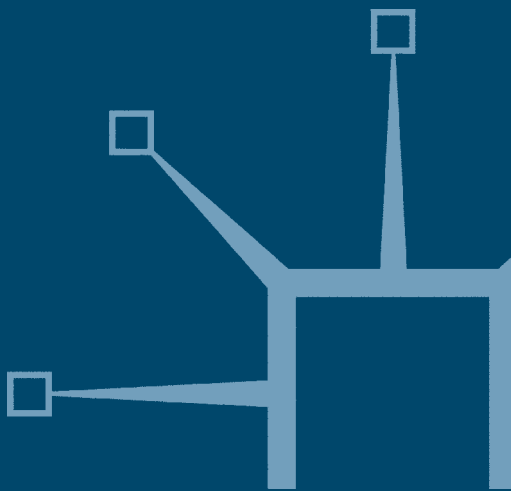
palgrave
macmillan

Evaluating Econometric Forecasts of Economic and Financial Variables

Michael P. Clements



Evaluating Econometric Forecasts of Economic and Financial Variables

Palgrave Texts in Econometrics

Series Editor: **Kerry Patterson**

Titles include:

Simon P. Burke
MODELLING NON-STATIONARY ECONOMIC TIME SERIES

Michael P. Clements
EVALUATING ECONOMETRIC FORECASTS OF ECONOMIC AND FINANCIAL VARIABLES

Terence C. Mills
MODELLING TRENDS AND CYCLES IN ECONOMIC TIME SERIES

Kerry Patterson
UNIT ROOTS IN ECONOMIC TIME SERIES

Jan Podivinsky
MODELLING VOLATILITY

Palgrave Texts in Econometrics
Series Standing Order ISBN 978-1-4039-0172-9 Hardback
Series Standing Order ISBN 978-1-4039-0173-6 Paperback
(*outside North America only*)

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and the ISBN quoted above.

Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke, Hampshire RG21 6XS, England

Evaluating Econometric Forecasts of Economic and Financial Variables

Michael P. Clements
University of Warwick

palgrave
macmillan

Softcover reprint of the hardcover 1st edition 2005 978-1-4039-4156-5

First published in 2005 by
PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York, N.Y. 10010
Companies and representatives throughout the world.

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN 978-1-4039-4157-2 ISBN 978-0-230-59614-6 (eBook)

DOI 10.1057/9780230596146

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Clements, Michael P.
Evaluating econometric forecasts of economic and financial variables / Michael P. Clements.
p. cm. – (Palgrave texts in econometrics)
Includes bibliographical references and index.

1. Economic forecasting – Econometric models – Evaluation.
I. Title. II. Series.

HB3730.C556 2005
330′.01′5195—dc22 2004054893

10 9 8 7 6 5 4 3 2 1
14 13 12 11 10 09 08 07 06 05

To Carolyn, Anna and William

This page intentionally left blank

Contents

List of Tables

List of Figures

Author's Preface and Acknowledgements

I am grateful to Kerry Patterson for encouraging me to write this book, and to Amanda Watkins at Palgrave Macmillan for her support in this venture. After co-writing two books on economic forecasting (*Forecasting Economic Time Series*, 1998, CUP; *Forecasting Non-Stationary Economic Time Series*, 1999, MIT Press) and co-editing another (*A Companion to Economic Forecasting*, 2002, Blackwells) I believed my book-writing days were over, at least for a while. But time passes and it is now six years since the first of the two books, and longer since the work for the first book was undertaken.

Much of my recent research has been on forecast evaluation, but going beyond the traditional concern with 'the most likely outcome', to consider interval forecasts and probability distributions. Writing a book on the relatively narrowly defined area of 'forecast evaluation' has allowed me to sort out and order my own thoughts on the differences and similarities that arise in the evaluation of diverse types of forecasts, and for this opportunity I certainly owe thanks to Kerry.

I have drawn on published work in preparing this book. As an acknowledgement to my co-authors and the publishers of journals, I note that material from the following articles has been included in this book. Section 2.5.5 draws on 'A Monte Carlo study of the forecasting performance of empirical SETAR models' which appeared in the *Journal of Applied Econometrics* (**14**, 1999, 123–141) (John Wiley and Sons) and was written jointly with Jeremy Smith. Sections 5.3 and 5.10 draw on 'Evaluating multivariate forecast densities: A comparison of two approaches,' published in the *International Journal of Forecasting* (**18**, 2002, 397–407) (Elsevier), also joint work with Jeremy Smith. The material in Section 4.2 draws on 'Bootstrapping prediction intervals for autoregressive models' which appeared in *International Journal of Forecasting* (**17**, 2001, 247–276) (Elsevier) and was written jointly with Nick Taylor. The material in Sections 4.5 and 4.7 was published as 'Evaluating interval forecasts for financial data' in the *Journal of Applied Econometrics* (**18**, 2003, 445–456) (John Wiley and Sons) and was also with Nick Taylor. In addition, Sections 5.7 and 6.4 to 6.7 reproduce material from 'Evaluating the Bank

of England density forecasts of inflation', published in the *Economic Journal* (**114**, 2004, 855–877), © 2004 by the Royal Economic Society (Blackwell Publishing).

