixed effects
    for (i in 1:nalphas){alpha[i]~dnorm(0,1.0E-6)}
    for (i in 1:nbetas){beta[i]~dnorm(0,1.0E-6)}

    #Deterministic transformations (obtain variances)
    sigma2eps<-1/taueps
    sigma2x<-1/taux
    sigma2u<-1/tauu
    sigma2r<-1/taur
}#END MODEL
```

References

Abrevaya, J., & Hausman, J. A. (2004). Response error in a transformation model with an application to earnings equation estimation. *Econometrics Journal*, 7, 366–388.

Aitkin, M., & Rocci, R. (2002). A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, 12, 163–174.

Albert, P. S. (1999). A mover-stayer model for longitudinal marker data. *Biometrics*, 55, 1252–1257.

Albert, P. S., & Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60, 427–435.

Amemiya, Y. (1985). Instrumental variable estimator for the nonlinear errors in variables model. *Journal of Econometrics*, 28, 273–289.

Amemiya, Y. (1990a). Instrumental variable estimation of the nonlinear measurement error model. In P. J. Brown & W. A. Fuller, (Eds.), *Statistical Analysis of Measurement Error Models and Application*, Providence, RI: American Mathematics Society.

Amemiya, Y. (1990b). Two stage instrumental variable estimators for the nonlinear errors in variables model. *Journal of Econometrics*, 44, 311–332.

Amemiya, Y., & Fuller, W. A. (1988). Estimation for the nonlinear functional relationship. *Annals of Statistics*, 16, 147–160.

ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: Design and objectives. *International Journal of Epidemiology*, 129, 687–702.

Armstrong, B. (1985). Measurement error in generalized linear models. *Communications in Statistics, Series B*, 14, 529–544.

Augustin, T. (2004). An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics*, 31, 43–50.

Baker, S. G., Wax, Y., & Patterson, B. H. (1993). Regression analysis of grouped survival data: Informative censoring and double sampling. *Biometrics*, 49, 379–389.

Beaton, G. H., Milner, J., & Little, J. A. (1979). Sources of variation in 24-hour dietary recall data: Implications for nutrition study design and interpretation. *American Journal of Clinical Nutrition*, 32, 2546–2559.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis,* 2nd ed., New York: Springer-Verlag.

Berkson, J. (1950). Are there two regressions? *Journal of the American Sta-*

tistical Association, 45, 164–180.

Berry, S. A., Carroll, R. J., & Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97, 160–169.

Bickel, P. J., & Ritov, Y. (1987). Efficient estimation in the errors in variables model. *The Annals of Statistics*, 15, 513–540.

Boggs, P. T., Spiegelman, C. H., Donaldson, J. R., & Schnabel, R. B. (1988). A computational examination of orthogonal distance regression. *Journal of Econometrics*, 38, 169–201.

Bowman, A. W., & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford: Clarendon Press.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Breslow, N., & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika*, 82, 81-91.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics*, 7, 434–455

Brown, B. W., & Mariano, R. S. (1993). Stochastic simulations for inference in nonlinear errors-in-variables models. *Handbook of Statistics*, Vol. 11, 611–627. North Holland: New York.

Buonaccorsi, J. P. (1991). Measurement error, linear calibration and inferences for means. *Computational Statistics and Data Analysis*, 11, 239–257.

Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model. *Journal of the American Statistical Association*, 91, 633–642.

Buonaccorsi, J. P., & Tosteson, T. (1993). Correcting for nonlinear measurement error in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods*, 22, 2687–2702.

Buonaccorsi, J. P., Demidenko, E., & Tosteson, T. (2000). Estimation in longitudinal random effects models with measurement error. *Statistica Sinica*, 10, 885–903.

Buonaccorsi, J. P., Laake, P., & Veirød, M. (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, 61, 831–836.

Burr, D. (1988). On errors-in-variables in binary regression—Berkson case. *Journal of the American Statistical Association*, 83, 739–743.

Buzas, J. S. (1997). Instrumental variable estimation in nonlinear measurement error models. *Communications in Statistics, Part A*, 26, 2861–2877.

Buzas, J. S., & Stefanski, L. A. (1996a). A note on corrected score estimation. *Statistics & Probability Letters*, 28 , 1–8.

Buzas, J. S., & Stefanski, L. A. (1996b). Instrumental variable estimation in a probit measurement error model. *Journal of Statistical Planning and Inference*, 55, 47–62.

Buzas, J. S., & Stefanski, L. A. (1996c). Instrumental variables estimation in generalized linear measurement error models. *Journal of the American Statistical Association*, 91, 999–1006.

Buzas, J. S., Stefanski, L. A., & Tosteson, T. D. (2004). Measurement Error. In

W. Ahrens & I. Pigeot (Eds.), *Handbook of Epidemiology*, London: Springer.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis,* 2nd ed., London & New York: Chapman & Hall.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8, 1075–1093.

Carroll, R. J. (1997). Surprising effects of measurement error on an aggregate data estimation. *Biometrika*, 84, 231–234.

Carroll, R. J. (1999). Risk assessment with subjectively derived doses. In E. Ron & F. O. Hoffman (Eds.), *Uncertainties in Radiation Dosimetry and Their Impact on Dose response Analysis*, National Cancer Institute Press.

Carroll, R. J. (2003). Variances are not always nuisance parameters: The 2002 R. A. Fisher Lecture. *Biometrics*, 59, 211–220.

Carroll, R. J., & Galindo, C. D. (1998). Measurement error, biases and the validation of complex models. *Environmental Health Perspectives*, 106 (Supplement 6), 1535–1539.

Carroll, R. J., & Gallo, P. P. (1982). Some aspects of robustness in functional errors-in-variables regression models. *Communications in Statistics, Series A, 11*, 2573–2585.

Carroll, R. J., & Gallo, P. P. (1984). Comparisons between maximum likelihood and method of moments in a linear errors-in-variables regression model. In T. J Santner & A. C. Tamhane (Eds.), *Design of Experiment: Ranking and Selection*, New York: Marcel Dekker.

Carroll, R. J., & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184–1186.

Carroll, R. J., & Hall, P. (2004). Low-order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society, Series B*, 66, 31–46.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and Weighting in Regression.* London: Chapman & Hall.

Carroll, R. J., & Ruppert, D. (1991). Prediction and tolerance intervals with transformation and/or weighting. *Technometrics*, 33, 197–210.

Carroll, R. J., & Ruppert, D. (1996). The use and misuse of orthogonal regression estimation in linear errors-in-variables models. *The American Statistician*, 50, 1–6.

Carroll, R. J., & Spiegelman, C. H. (1986). The effect of small measurement error on precision instrument calibration. *Journal of Quality Technology, 18*, 170–173.