As well as expressing my gratitude to Jeremy and Nick, I would also mention Philip Hans Franses, Ana Beatriz Galvão, David Harvey, David Hendry, Hans-Martin Krolzig, Marianne Sensier, Norman Swanson, Farshid Vahid, Dick van Dijk, Ken Wallis and Robert Witt, all of whom I have had the pleasure of working with in recent years. Those collaborations have undoubtedly informed the writing of this book. In particular, I wish to acknowledge the support and guidance I have received throughout my academic career from my co-author and co-editor of the earlier three ventures. David Hendry has been truly inspirational.

Finally a word on computer software. Most of the computations were carried out using programs written in the GAUSS programming language, Aptech Systems Inc. A number of the graphs were produced using GiveWin (Doornik and Hendry (2001)). GiveWin and PcGive were also used for the estimation and forecasting of the volatility models in Chapter 3. Some sample programs are included and described at the back of the book.

1
Introduction

By a forecast will be meant any statement about 'the future', where the future is relative to the analyst's viewpoint. So as well as the common sense notion of a forecast of what will happen tomorrow, or next Saturday, the term will equally apply to the outcome of the 1997 General Election made *now* but based on what was known at the end of 1996, for example. Forecasts are often constructed *ex post* as a way of evaluating a particular forecasting model or forecasting device, presumably with the hope that the past forecast performance of the model will serve as a useful guide to how well it might forecast in the future. In any event, forecasting the past as the 'relative future' means that forecasts can be evaluated as they are made, without having to wait to see what actually happens tomorrow, or on the coming Saturday, and a large sample of forecasts can be generated (with associated outcomes available), which might allow a statistical analysis of the forecast performance of the model. My forecast of rain might turn out to be wrong, but that might just be bad luck. Suppose my forecasting model is that I forecast rain in the afternoon if at 11 a.m. in the morning the cows in a certain field are lying down. Given daily observations on afternoon rainfall and the morning stance of cows over the last year, one could devise a statistical test of whether my forecasting model was a good predictor of meteorological conditions.

Forecasts can be made about anything, and using a variety of means: systems of dynamic equations, back-of-the-envelope calculations, tea-leaf dregs, goats' entrails. Our subject matter will be economic and financial variables, such as output growth rates, unemployment and inflation rates, and stock returns. The means will include econometric models and survey-based expectations. The key issue will be the evaluation of the forecasts. That is, how we judge 'good' in relation to forecasts, and how we decide whether a certain set of forecasts have this property.

The method of evaluation may depend on whether the forecasts are model-based, as well as depending on the type of forecast being made. A forecast defined as 'any statement about the future' includes statements such as: the probability that it will rain is 80% (a probability forecast); that it will rain (an event forecast); that there will be $\frac{1}{4}$ cm of rain (a point forecast); that there is a 75% probability that there will be between 0 cm and $\frac{3}{4}$ cm of rain (an interval forecast).

The material in this book will be structured by the type of forecast. Chapter 2 begins with the evaluation of point forecasts. These are typically quantitative forecasts of the level or rate of change of a continuous variable, such as the level of output or the rate of growth of output, the price level or the inflation rate, but we will also include in this chapter 'direction of change, tests, although these might be more correctly thought of as event forecasts. It is probably fair to say that the traditional concern of economic forecasting has been the production and evaluation of point forecasts, and that it is only relatively recently that there has been a general recognition that some measure of the degree of uncertainty surrounding a 'central tendency' will enhance the value or usefulness of the forecast. For example, the government might react rather differently to a point forecast that inflation will be $2\frac{1}{2}$% next year, but that the forecaster believes there is a 40% chance that it will exceed 5%, compared to the same point forecast and the assertion that the outcome will almost certainly be within $\pm\frac{1}{2}$ percentage point of $2\frac{1}{2}$%. These issues are taken up in subsequent chapters.

Chapter 3 switches attention from the evaluation of the (conditional) mean of the random variable to the evaluation of forecasts of the conditional variance of the process. For a large number of financial time series, as well as some macroeconomic time series (such as inflation), the conditional variance (or volatility) varies over time in a way that is in part predictable from the past of the process. Models of conditional variance are briefly reviewed as a precursor to a discussion of forecast evaluation. A complicating factor is that volatility is not observed.

In Chapter 4 interval forecasts or prediction intervals come under the spotlight. An interval forecast is a formal method of conveying forecast uncertainty. An interval forecast can be used to express the uncertainty surrounding a point forecast of the conditional mean, or of a volatility forecast. Viewed as an estimate of a quantile of the conditional distribution of the random variable, a one-sided interval forecast is an estimate of the 'Value-at-Risk' in the financial risk management literature.

Chapter 5 considers the evaluation of forecast densities, or forecast probability distributions. We review methods of evaluation that make no

recourse to the method of construction of the forecasts. These methods are clearly appropriate when the forecasts come from surveys, or when the models or methods underlying their construction are unknown to the econometrician. A number of recent papers have proposed the evaluation of models' forecast densities as model specification tests, and these are also reviewed.