Carroll, R. J., & Spiegelman, C. H. (1992). Diagnostics for nonlinearity and heteroscedasticity in errors in variables regression. *Technometrics*, 34, 186–196.

Carroll, R. J., & Stefanski, L. A. (1990) Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652–663.

Carroll, R. J., & Stefanski, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13, 1265–1282.

Carroll, R. J., & Stefanski, L. A. (1997). Asymptotic theory for the Simex estimator in measurement error models. In S. Panchapakesan & N. Balakrishnan (Eds.) *Advances in Statistical Decision Theory and Applications*, (pp. 151–164), Basel: Birkhauser Verlag.

Carroll, R. J., & Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, 53, 573–585.

Carroll, R. J., Eltinge, J. L., & Ruppert, D. (1993). Robust linear regression in replicated measurement error models. *Letters in Statistics & Probability*, 16, 169–175.

Carroll, R. J., Freedman, L., & Hartman, A. (1996). The use of semiquantitative food frequency questionnaires to estimate the distribution of usual intake. *American Journal of Epidemiology*, 143, 392–404.

Carroll, R. J., Freedman, L. S., & Pee, D. (1997). Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics*, 53, 1440–1451.

Carroll, R. J., Gail, M. H., & Lubin, J. H. (1993). Case-control studies with errors in predictors. *Journal of the American Statistical Association*, 88, 177–191.

Carroll, R. J., Gallo, P. P., & Gleser, L. J. (1985). Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association,* 80, 929–932.

Carroll, R. J., Hall, P. G., & Ruppert, D. (1994). Estimation of lag in misregistration problems for averaged signals. *Journal of the American Statistical Association*, 89, 219–229.

Carroll, R. J., Knickerbocker, R. K., & Wang, C. Y. (1995). Dimension reduction in semiparametric measurement error models. *The Annals of Statistics*, 23, 161–181.

Carroll, R. J., Maca, J. D., & Ruppert, D. (1998). Nonparametric regression splines for generalized linear measurement error models. In *Econometrics in Theory and Practice: Festschrift in the Honour of Hans Schneeweiss*, (pp. 23–30), Physica Verlag.

Carroll, R. J., Maca, J. D., & Ruppert, D. (1999). Nonparametric regression with errors in covariates. *Biometrika*, 86, 541–554.

Carroll, R. J., Roeder, K., & Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics*, 55, 44–54.

Carroll, R. J., Wang, C. Y., & Wang, S. (1995). Asymptotics for prospective analysis of stratified logistic case-control studies. *Journal of the American Statistical Association*, 90, 157–169.

Carroll, R. J., Wang, S., & Wang, C. Y. (1995). Asymptotics for prospective analysis of stratified logistic case-control studies. *Journal of the American Statistical Association*, 90, 157–169.

Carroll, R. J., Freedman, L. S., Kipnis, V. & Li, L. (1998). A new class of measurement error models, with applications to estimating the distribution of usual intake. *Canadian Journal of Statistics*, 26, 467–477.

Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, 91, 242–250.

Carroll, R. J., Midthune, D., Freedman, L. S., & Kipnis, V. (2006). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, to appear.

Carroll, R. J., Ruppert, D., Tosteson, T. D., Crainiceanu, C., & Karagas, M. R. (2004). Nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, 99, 736–750.

Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T., & Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika, 71*, 19–26.

Cheng, C.-L., & van Ness, J. W. (1988). Generalized M-estimators for errors in variables regression. *The Annals of Statistics*, 20, 385–397.

Cheng, C.-L., & Schneeweiss, H. (1998). Polynomial regression with errors in the variables. *Journal of the Royal Statistical Society, Series B*, 60, 189–199.

Cheng, C.-L., & Tsai, C.-L. (1992). Diagnostics in measurement error models. Unpublished.

Cheng, C.-L., Schneeweiss, H., & Thamerus, M. (2000). A small sample estimator for a polynomial regression with errors in the variables. *Journal of the Royal Statistical Society, Series B*, 62, 699–709 .

Clayton, D. G. (1991). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In J. H. Dwyer, M. Feinleib, P. Lipsert, P., et al. (Eds.), *Statistical Models for Longitudinal Studies of Health*, (pp. 301–331). New York: Oxford University Press.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W., & Devlin, S. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596–610.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, 10, 637–666.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314–1328.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society, Series B*, 50, 225–265.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Federal Proceedings*, 21, 58–61.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883–904.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman & Hall.

Crainiceanu, C., Ruppert, D., & Wand, M. (2005). Bayesian Analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software.*

Volume 14, Issue 14. (http://www.jstatsoft.org/)

Crouch, E. A., & Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: Applications to logistic-normal models. *Journal of the American Statistical Association*, 85, 464–467.

Dafni, U. & Tsiatis, A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54, 1445–1462.

Dagalp, R. E. (2001). *Estimators for Generalized Linear Measurement Error Models with Interaction Terms*. Unpublished Ph.D. thesis, North Carolina State University, Raleigh, NC.

Davidian, M., & Carroll R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1091.

Davidian, M., & Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80, 475–488.

Davis, S., Kopecky, K., & Hamilton, T. (2002). *Hanford Thyroid Disease Study Final Report*. Centers for Disease Control and Prevention (CDC).

Delaigle, A., & Gijbels, I. (2004a). Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, 45, 249–267.

Delaigle, A., & Gijbels, I. (2004b). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56, 19–47.

Delaigle, A., & Gijbels, I. (2005). Data-driven boundary estimation in deconvolution problems. *Computational Statistics and Data Analysis*, 50, 1965–1994

Delaigle, A., & I. Gijbels (2006). Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica*, to appear.

Delaigle, A., Hall, P., & Qui, P. (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *Journal of the Royal Statistical Society, Series B*, 68, 201–220.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: John Wiley & Sons.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, **76**, 341–353.

Desmond, A. F. (1989). Estimating Equations, Theory of, In S. Kotz & N. L. Johnson, (Eds.) *Encyclopedia of Statistical Sciences,* (pp. 56–59), New York: John Wiley & Sons.

Devanarayan, V. (1996). *Simulation Extrapolation Method for Heteroscedastic Measurement Error Models with Replicate Measurements*. Unpublished Ph.D. thesis, North Carolina State University, Raleigh, NC.

Devanarayan, V., & Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59, 219–225.

Dominici F., Zeger S. L., & Samet, J. M. (2000). A Measurement error model for time-series studies of air pollution and mortality. *Biostatistics*, l, 157–175.

Dosemeci, M., Wacholder, S., & Lubin, J. H. (1990). Does nondifferential misclassification of exposure always bias a true effect towards the null value? *American Journal of Epidemiology*, 132, 746–748.

Drum, M., & McCullagh, P. (1993). Comment on the paper by Fitzmaurice, Laird, & Rotnitzky. *Statistical Science*, 8, 300–301.

Eckert, R. S., Carroll, R. J., & Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics*, 53, 262–272.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Efron, B. (1994). Missing data, imputation and the bootstrap. *Journal of the American Statistical Association*, 463–475.

Efron, B. (2005). Bayesians, frequentists and scientists. *Journal of the American Statistical Association*, 100, 1–5.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expectedFisher information. *Biometrika*, 65, 457–487.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Ekholm, A., Green, M., & Palmgren, J. (1986). Fitting exponential family nonlinear models in GLIM 3.77. *GLIM Newsletter*, 13, 4–13.

Ekholm, A., & Palmgren, J. (1982). A model for a binary response with misclassification. In *GLIM-82*, editor R. Gilchrist. Heidelberg: Springer.

Ekholm, A., & Palmgren, J. (1987). Correction for misclassifiction using doubly sampled data. *Journal of Official Statistics*, 3, 419–429.

Fan, J. (1991a). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19, 1257–1272.

Fan, J. (1991b). Asymptotic normality for deconvolving kernel density estimators. *Sankhyā, Series A*, 53, 97–110.

Fan, J. (1991c). Global behavior of deconvolution kernel estimates. *Statistica Sinica*, 1, 541–551.

Fan, J. (1992a). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20, 23–37.

Fan, J., & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Fan, J., & Masry, E. (1993). Multivariate regression estimation with errors-invariables: asymptotic normality for mixing processes *Journal of Multivariate Analysis*, 43, 237–271.

Fan, J., & Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, 21, 1900–1925.

Fan, J., Truong, Y. K., & Wang, Y. (1991). Nonparametric function estimation involving errors-in-variables. In G. Roussas (Ed.), *Nonparametric Functional Estimation and Related Topics* (pp. 613–627), Dordrecht: Kluwer Academic Publishers.

Freedman, L. S., Carroll, R. J., & Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology*, 134, 510–520.

Freedman, L. S., Feinberg, V., Kipnis, V., Midthune, D., & Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60, 171–181.

Freedman, L., Schatzkin, A., & Wax, Y. (1990). The effect of dietary measurement error on the sample size of a cohort study. *American Journal of Epidemiology*, 132, 1185–1195.

Fuller, W. A. (1980). Properties of some estimators for the errors in variables model. *The Annals of Statistics*, 8, 407–422.

Fuller, W. A. (1984). Measurement error models with heterogeneous error variances. In Y. P. Chaubey & T. D. Dwivedi (Eds.) *Topics in Applied Statistics*, (pp. 257–289). Montreal: Concordia University.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Fung, K. Y., & Krewski, D. (1999). On measurement error adjustment methods in Poisson regression. *Environmetrics*, 10, 213–224.

Gail, M. H., Tan, W. Y., & Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75, 57–64.

Gail, M. H., Wieand, S. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71, 431–444.

Gallant, A. R., & Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica*, 55, 363–390.

Gallo, P. P. (1982). Consistency of some regression estimates when some variables are subject to error. *Communications in Statistics, Series A*, 11, 973–983.

Ganguli, B., Staudenmayer, J., & Wand, M. P. (2005). Additive models with predictors subject to measurement error. *Australian and New Zealand Journal of Statistics*, 47, 193–202.

Ganse, R. A., Amemiya, Y., & Fuller, W. A. (1983). Prediction when both variables are subject to error, with application to earthquake magnitude. *Journal of the American Statistical Association*, 78, 761–765.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., & Rubin, D. R. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. London & New York: Chapman & Hall.

Geman, S., & Geman D. (1984). Stochastic relaxation,, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte-Carlo. *Statistical Science*, 7, 473–511.

Geyer, C. J., & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B*, 54, 657–700.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London & New York: Chapman & Hall.

Gleser, L. J. (1981). Estimation in a multivariate errors in variables regression model: Large sample results. *The Annals of Statistics*, 9, 24–44.

Gleser, L. J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In P. J. Brown & W. A. Fuller (Eds.) *Statistical Analysis of Measurement Error Models and Application*. American Mathematics Society, Providence.

Gleser, L. J. (1992). The importance of assessing measurement reliability in multivariate regression, *Journal of the American Statistical Association*, 87, 696–707.

Gleser, L. J., Carroll, R. J., & Gallo, P. P. (1987). The limiting distribution of least squares in an errors-in-variables linear regression model. *Annals of Statistics*, 15, 220–233.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208–1211.

Godambe, V. P. (1991). *Estimating functions*. New York: Clarendon Press.

Gong, G., & Samaniego, F. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 9, 861–869.

Gössi, C., & Küchenhoff, H. (2001). Bayesian analysis of logistic regression with an unknown change point and covariate measurement error. *Statistics in Medicine*, 20, 3109–3121.

Gould, W. R., Stefanski, L. A., & Pollock, K. H. (1997). Effects of measurement error on catch-effort estimation. *Canadian Journal of Fisheries and Aquatic Science*, 54, 898–906.

Gould, W. R., Stefanski, L. A., & Pollock, K. H. (1999). Use of simulation-extrapolation estimation in catch-effort analyses. *Canadian Journal of Fisheries and Aquatic Sciences*, 56, 1234–1240.

Gray H. L., Watkins, T. A., & Schucany W. R. (1973). On the jackknife statistic and its relation to UMVU estimators in the normal case. *Communications in Statistics, Theory & Methods*, 2, 285–320.

Green, P., & Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. New York: Chapman & Hall.

Greene, W. F., & Cai, J. (2004). Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, 60, 987–996.

Greenland, S. (1988a). Statistical uncertainty due to misclassification: Implications for validation substudies. *Journal of Clinical Epidemiology*, 41, 1167–1174.

Greenland, S. (1988b). On sample size and power calculations for studies using confidence intervals. *American Journal of Epidemiology*, 128, 231–236.

Greenland, S. (1988c). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, 7, 745–757.

Greenland, S., & Kleinbaum, D. G. (1983). Correcting for misclassification in two-way tables and pair-matched studies. *International Journal of Epidemiology*, 12, 93–97.

Griffiths, P., & Hill, I. D. (1985). *Applied Statistics Algorithms*. London: Horwood.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton: CRC/Chapman & Hall.

Gustafson, P. (2005) On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables (with discussion),. *Statistical Science*, 20, 111–140.

Gustafson, P., Le, N. D., & Vallée, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatistics*, 3, 229–243.

Hall, P. (1989). On projection pursuit regression. *The Annals of Statistics*, 17, 573–588.

Hall, P. G. (1992). *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Härdle, W. (1990). *Applied Nonparametric Regression.* New York: Cambridge University Press.

Hanfelt, J. J. (2003). Conditioning to reduce the sensitivity of general estimating functions to nuisance parameters. *Biometrika*, 90, 517–531.

Hanfelt, J. J., & Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, 82, 461–477.