Finally, Chapter 6 recognizes that forecasts are generally used to guide actions (or decisions) in uncertain environments, and should ideally be evaluated in terms of the benefits (or costs) that result or are expected to result from using them in this way. This approach is still in its infancy in terms of applications in macroeconomics, but an exploration of the 'decision-based' approach and its connections with more standard approaches is illuminating.

2
Point Forecasts

Sections 2.1 and 2.2 consider the evaluation of sequences of point forecasts in terms of the first- and second-moment properties of the forecast errors. Section 2.3 allows that there is at least one rival set of forecasts of the variable of interest, and asks which of the two is better, as well as whether even the less good of the two provides some useful information. In Section 2.4 we explicitly allow that the forecasts have been generated by models. At this point, the question becomes not which of the sets of forecasts is best, but which of the models generates more accurate forecasts, as judged by out-of-sample tests of predictive ability. Section 2.5 considers a number of issues that arise in the evaluation of forecasts from non-linear models.

2.1 Realization-forecast regressions

Suppose that we have a sequence of pairs of forecasts and outturns, $\{y_{t+h|t}, y_{t+h}\}$, where, for example, $t = 1, 2, \ldots, T$ and h is a fixed integer. $y_{t+h|t}$ is the forecast of the value of the variable in period $t + h$ made at time t, and y_{t+h} is the realization, or out-turn. h is the forecast horizon.

An obvious property of a good sequence of forecasts is that there is no tendency to systematically over or underpredict, that is, that the forecasts are unbiased. Formally,

$$E_t(y_{t+h|t} - y_{t+h}) = 0, \tag{2.1}$$

where E_t denotes the mathematical expectation based on information up to period t. An equivalent way of writing $E_t(y_{t+h|t} - y_{t+h})$ is $E(y_{t+h|t} - y_{t+h}|\Omega_t)$, where the information set Ω_t being conditioned

on is made explicit. Notice that (2.1) implies that the forecasts are unconditionally unbiased, $E(y_{t+h|t} - y_{t+h}) = 0$.

The condition (2.1) should hold for each t, that is, for each forecast-out-turn pair. Because only one realization of the random variable occurs at each point in time, unbiasedness is tested by whether the sample mean of the forecast errors, $e_{t+h|t} \equiv y_{t+h} - y_{t+h|t}$, over $t = 1, 2, \ldots, T$, is significantly different from zero. Weak rationality or consistency often refers to the property that forecasters are not systematically mistaken in their forecasts.

Strong rationality, or efficiency, in addition requires that the forecast errors are uncorrelated with other series or information available at the time the forecasts were made. Otherwise it would be possible to exploit these relationships to produce superior forecasts, in a sense to be defined, and the original forecasts would be inefficient. There have been many studies of the rationality of macroeconomic forecasts based on these notions, including Mincer and Zarnowitz (1969), Figlewski and Wachtel (1981), Zarnowitz (1985), Keane and Runkle (1990), and see Stekler (2002) and Fildes and Stekler (2002) for recent reviews.

Forecasts could be unbiased and efficient but highly inaccurate. Unbiasedness would result if large positive and negative errors approximately cancel, so that the sample mean of the forecast errors is close to zero. So we will also need to pay attention to the variance of the observed sample of forecast errors about the mean. This last consideration will be put to one side until we consider forecast precision in Section 2.2.

Tests of rationality are often based on regression equations of the form:

$$y_{t+1} = \alpha + \beta y_{t+1|t} + \epsilon_{t+1} \tag{2.2}$$

for $t = 1, \ldots, T$. We assume that $h = 1$ to forestall the complications that arise for multi-step ($h > 1$) forecasts. These are addressed at the end of this section.

Clearly, the joint null hypothesis $\alpha = 0$ and $\beta = 1$ entails unbiasedness. From (2.2):

$$E_t(y_{t+1}) = \alpha + \beta E_t(y_{t+1|t}), \tag{2.3}$$

so $E_t(y_{T+1} - y_{T+1|t}) = 0$. However, as noted by Holden and Peel (1990), $\alpha = 0$ and $\beta = 1$ is a sufficient, but not a necessary, condition for unbiasedness, since (2.3) is satisfied more generally by:

$$\alpha = (1 - \beta)E_t(y_{t+1|t}). \tag{2.4}$$

A more satisfactory test of unbiasedness is via a test of $\tau = 0$ in the regression:

$$e_{t+1|t} \equiv y_{t+1} - y_{t+1|t} = \tau + \epsilon_{t+1}, \tag{2.5}$$

that is, from comparing the t-statistic of the null that $\tau = 0$ to the Student's t distribution or the standard normal distribution. The t-statistic is given by:

$$\frac{(1/T)\sum_{t=1}^{T} e_{t+1|t}}{\sqrt{(1/T)s}},$$

where:

$$s^2 = \frac{1}{T-1}\sum_{t=1}^{T}(e_{t+1|t} - \bar{e}_{t+1|t})^2$$

and $\bar{e}_{t+1|t}$ is the sample mean of the forecast errors. A standard textbook result is that the t-statistic has a Student's t distribution ($T-1$ degrees of freedom) when $\{\epsilon_{t+1}\}$ are independent identically distributed (i.i.d.), zero-mean, and come from a normal distribution. Without the assumption of normality the statistic converges in distribution to a standard normal. We will show in Section 2.4 that these standard distributional results for testing for unbiasedness, as well as other aspects of forecast accuracy, may no longer be applicable when the forecasts are derived from models with estimated parameters.