Hanfelt, J. J. ,& Liang, K. Y. (1997). Approximate likelihoods for generalized linear errors-in-variables models. *Journal of the Royal Statistical Society, Series B*, 59, 627–637.

Hansen, M. H., & Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion). *Statistical Science*, 17, 2–51.

Hasenabeldy, N., Fuller, W. A., & Ware, J. (1988). Indoor air pollution and pulmonary performance: investigating errors in exposure assessment. *Statistics in Medicine*, 8, 1109–1126.

Hastie, T., & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models*, New York: Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Hausman, J. A., Newey, W. K., Ichimura, H., & Powell, J. L. (1991). Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics*, 50, 273–295.

Heagerty, P. J. & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika*, 88, 973–985.

Henderson, M. M., Kushi, L. H., Thompson, D. J., et al. (1990). Feasibility of a randomized trial of a low-fat diet for the prevention of breast cancer: Dietary compliance in the Women's Health Trial Vanguard Study. *Preventive Medicine*, 19, 115–133.

Higdon, R., & Schafer, D. W. (2001). Maximum likelihood computations for regression with measurement error. *Computational Statistics & Data Analysis*, 35, 283–299.

Higgins, K. M., Davidian, M., & Giltinan, D. M. (1997). A two-step approach to measurement error in time-dependent covariates in nonlinear mixed effects models, with application to IGF-1 pharmacokinetics. *Journal of the American Statistical Association*, 92, 436–448.

Hildesheim, A., Mann, V., Brinton, L. A., Szklo, M., Reeves, W. C., & Rawls, W. E. (1991). Herpes Simplex Virus Type 2: A possible interaction with Human Papillomavirus Types 16/18 in the development of invasive cervical cancer. *International Journal of Cancer*, 49, 335–340.

Holcomb, J.P. (1999). Regression with covariates and outcome calculated from a common set of variables measured with error: Estimation using the SIMEX method. *Statistics in Medicine*, 18, 2847–2862.

Horowitz, J. L., & Markatou, M. (1993). Semiparametic estimation of regression models for panel data. Unpublished.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, 11, 271–283.

Hu, C. & Lin, D. Y. (2004). Semiparametric failure time regression with replicates of mismeasured covariates. *Journal of the American Statistical Association*, 99, 105–118.

Hu, P., Tsiatis, A. A., & Davidian, M. (1998). Estimating the parameters of the Cox model when covariate variables are measured with errors. *Biometrics*, 54, 1407–1419.

Huang, Y., & Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric correction approach. *Journal of the American Statistical Association*, 95, 1209–1219.

Huang, Y., & Wang, C. Y. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*, 96, 1469–1482.

Huang, X., Stefanski, L., & Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika*, 93, 53–64.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium*, 1, 221–233.

Hughes, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, 49, 1056–1066.

Hunter, W. G., & Lamboy, W. F. (1981). A Bayesian analysis of the linear calibration problem. *Technometrics*, 23, 323–328.

Hwang, J. T. (1986). Multiplicative errors in variables models with applications to the recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association*, 81, 680–688.

Hwang, W. H., & Huang, S.Y.H. (2003). Estimation in capture-recapture models when covariates are subject to measurement errors. *Biometrics*, 59, 1113–1122.

Hwang, J. T., & Stefanski, L. A. (1994). Monotonicity of regression functions in structural measurement error models. *Statistics & Probability Letters*, 20, 113–116.

Iturria, S., Carroll, R. J., & Firth, D. (1999). Multiplicative measurement error estimation: Estimating equations. *Journal of the Royal Statistical Society, Series B*, 61, 547–562.

Jeong, M., & Kim, C. (2003). Some properties of SIMEX estimator in partially linear measurement error model. *Journal of the Korean Statistical Society*, 32, 85–92.

Johnson, N. L., & Kotz, S. (1970). *Distributions in Statistics*, Vol. 2. Boston: Houghton-Mifflin.

Jones, D. Y., Schatzkin, A., Green, S. B., Block, G., Brinton, L. A., Ziegler, R. G., Hoover, R., & Taylor, P. R. (1987). Dietary fat and breast cancer in the National Health and Nutrition Survey I: Epidemiologic follow-up study. *Journal of the National Cancer Institute*, 79, 465–471.

Kangas, A. S. (1998). Effect of errors-in-variables on coefficients of a growth model and on prediction of growth. *Forest Ecology And Management*, 102, 203–212.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B., & Kjelsberg, M. O. (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT. *American Heart Journal*, 112, 825–836.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52, 93–100.

Kauermann, G., & Carroll, R. J. (2001). The Sandwich variance estimator: Efficiency properties and coverage probability of confidence intervals. *Journal of the American Statistical Association*, 96, 1387–1396.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69, 19–27.

Kerber, R. L., Till, J. E., Simon, S. L., Lyon, J. L. Thomas, D. C., Preston-Martin, S., Rollison, M. L., Lloyd, R. D., & Stevens, W. (1993). A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *Journal of the American Medical Association*, **270**, 2076–2083.

Ketellapper, R. H., & Ronner, R. E. (1984). Are robust estimation methods useful in the structural errors in variables model? *Metrika*, 31, 33–41.

Kim, C., Hong, C., & Jeong, M. (2000). Simulation-extrapolation via the Bezier curve in measurement error models. *Communications In Statistics—Simulation and Computation*, 29, 1135–1147.

Kim, J., & Gleser, L. J. (2000). SIMEX approaches to measurement error in ROC studies. *Communications in Statistics—Theory and Methods*, 29, 2473–2491.

Kipnis, V., Carroll, R. J., & Freedman, L. S. & Li, L. (1999). A new dietary measurement error model and its application to the estimation of relative risk: Application to four validation studies. *American Journal of Epidemiology*, 150, 642–651.

Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Day, N. E., Riboli, E., & Carroll, R. J. (2003). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition*, 5, 915–923.

Kipnis V., Midthune D., Freedman L. S., Bingham S., Schatzkin A., Subar A., & Carroll R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology*, 153, 394–403.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. Bingham, S., Schoeller, D. A., Schatzkin, A., & Carroll, R. J. (2003). The structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology*, 158, 14–21.

Ko, H., & Davidian, M. (2000) Correcting for measurement error in individual-level covariates in nonlinear mixed effects models. *Biometrics*, 56, 368–375.

Kopecky, K. J., Davis, S., Hamilton, T. E., Saporito, M. S., & Onstad, L. E. (2004). Estimation of thyroid radiation doses for the Hanford Thyroid Disease Study: Results and implications for statistical power of the epidemiological analyses. *Health Physics*, 87, 15–32.

Küchenhoff, H. (1990). *Logit- und Probitregression mit Fehlen in den Variabeln.* Frankfurt am Main: Anton Hain.

Küchenhoff, H., & Carroll, R. J. (1997). Segmented regression with errors in predictors: Semi-parametric and parametric methods. *Statistics in Medicine*, 16, 169–188.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, to appear.

Kukush, A., Schneeweiss, H., & Wolf, R. (2004). Three estimators for the Poisson regression model with measurement errors. *Statistical Papers*, 45, N3, 351–368.

Landin, R., Carroll, R. J., & Freedman, L. S. (1995). Adjusting for time trends when estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *Biometrics*, 51, 169–181.

Lechner, S., & Pohlmeier, W. (2004). To blank or not to blank? A comparison of the effects of disclosure limitation methods on nonlinear regression estimates. *Annals of the New York Academy of Sciences*, 3050, 187–200.

Li, B., & McCullagh, P. (1994). Potential functions and conservative estimating functions. *Annals of Statistics*, 22, 340–356.

Li, E., Wang, N., and Wang, N-Y. (2005). Joint models for a primary endpoint and multivariate longitudinal data. Manuscript.

Li, E., Zhang, D., & Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal parameters. *Biometrics* **60**, 1–7.

Li, E., Zhang, D., & Davidian, M. (2005). Likelihood and pseudo-likelihood methods for semiparametric joint models for a primary endpoint and longitudinal data. Manuscript.

Li, L., Palta, M., & Shao, J. (2004). A measurement error model with a Poisson distributed surrogate. *Statistics in Medicine*, 23, 2527-2536.

Li, L., Shao, J., & Palta, M. (2005). A longitudinal measurement error model with a semicontinuous covariate. *Biometrics*, 61, 828–830.

Li, Y., & Lin, X. (2000). Covariate measurement errors in frailty models for clustered survival data. *Biometrika*, 87, 849–866.

Li, Y., & Lin, X. (2003a). Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of The American Statistical Association*, 98, 191–203.

Li, Y., & Lin, X. (2003b). Testing the correlation for clustered categorical and censored discrete time-to-event data when covariates are measured without/with errors. *Biometrics*, 59, 25–35.

Liang, H. (2000). Asymptotic normality of parametric part in partly linear models with measurement error in the nonparametric part. *Journal of Statistical Planning & Inference*, 86, 51–62.

Liang, H., & Wang, N. (2005). Partially linear single-index measurement error models. *Statistica Sinica*, 15, 99–116.

Liang, H., Härdle, W., & Carroll, R. J. (1999). Large sample theory in a semiparametric partially linear errors in variables model. *The Annals of Statistics*, 27, 1519–1535.

Liang, H., Wu, H., & Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics*, 4, 297–312.

Liang, K. Y., & Liu, X. H. (1991). Estimating equations in generalized linear models with measurement error. In V. P. Godambe (Ed.) *Estimating Functions*, Oxford: Clarendon Press.

Liang, K. Y., & Tsou, D. (1992). Empirical Bayes and conditional inference with many nuisance parameters. *Biometrika*, 79, 261–270.

Liang, K. Y., & Zeger, S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* 10, 158–173.

Lin, D. Y., & Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88, 1341–1349.

Lin, X., & Breslow, N. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91, 1007–1016.

Lin, X., & Carroll, R. J. (1999). SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics*, 55, 613–619.

Lin, X., & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, 95, 520–534.

Lindley, D. V. (1953). Estimation of a functional relationship. *Biometrika*, 40, 47–49.

Lindley, D. V., & El Sayyad, G. M. (1968). The Bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society, Series B*, 30, 190–202.

Lindsay, B. G. (1982). Conditional score functions: Some optimality results. *Biometrika*, 69, 503–512.

Lindsay, B. G. (1983). The geometry of mixture likelihoods, Part I: A general theory. *The Annals of Statistics*, 11, 86–94.

Lindsay, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *The Annals of Statistics*, 13, 914–32.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.

Liu, M. C., & Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17, 399–410.

Liu, M. C., & Taylor, R. L. (1990). Simulation and computation of a nonparametric density estimator for the deconvolution problem. *Statistical Computation and Simulation*, 35, 145–167.

Liu, K., Stamler, J., Dyer, A., McKeever, J., & McKeever, P. (1978). Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. *Journal of Chronic Diseases*, 31, 399–418.

Liu, X., & Liang, K. Y. (1992). Efficacy of repeated measures in regression models with measurement error. *Biometrics*, 48, 645–654.

Lord, F. M. (1960). Large sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 55, 307–321.

Luo, M., Stokes, L., & Sager, T. (1998). Estimation of the CDF of a finite population in the presence of a calibration sample. *Environmental and Ecological Statistics*, 5, 277–289.

Lubin, J. H., Schafer, D. W. Ron, E. Stovall, M., & Carroll, R. J. (2004). A reanalysis of thyroid neoplasms in the Israeli tinea capitis study accounting for dose uncertainties. *Radiation Research*, 161, 359–368.

Lyles, R. H., & Kupper, L. L. (1997). A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics*, 53, 1008–1025.

Lyles, R., Williamson, J., Lin, H.-M., & Heilig, C. (2005). Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. *Biometrics*, 61, 287–294.

Lyles, R. H., Munoz, A., Xu, J., Taylor, J. M. G., & Chmiel, J. S. (1999). Adjusting for measurement error to assess health effects of variability in biomarkers. *Statistics in Medicine*, 18, 1069–1086.

Lyon, J. L., Alder, S. C., Stone, M. B., Scholl, A., Reading, J. C. Holubkov, R., Sheng, X. White, G. L., Hegmann, K. T., Anspaugh, L., Hoffman, F. O., Simon, S. L., Thomas, B., Carroll, R. J., & Meikle, A. W. (2006). Thyroid disease associated with exposure to the Nevada Test Site radiation: A reevaluation based on corrected dosimetry and examination data. Preprint.

Ma, Y., & Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, to appear.

MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A., & Stamler, J. (1990). Blood pressure, stroke and coronary heart disease: Part 1, prolonged differences in blood pressure: Prospective observational studies corrected for the regression dilution bias. *Lancet*, 335, 765–774.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54, 173–205.

Mallick, B. K., & Gelfand, A. E. (1996). Semiparametric errors-in-variables models: A Bayesian approach. *Journal of Statistical Planning and Inference*, 52, 307–321.

Mallick, B., Hoffman, F., & Carroll, R. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. *Biometrics*, 58, 13–20.

Marazzi, A. (1980). ROBETH, a subroutine library for robust statistical procedures. *COMPSTAT 1980, Proceedings in Computational Statistics*, Vienna: Physica.

Marcus, A. H., & Elias, R. W. (1998) Some useful statistical methods for model validation. *Environmental Health Perspectives*, 106, 1541–1550.

Marschner, I. C., Emberson, J., Irwig, L., & Walter, S. D. (2004). The number needed to treat (NNT) can be adjusted for bias when the outcome is measured with error. *Journal of Clinical Epidemiology*, 57, 1244–1252.