If unbiasedness is not rejected, then this is typically formulated as part of the maintained hypothesis, and various tests of the forecast error being uncorrelated with the past of the process, past errors, or in fact any variables known at t, can be conducted. Tests that include other variables are often termed orthogonality tests. An example of which would be $H_0: \boldsymbol{\gamma} = 0$ in:

$$e_{t+1|t} = \boldsymbol{\gamma}' \mathbf{z}_t + \epsilon_{t+1}, \tag{2.6}$$

where $\mathbf{z}_t$ is the designated vector of variables known at period T.

Although a test of the joint hypothesis $\alpha = 0$ and $\beta = 1$ is often described as a test of unbiasedness, it can also be viewed as a test of efficiency, in the sense of checking that forecasts and their errors are uncorrelated. If there is a systematic relationship between the two, then

this could be exploited to help predict future errors, and could be used to adjust the forecast-generating mechanism accordingly. From (2.2):

$$e_{t+1|t} = y_{t+1} - y_{t+1|t} = \alpha + (\beta - 1)y_{t+1|t} + \epsilon_{t+1}, \tag{2.7}$$

so that the forecast error and forecast are uncorrelated:

$$E(y_{t+1|t}, e_{t+1|t}) = \alpha E(y_{t+1|t}) + (\beta - 1)\, E(y^2_{t+1|t}) + E(y_{t+1|t}\epsilon_{t+1}) = 0$$

when $\alpha = 0$ and $\beta = 1$.

The properties of unbiasedness and efficiency are often presented as minimum requirements for optimal or rational forecasts. However, the identification of the unbiasedness property with optimality requires that the loss function is symmetric, as in the case of quadratic costs (see Zellner (1986), and the discussion in Section 6.8).

2.1.1 Testing the rationality of multi-step forecasts

Consider the case where $h > 1$ in the realization-forecast regression:

$$y_{t+h} = \alpha + \beta y_{t+h|t} + \epsilon_{t+h}. \tag{2.8}$$

When the forecast horizon, h, exceeds the frequency at which forecasts are made (assumed to be 1, with forecasts made at $t, t+1, t+2$, etc.), forecasts will overlap in the sense of being made before the realization paired to the previous forecast is known. Thus, for example, the 2-step ahead forecast error $e_{t+1|t-1}$ will be unknown when the forecast $y_{t+2|t}$ is made. In that case, the efficient use of information does not rule out serial correlation in the error process in (2.8) of order 1 (more generally, moving average of order $h-1$ for h-step ahead forecasts).

This can perhaps be made clearer by supposing for the moment that forecasts are model based. Specifically, the forecasting model is an AR(1):

$$y_t = \phi y_{t-1} + \upsilon_t, \tag{2.9}$$

which happens to coincide with the data generating process. We assume $\{\upsilon_t\}$ is an i.i.d. zero-mean series with $E(\upsilon_t|y_{t-1}, y_{t-2}, \ldots) = 0$, and $|\phi| < 1$. Then letting $y_{t+h|t} = E_t(y_{t+h})$:

$$y_{t+h|t} = \phi^h y_t \tag{2.10}$$

and:

$$e_{t+h|t} = y_{t+h} - y_{t+h|t} = \sum_{i=0}^{h-1} \phi^i \upsilon_{t+h-i}. \tag{2.11}$$

Consider the h-step forecast error made s periods later (substitute t by $t+s$):

$$e_{t+s+h|t+s} = y_{t+s+h} - y_{t+s+h|t+s} = \sum_{i=0}^{h-1} \phi^i \upsilon_{t+s+h-i}. \tag{2.12}$$

The forecast errors will be correlated whenever some same-dated υ's are common to the summations in (2.11) and (2.12). For $s > 0$, the correlation will only be zero when $s > h-1$, which holds for $h = 1$. For $s = 1$ (consecutive forecasts) and $h > 1$ simple algebra gives:

$$E(e_{t+h|t} e_{t+s+h|t+s}) = \frac{\phi(1-\phi^{2(h-1)})}{1-\phi^2}$$

assuming $\sigma_\upsilon^2 \equiv E(\upsilon_t^2) = 1$. When $\alpha = 0$ and $\beta = 1$, the error terms in regression (2.8) are the forecast errors, so the regression errors will exhibit the correlation patterns described above.

While the coefficient estimates obtained from ordinary least squares (OLS) on (2.8) will remain unbiased, the estimate of the covariance matrix of the parameter estimates (necessary for tests of the significance of the parameters in (2.8)) will be inconsistent. This is typically dealt with by using Newey and West (1987) standard errors which correct for autocorrelation and heteroskedasticity, implemented as follows.