Marsh-Manzi, J., Crainiceanu, C. M., Astor, B. C., Powe, N. R., Klag, M. J., Taylor, H. A., & Coresh, J. (2005). Increased risk of CKD progression and ESRD in African Americans: The Atherosclerosis Risk in Communities (ARIC) Study. Submitted.

Masry, E., & Rice, J. A. (1992). Gaussian deconvolution via differentiation. *Canadian Journal of Statistics*, 20, 9–21.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 239–265.

McLeish, D. L., & Small, C. G. (1988). *The Theory and Applications of Statistical Inference Functions*. New York: Springer-Verlag.

McShane, L., Midthune, D. N., Dorgan, J. F., Freedman, L. S., & Carroll, R. J. (2001). Covariate measurement error adjustment for matched case-control studies. *Biometrics*, 57, 62–73.

Mengersen, K. L., Robert, C. P., & Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: A reviewww, In J. M. Bernardo, J. O. Berger, A. F. Dawid & A. F. M. Smith, (Eds.), *Bayesian Statistics 6* Oxford: Oxford University Press.

Miller, R. G. (1998). Survival Analysis. New York: John Wiley & Sons.

Monahan, J., & Stefanski, L. A. (1992). Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. In N. Balakrishnan (Ed.) *Handbook of the Logistic Distribution* (pp. 529–540). New York: Marcel Dekker.

Müller, H-G. (1988). *Nonparametric Analysis of Longitudinal Data.* Berlin: Springer-Verlag.

Müller, P., & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, 84, 523–537.

Nakamura, T. (1990). Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77, 127–137.

Nakamura, T. (1992). Proportional hazards models with covariates subject to measurement error. *Biometrics*, 48, 829–838.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, 58, 675–683.

Newey, W. K. (1991). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5, 99–135.

Novick, S. J., & Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97, 472–481.

Nummi, T. (2000). Analysis of growth curves under measurement errors. *Journal of Applied Statistics*, 27, 235–243.

Palmgren, J. (1987). Precision of double sampling estimators for comparing two probabilities. *Biometrika*, 74, 687–694.

Palmgren, J., & Ekholm, A. (1987). Exponential family non-linear models for categorical data with errors of observation. *Applied Stochastic Models and Data Analysis*, 3, 111–124.

Palta, M., & Lin, C.-Y. (1999). Latent variables, measurement error and methods for analysing longitudinal binary and ordinal data. *Statistics in Medicine*, 18, 385–396.

Paulino, C. D., Soares, P., & Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, 59, 670–675.

Pearson, K. (1902). On the mathematical theory of errors of judgment. *Philosophical Transactions of the Royal Society of London A*, 198, 235–299.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, 79, 355–365.

Pepe, M. S., & Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data. *Journal of the American Statistical Association*, 86, 108–113.

Pepe, M. S., Reilly, M., & Fleming, T. R. (1994). Auxilliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42, 137–160.

Pepe, M. S., Self, S. G., & Prentice, R. L. (1989). Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine*, 8, 1167–1178.

Pierce, D. A., & Kellerer, A. M. (2004). Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika*, 91, 863–876.

Pierce, D. A., Stram, D. O., Vaeth, M., & Schafer, D. (1992). Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors. *Journal of the American Statistical Association*, 87, 351–359.

Polson, N. G. (1996). Convergence of Markov chain Monte Carlo algorithms, In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics, 5* Oxford: Oxford University Press.

Polzehl, J., & Zwanzig, S. (2004). On a symmetrized simulation extrapolation estimator in linear errors-in-variables models. *Computational Statistics & Data Analysis*, 47, 675–688.

Powe N. R., Klag M. J., Sadler J. H., Anderson G. F., Bass E. B., Briggs W. A., Fink N. E., Levey A. S., Levin N. W., Meyer K. B., Rubin H. R., & Wu A. W. (1996). Choices for healthy outcomes in caring for end stage renal disease. *Seminars in Dialysis* 9, 9–11.

Prentice, R. L. (1976). Use of the logistic model in retrospective studies, *Biometrics*, 32, 599–606.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69, 331–342.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8, 431–440.

Prentice, R. L., Pepe, M., & Self, S. G. (1989). Dietary fat and breast cancer: Areview of the literature and a discussion of methodologic issues. *Cancer Research*, 49, 3147–3156.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.

Prescott, G. J., & Garthwaite, P. H. (2002). A simple Bayesian analysis of misclassified binary data with a validation substudy. *Biometrics*, 58, 454–458.

Ramalho, E. A. (2002). Regression models for choice-based samples with misclassification in the response variable. *Journal of Econometrics*, 106, 171–201.

Rao, C. R. (1947). Large-sample test of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings Cambridge Philosophical Society*, 44, 50–57.

Rathouz, P. J., & Liang, K. Y. (1999). Reducing sensitivity to nuisance parameters in semiparametric models: A quasi-score method. *Biometrika*, 86 857–869.

Reeves, G. K., Cox, D. R., Darby, S. C., & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*, 17, 2157–2177.

Reilly, M., & Pepe, M. S. (1994). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82, 299–314.

Reiser, B. (2000). Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine*, 19, 2115–2159.

Richardson, S., & Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138, 430–442.

Richardson, S., & Gilks, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, 12 , 1703–1722.

Richardson, S., Leblond, L., Jaussent, I., & Green, P. J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society, Series A*, 165, 549–566.

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy Gibbs stopper. *Journal of the American Statistical Association*, 87, 861–868.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7, 110–120.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis Hastings algorithms. *Statistical Science*, 16, 351–367.

Rocke, D.,, & Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8, 557–569.

Roeder, K., Carroll, R. J., & Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91, 722–732.

Ronchetti, E. (1982). Robust testing in linear models: The infinitesimal approach. Ph.D. Thesis. ETH, Zurich.

Rosner, B., Spiegelman, D., & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology*, 132, 734–745.

Rosner, B., Willett, W. C., & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8, 1051–1070.

Rudemo, M., Ruppert, D., & Streibig, J. C. (1989). Random effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45, 349–362.

Ruppert, D. (1985). M-estimators, In S. Kotz & N. L. Johnson (Eds.) *Encyclopedia of Statistical Sciences, vol. 5*, (pp. 443–449). New York: John Wiley & Sons.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, *Journal of the American Statistical Association*, 92, 1049–1062.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D., & Carroll, R. J. (2000). Spatially adaptive penalties for spline fitting. *Australia and New Zealand Journal of Statistics*, 42, 205–223.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D., Carroll, R. J., & Cressie, N. (1989). A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics*, 45, 637–362.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge, UK.

Satten, G. A., & Kupper, L. L. (1993). Inferences about exposure-disease association using probability of exposure information. *Journal of the American Statistical Association*, 88, 200–208.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74, 385–391.

Schafer, D. W. (1992). Replacement methods for measurement error models. Unpublished.