Let $\mathbf{x}_{t+h} = (1 \; y_{t+h|t})'$ and $\boldsymbol{\gamma} = (\alpha \; \beta)'$, then we can write (2.8) as:

$$y_{t+h} = \mathbf{x}_{t+h}' \boldsymbol{\gamma} + \epsilon_{t+h}.$$

The OLS estimator of the covariance matrix of $\hat{\boldsymbol{\gamma}}$, $\hat{V}(\hat{\boldsymbol{\gamma}})$, is given by:

$$\hat{V}(\hat{\boldsymbol{\gamma}}) = s^2 \left(\sum_{t=1}^{T} \mathbf{x}_{t+h} \mathbf{x}_{t+h}' \right)^{-1},$$

where s^2 is the usual OLS estimator of the error variance σ^2 of $\{\epsilon_{t+h}\}$. The OLS estimator $\hat{V}(\hat{\boldsymbol{\gamma}})$ assumes that the errors $\{\epsilon_{t+h}\}$ are serially uncorrelated, $E(\epsilon_{t+h}\epsilon_{t+h-s}) = 0$ for all $s \neq 0$. When $h > 1$, we have shown that

$E(\epsilon_{t+h}\,\epsilon_{t+h-s}) = 0$ only for $s > h-1$. The Newey–West covariance matrix, $\hat{V}^*(\hat{\boldsymbol{\gamma}})$, is given by:

$$\hat{V}^*(\hat{\boldsymbol{\gamma}}) = \left(\sum_{t=1}^{T} \mathbf{x}_{t+h}\mathbf{x}'_{t+h}\right)^{-1} T\mathbf{S}^* \left(\sum_{t=1}^{T} \mathbf{x}_{t+h}\mathbf{x}'_{t+h}\right)^{-1}, \tag{2.13}$$

where:

$$\mathbf{S}^* = \frac{1}{T}\sum_{t=1}^{T} \hat{\epsilon}^2_{t+h}\mathbf{x}_{t+h}\mathbf{x}'_{t+h} + \frac{1}{T}\sum_{j=1}^{h-1} w_j \sum_{s=j+1}^{T} \hat{\epsilon}_{s+h}\hat{\epsilon}_{s+h-j}(\mathbf{x}_{s+h}\mathbf{x}'_{s+h-j} + \mathbf{x}_{s+h-j}\mathbf{x}'_{s+h}). \tag{2.14}$$

When $h = 1$ (or equivalently, $w_j = 0$ all j) the second term in (2.14) is zero and the standard errors computed as the square roots of the diagonal elements of $\hat{V}^*(\hat{\boldsymbol{\gamma}})$ are known as heteroskedasticity-consistent standard errors (HCSEs), due to White (1980). The HCSEs allow valid inference when used in place of the square root of the elements of $\hat{V}(\hat{\boldsymbol{\gamma}})$, in the event that the $\{\epsilon_{t+h}\}$ are heteroskedastic but uncorrelated. Note that the type of heteroskedasticity does not need to be specified.

More generally, when $h > 1$ the w_j need to be specified. Setting $w_j = 1$ is perhaps natural, but may lead to estimates of the covariance matrix that are not positive definite. Bartlett weights $w_j = 1 - j/h$ ensure the estimate of the covariance matrix is positive definite and capture the notion that autocorrelations decline with j. The latter assumption will hold when the autocorrelation is induced by the overlapping nature of forecasts. Verbeek (2000), pp. 103–5 provides an intuitive explanation for the form of $\mathbf{S}^*$ given in (2.14). The standard errors from (2.13) with $h > 1$ are known as heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors.

2.2 Forecast precision

Evaluating whether forecasters make efficient use of available information, and whether agents are able to avoid making systematic errors in their predictions, is obviously of interest, not least because these findings may bear more generally on the assumptions of rationality made in the wider economics literature. However, forecasts that are not systematically biased may nevertheless be wildly inaccurate. That is, the variance of forecast errors may be 'large'. But without knowledge of the process generating the data, and the intrinsic difficulty of forecasting

that particular series, it is difficult to judge what a good forecast-error variance is.

Suppose the process were known to be, say, a first-order autoregression (AR(1)):

$$y_t = \phi y_{t-1} + \upsilon_t, \quad |\phi| < 1 \tag{2.15}$$

with $\{\upsilon_t\}$ an i.i.d. zero-mean series, $\sigma_\upsilon^2 = E(\upsilon_t^2)$, and with $E(\upsilon_t \mid y_{t-1}, y_{t-2}, \dots) = 0$, as in Section 2.1.1. The minimum attainable forecast-error variance is given by:

$$V(e_{t+h|t}) \equiv E[(e_{t+h|t} - E(e_{t+h|t}))^2] = E\left[\left(\sum_{i=0}^{h-1} \phi^i \upsilon_{t+h-i}\right)^2\right] = \sigma_\upsilon^2 \frac{(1-\phi^{2h})}{1-\phi^2} \tag{2.16}$$

for an h-step ahead forecast. This could be used as a benchmark against which the forecast-error variances of other models could be judged. For $h = 1$ (2.16) collapses to the variance of the disturbance term in (2.15), $V(e_{t+1|t}) = \sigma_\upsilon^2$.

In the event that the parameters of the model (here ϕ and σ_υ^2) were unknown a benchmark forecast-error variance could still be derived. We can calculate the minimum attainable forecast-error variance using the true model but in ignorance of the values of the parameters. For the model in (2.15), the forecast error for the forecast $\hat{y}_{t+h|t} = \hat{\phi}^h y_t$ is given by:

$$\hat{e}_{t+h|t} \equiv y_{t+h} - \hat{y}_{t+h|t} = (\phi^h - \hat{\phi}^h) y_t + \sum_{i=0}^{h-1} \phi^i \upsilon_{t+h-i}, \tag{2.17}$$

where $\hat{\phi}$ is an estimator of ϕ.[1] Taking the variance of $\hat{e}_{t+h|t}$ conditional on y_t yields:

$$V(\hat{e}_{t+h|t}) \simeq V(\phi^h - \hat{\phi}^h) y_t^2 + \hat{\sigma}_\upsilon^2 \frac{(1-\phi^{2h})}{(1-\phi^2)} \tag{2.18}$$

where in addition $\hat{\sigma}_\upsilon^2$ replaces σ_υ^2. An approximation to $V(\phi^h - \hat{\phi}^h)$ can be obtained as follows.[2] Using the OLS estimator of ϕ we have:

$$\hat{\phi} = \phi + \delta, \tag{2.19}$$

where δ is $O_p(1/\sqrt{T})$, so that powers of δ are asymptotically negligible. $\hat{\phi}^h$ is approximated by the expansion:

$$\hat{\phi}^h = (\phi + \delta)^h \simeq \phi^h + h\delta\phi^{h-1} = \phi^h + h\phi^{h-1}(\hat{\phi} - \phi). \tag{2.20}$$

Therefore:

$$V(\hat{\phi}^h - \phi^h) \simeq V[h\phi^{h-1}(\hat{\phi} - \phi)] = h^2\phi^{2(h-1)}V(\hat{\phi}), \tag{2.21}$$

and substitution into (2.18) results in:

$$V(\hat{e}_{t+h|t}) \simeq h^2\phi^{2(h-1)}V(\hat{\phi})y_t^2 + \hat{\sigma}_v^2\frac{(1-\phi^{2h})}{(1-\phi^2)}. \tag{2.22}$$

The asymptotic variance of the estimated parameters is given by:

$$V(\hat{\phi}) = \sigma^2 E\left(\sum_{t=2}^{T} y_{t-1}^2\right)^{-1} \simeq \frac{1}{T}(1-\phi^2) \tag{2.23}$$

so that plugging this into (2.22) gives the *approximate* forecast-error variance as:

$$V(\hat{e}_{t+h|t}) \simeq \frac{1}{T}h^2\phi^{2(h-1)}(1-\phi^2)y_t^2 + \hat{\sigma}_v^2\frac{(1-\phi^{2h})}{(1-\phi^2)}. \tag{2.24}$$

From (2.24) it is apparent that the effect of parameter estimation uncertainty is of order T^{-1}, and so should be 'small' when T is of a reasonable size.

In practice the specification of the model (e.g. that it is an AR(1) rather than an ARMA(2, 1) say, or an ARMA with extraneous explanatory variables) will not be known, so that the notion of using some measure of the minimum value of the forecast-error variance as a benchmark for assessing forecasts would appear to be unworkable. In the next section we will show that rival models' forecast error variances can be used in the spirit of 'encompassing' to fill in this missing information.

The forecast bias and forecast-error variance combine to give the expected squared forecast error:

$$E(e_{t+h|t}^2) = V(e_{t+h|t}) + [E(e_{t+h|t})]^2. \tag{2.25}$$

When the forecasts are unbiased, choosing the forecast with the smallest forecast-error variance will amount to choosing the forecast with the

smallest expected squared error. Such a strategy suggests that the cost function (which attaches costs or losses to making forecast errors of different magnitudes) is quadratic, so that large errors are proportionately more serious than small, and that over- and under-predictions of the same magnitude have equal costs. Quadratic cost functions are mathematically tractable and underpin OLS, for example. Cost functions and the extent to which squared-error loss is unduly restrictive will be considered in Section 3.3 and in Chapter 6.

Assuming squared-error loss, we will show in Section 2.5.1 that the conditional expectation is the optimal forecast, in that it is the minimum MSE predictor (the MMSEP). In the previous section, the forecast given by (2.10) is the conditional expectation for the model (2.9) and is therefore the MMSEP.

The sample counterpart of (2.25) for the sample of T h-step ahead forecasts is the mean squared forecast error (MSFE):

$$\text{MSFE}_h = \frac{1}{T}\sum_{t=1}^{T} e_{t+h|t}^2 \tag{2.26}$$

while the square root of this quantity is the root MSFE (RMSFE).

2.3 Rival forecasts, forecast combination and encompassing

2.3.1 Tests of comparative forecast accuracy

Assuming that the loss function is squared-error loss, the corresponding sample measure of forecast accuracy, (R)MSFE, can be calculated for each set of forecasts, and the set with the smallest MSFE might be deemed the most accurate. Of interest though is whether differences between rival forecasts can be attributed to sampling variability, or whether any apparent differences are statistically significant once this variability has been taken into account.