Schafer, D. W. (1993). Likelihood analysis for probit regression with measurement errors. *Biometrika*, 80, 899–904.

Schafer, D. W. (2002). Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computation and Simulation*, 72, 33–45.

Schafer, D., & James, I. R. (1991). Weibull regression with covariate measurement errors and assessment of unemployment duration dependence. Unpublished.

Schafer, D. W., & Purdy, K. (1996). Likelihood analysis for errors-in-variables regression with replicate measurement. *Biometrika*, 83, 813–824.

Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M., & Carroll, R. J. (2001). Thyroid cancer following scalp irradiation: A reanalysis accounting for uncertainty in dosimetry. *Biometrics*, 57, 689–697.

Schafer, D. W., Stefanski, L. A., & Carroll, R. J. (1999). Consideration of measurement errors in the international radiation study of cervical cancer. In E. Ron & F. O. Hoffman (Eds.) *Uncertainties in Radiation Dosimetry and Their Impact on Dose response Analysis*, National Cancer Institute Press.

Schafer J. (1997). *The Analysis of Incomplete Multivariate Data*, New York: Chapman & Hall/CRC.

Schatzkin, A., Kipnis, V., Subar, A. F., Midthune, D., Carroll, R. J., Bingham, S., Schoeller, D. A., Troiano, R., & Freedman, L. S. (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: Results from the biomarker-based OPEN study. *International Journal of Epidemiology*, 32, 1054–1062.

Schennach, S. M. (2006). Instrumental variables estimation of nonlinear errors-in-variables models. Preprint available at http://home.uchicago.edu/∼smschenn/cv_schennach.pdf.

Schennach, S. M. (2004a). Estimation of nonlinear models with measurement error. *Econometrica*, 72, 33–75.

Schennach, S. M. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory*, 20, 1046–1093.

Schmid, C. H., & Rosner, B. (1993). A Bayesian approach to logistic regression models having measurement error following a mixture distribution. *Statistics in Medicine*, 12, 1141–1153.

Schmid, C. H., Segal, M. R., & Rosner, B. (1994). Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *Journal of Statistical Planning and Inference*, 42, 1–18

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood criterion. *Biometrika*, 67, 93–101.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*, New York: John Wiley & Sons.

Sepanski, J. H. (1992). Score tests in a generalized linear model with surrogate covariates, *Statistics & Probability Letters*, 15, 1–10.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, London & New York: Chapman & Hall.

Simon, S. L., Till, J. E., Lloyd, R. D., Kerber, R. L., Thomas, D. C., Preston-Martin, S., Lyon, J. L., & Stevens, W. (1995). The Utah Leukemia case-control study: dosimetry methodology and results. *Health Physics*, **68**, 460–471.

Small, C. G., Wang, J., & Yang, Z. (2000). Eliminating multiple root problems in estimation, *Statistical Science*, 15, 313-341.

Smith, A. F. M., & Gelfand, A. E. (1992) Bayesian statistics without tears: A sampling-resampling perspective. *American Statistician*, 46, 84–88.

Solow, A. R. (1998). On fitting a population model in the presence of observation error. *Ecology*, 79, 1463–1466.

Song, X., & Huang, Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, 61, 702–714.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64, 583–616.

Spiegelman, C. H. (1986). Two pitfalls of using standard regression diagnostics when both X and Y have measurement error. *The American Statistician*, 40, 245–248.

Spiegelman, D. (1994). Cost-efficient study designs for relative risk modeling with covariate measurement error. *Journal of Statistical Planning and Inference*, 42, 187–208.

Spiegelman, D., & Casella, M. (1997). Fully parametric and semi-parametric regression models for common events with covariate measurement error in main study/validation study designs. *Biometrics*, 53, 395–409.

Sposto, R., Preston, D. L., Shimizu, Y., & Mabuchi, K. (1992). The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in A-bomb survivors. *Biometrics*, 48, 605–618.

Staudenmayer, J., & Spiegelman, D. (2002). Segmented regression in the presence of covariate measurement error in main study/validation study designs. *Biometrics*, 58, 871–877.

Staudenmeyer, J., & Ruppert, D. (2004). Local polynomial regression and simulation-extrapolation. *Journal of the Royal Statistical Society, Series B*, 66, 17–30.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, 72, 583–592.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, 18, 4335–4358.

Stefanski, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics & Probability Letters*, 9, 229–235.

Stefanski, L. A., & Bay, J. M. (1996). Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. *Biometrika*, 83, 407–417.

Stefanski, L. A., & Buzas, J. S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association*, 90, 541–550.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics*, 13, 1335–1351.

Stefanski, L. A., & Carroll, R. J. (1986). Deconvoluting kernel density estimators. *Statistics*, 21, 169–184.

Stefanski, L. A., & Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74, 703–716.

Stefanski, L. A., & Carroll, R. J. (1990a). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society, Series B*, 52, 345–359.

Stefanski, L. A., & Carroll, R. J. (1990b). Structural logistic regression measurement error models. In P. J. Brown & W. A. Fuller (Eds.) *Statistical analysis of measurement error models and applications: Proceedings of the AMS-IMS-SIAM joint summer research conference held June 10-16, 1989, with support from the National Science Foundation and the U.S. Army Research Office*, Providence, RI: American Mathematical Society.

Stefanski, L. A., & Carroll, R. J. (1990c). Deconvoluting kernel density estimators. *Statistics*, 21, 165–184.

Stefanski, L. A., & Carroll, R. J. (1991). Deconvolution based score tests in measurement error models. *The Annals of Statistics*, 19, 249–259.

Stefanski, L. A., & Cook, J. (1995). Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90, 1247–1256.

Stefanski, L. A., Novick, S. J., & Devanarayan, V. (2005). Estimating a nonlinear function of a normal mean. *Biometrika*, 92, 732–736.

Stephens, D. A., & Dellaportas, P. (1992). Bayesian analysis of generalized linear models with covariate measurement error. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 813–820), Oxford: Oxford University Press.

Stevens, W., Till, J. E., Thomas, D. C., et al. (1992). Assessment of leukemia and thyroid disease in relation to fallout in Utah: Report of a cohort study of thyroid disease and radioactive fallout from the Nevada test site. Salt Lake City: University of Utah.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25, 1371–1425.

Stram, D. O., & Kopecky, K. J. (2003). Power and uncertainty analysis of epidemiological studies of radiation-related disease risk in which dose estimates are based on a complex dosimetry system: some observations. *Radiation Research*, 160, 408–417.

Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D. A., Bingham, S., Sharbaugh, C. O., Trabulsi, J., Runswick, S., Ballard-Barbash, R., Sunshine, J., & Schatzkin, A. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The Observing Protein and Energy Nutrition (OPEN) study. *American Journal of Epidemiology*, 158, 1–13.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., & Rosenfeld, S. (2001). Comparative validation of the Block, Willett and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology*, 154, 1089–1099.