The two sets of forecasts will be distinguished by 'hats' and 'tildes', that is, as $\hat{y}_{t+h|t}$ and $\tilde{y}_{t+h|t}$. The corresponding forecast errors are collected in the column vectors $\hat{\mathbf{e}}$ and $\tilde{\mathbf{e}}$, where $\hat{\mathbf{e}} = (\hat{e}_{1+h|1}, \ldots, \hat{e}_{T+h|T})'$ and $\tilde{\mathbf{e}} = (\tilde{e}_{1+h|1}, \ldots, \tilde{e}_{T+h|T})'$. If we assume that the forecast errors are zero-mean, normally distributed and serially uncorrelated (implying $h = 1$) then the following test due to Granger and Newbold (1977) (and sometimes known as the Morgan–Granger–Newbold test in recognition of

Morgan 1940) is the uniformly most powerful unbiased. Under unbiasedness, equality of MSFE of $\hat{\mathbf{e}}$ and $\tilde{\mathbf{e}}$ amounts to equality of variances. The test of equal variances can be implemented by the use of an orthogonalizing transformation to construct $u_{1,t+1|t} = \hat{e}_{t+1|t} - \tilde{e}_{t+1|t}$ and $u_{2,t+1|t} = \hat{e}_{t+1|t} + \tilde{e}_{t+1|t}$, and then test for zero correlation between $u_{1,t+1|t}$ and $u_{2,t+1|t}$.

Note that:

$$E(u_{1,t+1|t}\, u_{2,t+1|t}) = E(\hat{e}^2_{t+1|t}) - E(\tilde{e}^2_{t+1|t})$$

so:

$$E(u_{1,t+1|t} u_{2,t+1|t}) = 0 \Longrightarrow E(\hat{e}^2_{t+1|t}) = E(\tilde{e}^2_{t+1|t}).$$

The test statistic is:

$$\frac{r}{\sqrt{(T-1)^{-1}(1-r^2)}} \sim t_{T-1}, \tag{2.27}$$

where:

$$r = \frac{\mathbf{u}_1' \mathbf{u}_2}{\sqrt{\mathbf{u}_1' \mathbf{u}_1 \mathbf{u}_2' \mathbf{u}_2}}$$

and $\mathbf{u}_i' = (u_{i,2|1}, \ldots, u_{i,T+1|T})$, $i = 1, 2$.

Given the restrictive nature of the assumptions that underpin this test, and especially that it is only applicable for $h = 1$, more general approaches are often required. Diebold and Mariano (1995) introduce a test statistic that does not require zero-mean, normally distributed and serially uncorrelated forecast errors, so that it is applicable when $h > 1$. Nor does the test statistic require that the loss function is squared-error loss. So assume an arbitrary loss function $g(x)$, where $g(x) = x^2$ for squared-error loss, for example, and x is either $\hat{e}_{t+h|t}$ or $\tilde{e}_{t+h|t}$. Next, define the loss differential as $d_{t+h|t} \equiv [g(\hat{e}_{t+h|t}) - g(\tilde{e}_{t+h|t})]$, so that equal forecast accuracy entails the condition that $E(d_{t+h|t}) = 0$. Given a covariance-stationary sample realization $\{d_{t+h|t}\}$, the asymptotic distribution of the sample mean loss differential $\bar{d}$:

$$\bar{d} = \frac{1}{T} \sum_{t=1}^{T} d_{t+h|t}$$

is given by:

$$\sqrt{T}(\bar{d} - \mu) \xrightarrow{D} N(0, 2\pi f_d(0)),$$

where:

$$f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau),$$

is the spectral density of the loss differential at frequency zero, γ_d is the autocovariance function, and $\xrightarrow{D}$ denotes convergence in distribution. The large-sample statistic that Diebold and Mariano (1995) propose for testing the null of equal forecast accuracy is:

$$\frac{\bar{d}}{\sqrt{\frac{1}{T} 2\pi \hat{f}_d(0)}} \overset{app}{\sim} N(0, 1),$$

where $\hat{f}_d(0)$ is a consistent estimate of $f_d(0)$, based on a weighted sum of the sample autocovariances. The notation '$\overset{app}{\sim}$' indicates that this statistic approximately follows a standard normal distribution.

It may be more natural to some readers to think of the denominator of the test statistic explicitly as the square root of the estimated variance of $\bar{d}$:

$$\frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \overset{app}{\sim} N(0, 1). \tag{2.28}$$

As an example, consider only $h = 1$ forecasts, so that $\{\hat{e}_{t+1|t}\}$ or $\{\tilde{e}_{t+1|t}\}$ are serially uncorrelated, and treating $\{d_{t+1|t}\}$ as being serially uncorrelated, we obtain:

$$V(\bar{d}) = V\left(\frac{1}{T}\sum_{t=1}^{T} d_{t+1|t}\right) = \frac{1}{T^2}[TV(d_{t+1|t})] = \frac{1}{T}V(d_{t+1|t}) \tag{2.29}$$

because all the covariance terms of the form $C(d_{t+1|t}d_{t+1-i|t-i})$ are zero, $i \neq 0$. We can estimate $V(d_{t+1|t})$ by $\hat{V}(d_{t+1|t})$, the sample moment:

$$\hat{V}(d_{t+1|t}) = \frac{1}{T}\sum_{t=1}^{T} \left(d_{t+1|t} - \bar{d}\right)^2$$

and plugging this into (2.29) gives an estimate $\hat{V}(\bar{d})$ of $V(\bar{d})$, and the test statistic becomes

$$\frac{T\bar{d}}{\sqrt{\sum_{t=1}^{T}(d_{t+1|t}-\bar{d})^2}} \overset{app}{\sim} N(0,1).$$