Tadesse, M., Ibrahim, J., Gentleman, R., Chiaretti, S., Ritz, J., & Foa, R. (2005), Bayesian error-in-variable survival model for the analysis of GeneChip arrays. *Biometrics*, 61, 488–497.

Tan, C. Y., & Iglewicz, B. (1999). Measurement-methods comparisons and linear statistical relationship. *Technometrics*, 41, 192–201.

Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.) New York: Springer-Verlag.

Taupin, M. L. (2001). Semi-parametric estimation in the nonlinear structural errors-in-variables model. *The Annals of Statistics*, 29, 66–93.

Thisted, R. A. (1988). *Elements of Statistical Computing*, New York & London: Chapman & Hall.

Thomas, D. C., Gauderman, J., & Kerber, R. (1993). A nonparametric Monte-Carlo approach to adjustment for covariate measurement errors in regression analysis. Unpublished.

Thomas, D., Stram, D., & Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health*, 14, 69–93.

Thompson, F. E., Sowers, M. F., Frongillo, E. A., & Parpia, B. J. (1992). Sources of fiber and fat in diets of U.S. women aged 19–50: Implications for nutrition education and policy. *American Journal of Public Health*, 82, 695–718.

Tosteson, T., & Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika*, 75, 507–514.

Tosteson, T. D., & Ware, J. H. (1990). Designing a logistic regression study using surrogate measures of exposure and outcome. *Biometrika*, 77, 11–20.

Tosteson, T., Buonaccorsi, J., & Demidenko, E. (1998). Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statistics in Medicine*, 17, 1959–1971.

Tosteson, T., Stefanski, L. A., & Schafer D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine*, 8, 1139–1147.

Tosteson, T., Buzas, J., Demidenko, E., & Karagas, M. (2003). Power and sample size calculations for generalized regression models with covariate measurement error. *Statistics in Medicine*, 22, 1069–1082.

Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88, 447–458.

Tsiatis, A. A., & Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* 91, 835–848.

Turnbull, B. W., Jiang, W., & Clark, L. C. (1997). Regression models for recurrent event data: Parametric random effects models with measurement error. *Statistics in Medicine*, 16, 853–864.

Ulm, K. (1991). A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine*, 10, 341–349.

United States Renal Data System. (2003). *USRDS 2003 Annual Data Report*. Bethesda, MD: National Institute of Health, National Institute of Diabetes and Digestive and Kidney Disease.

van der Vaart, A. (1988). Estimating a real parameter in a class of semiparametric models. *The Annals of Statistics*, 16, 1450–1474.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Wand, M. P. (1998). Finite sample performance of deconvolving density estimators. *Statistics & Probability Letters*, 37, 131–139.

Wang, C. Y., & Carroll, R. J. (1994). Robust estimation in case-control studies with errors in predictors. In J. O. Berger & S. S. Gupta (Eds.), *Statistical Decision Theory and Related Topics, V*, New York: Springer-Verlag

Wang, C. Y., & Pepe, M. S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, 62, 509–524.

Wang, C. Y., Wang, N., & Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56, 487–495.

Wang, C. Y., Wang, S., & Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*, 77, 65–86.

Wang, C. Y., Hsu, L., Feng, Z. D., & Prentice, R. L. (1997). Regression calibration in failure time regression, *Biometrics*, 53, 131–145.

Wang, N., & Davidian, M. (1996). A note on covariate measurement error in nonlinear mixed effects models, *Biometrics*, 83, 801–812.

Wang, N., Carroll, R. J., & Liang, K. Y. (1996). Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics*, 52, 401–411.

Wang, N. Lin, X., & Gutierrez, R. (1999). A bias correction regression calibration approach in generalized linear mixed measurement error model. *Communication in Statistics, Series A, Theory and Methods*, 28, 217–232.

Wang, N., Lin, X., Gutierrez, R. G., & Carroll, R. J. (1998). Generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93, 249–261.

Wannemuehler, K. A., & Lyles, R. H. (2005). A unified model for covariate measurement error adjustment in an occupational health study while accounting for non-detectable exposures. *Applied Statistics, Journal of the Royal Statistical Society, Series C*, 54, 259–271.

Wasserman, L., & Roeder, K. (1997). Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 90, 1247–1256.

Weinberg, C. R., Umbach, D. M., & Greenland, S. (1993). When will nondifferential misclassification preserve the direction of a trend? *American Journal of Epidemiology*, 140, 565–571.

Weinberg, C. R., & Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept models. *Biometrika*, 80, 461–465.

Whittemore, A. S. (1989). Errors in variables regression using Stein estimates. *American Statistician*, 43, 226–228.

Whittemore, A. S., & Gong, G. (1991). Poisson regression with misclassified counts: Application to cervical cancer mortality rates. *Applied Statistics*, 40, 81–93.

Whittemore, A. S., & Keller, J. B. (1988). Approximations for regression with covariate measurement error. *Journal of the American Statistical Association*, 83, 1057–1066.

Willett, W. C. (1989). An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statistics in Medicine*, 8, 1031–1040.

Willett, W. C., Meir, J. S., Colditz, G. A., Rosner, B. A., Hennekens, C. H., & Speizer, F. E. (1987). Dietary fat and the risk of breast cancer. *New England Journal of Medicine*, 316, 22–25.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H., & Speizer, F. E. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology*, 122, 51–65.

Wolter, K. M., & Fuller, W. A. (1982a). Estimation of the quadratic errors in variables model. *Biometrika*, 69, 175–182.

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *Journal of the American Statistical Association*, 97, 955–964.

Wu, M. L., Whittemore, A. S., & Jung, D. L. (1986). Errors in reported dietary intakes. *American Journal of Epidemiology*, 124, 826–835.

Zamar, R. H. (1988). Orthogonal regression M-estimators. *Biometrika*, 76, 149–154.

Zamar, R. H. (1992). Bias-robust estimation in the errors in variables model. *The Annals of Statistics*, 20, 1875–1888.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics*, 18, 806–831.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57, 795–802.

Zhao, L. P., Prentice, R. L., & Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B*, 54, 805–812.

Zhou, H., & Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, 82, 139–149.

Zhou, H., & Wang, C. Y. (2000). Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society, Series B*, 62, 657–665.

Zhu, L., & Cui, H. (2003). A Semi-parametric regression model with errors in variables. *Scandinavian Journal of Statistics*, 30, 429–442.

Zidek, J. V., Le, N. D., Wong, H., & Burnett, R. T. (1998). Including structural measurement errors in the nonlinear regression analysis of clustered data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 26, 537–548.

Zidek, J. V., White, R., Le, N. D., Sun, W., & Burnett, R. T. (1998). Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Ecological and Environmental Statistics*, 5, 99–115.