For $h > 1$, forecast errors will be serially correlated of order $h-1$, and the estimate of $V(\bar{d})$ will need to include sample autocovariances:

$$\hat{\gamma}_i = \frac{1}{T}\sum_{t=i+1}^{T}\left(d_{t+h|t}-\bar{d}\right)\left(d_{t+h-i|t-i}-\bar{d}\right)$$

for $i < h$.

2.3.2 Forecast combination (or pooling) and encompassing

The basic idea behind the combination (or 'pooling') of forecasts is that, although one forecast may be superior to another (on MSFE, say) using the tests of equal forecast accuracy discussed in the previous section, a combined forecast of the two may be better still. There is an extensive literature: see *inter alia* Diebold and Lopez (1996) and Newbold and Harvey (2002) for recent surveys, and Clemen (1989) for an annotated bibliography. Studies such as Newbold and Granger (1974) provided early evidence of the efficacy of combining, and Stock and Watson (1999a) and Fildes and Ord (2002), for example, suggest that simple forms of combination (that do not take into account the relative past performances of the forecasts) often work just as well as more elaborate schemes.

Following Nelson (1972) and Granger and Newbold (1973), a forecast is said to be conditionally efficient if the variance of the forecast error from a combination of that forecast and a rival forecast is not significantly less than that of the original forecast alone. Chong and Hendry (1986) apply the principle of encompassing to the evaluation of forecasts. This principle holds that an empirical model should be able to account for the findings of rival models (see Hendry and Richard (1982, 1989), Mizon (1984) and Mizon and Richard (1986)). The reasoning is as follows. If the investigator knew the actual mechanism that gave rise to the observed data series, then the properties of a particular (mis-specified) model could be deduced analytically: we could work out the forecast-error variance for that model. Of course the data generating process is not known, but the above thought-experiment indicates that if a model closely approximates the data generation process (DGP), it should be possible to deduce

the properties of other models. One implication is that no other model should have a smaller forecast-error variance. A model is said to forecast encompass a rival if the rival's forecasts have no additional explanatory power, in the sense of contributing to a lower MSFE or forecast-error variance when used in combination with the original set of forecasts. Testing for forecast encompassing and the procedure for computing conditional efficiency are formally equivalent.

Suppose we have two sets of forecasts, $\hat{y}_{t+h|t}$ and $\tilde{y}_{t+h|t}$. We shall consider $h = 1$, and so compress the notation that explicitly states the origin and horizon, so that 1-step forecasts of period $t+1$ are simply $\hat{y}_{t+1}$ and $\tilde{y}_{t+1}$. A further simplification can be achieved by defining $f_{1t} \equiv \hat{y}_t$ and $f_{2t} \equiv \tilde{y}_t$. The forecast combination is given by:

$$f_{ct} = (1 - \lambda) f_{1t} + \lambda f_{2t}. \tag{2.30}$$

As written, the weights on the individual forecasts sum to unity, and there is no intercept. This is appropriate if the individual forecasts are unbiased. We might also impose the requirement that $0 \leq \lambda \leq 1$.

Given a squared-error loss function, the weight λ is chosen to minimize the MSFE of the combined predictor, f_{ct}. Assuming unbiasedness, this is equivalent to minimizing the forecast-error variance. Subtracting y_t from both sides of (2.30) and multiplying through by -1 results in:

$$e_{ct} = (1 - \lambda) e_{1t} + \lambda e_{2t}, \tag{2.31}$$

where $e_{ct} \equiv y_t - f_{ct}$ and $e_{it} \equiv y_t - f_{it}$, $i = 1, 2$. The variance of the combined forecast error is:

$$V(e_{ct}) = (1 - \lambda)^2 V(e_{1t}) + \lambda^2 V(e_{2t}) + 2\lambda(1 - \lambda) C(e_{1t}, e_{2t}). \tag{2.32}$$

Choosing λ to minimize $V(e_{c,t})$ leads to:

$$\lambda^* = \frac{V(e_{1t}) - C(e_{1t}, e_{2t})}{V(e_{1t}) + V(e_{2t}) - 2C(e_{1t}, e_{2t})}. \tag{2.33}$$

Substituting (2.33) into (2.32) we can obtain the variance using λ^* as the combination weight:

$$V[e_{ct}; \lambda^*] = \frac{(1 - \rho^2) V_1 V_2}{V_1 + V_2 - 2\rho\sqrt{V_1 V_2}},$$

where $V_i = V[e_{it}]$, $i = 1, 2$, $\rho = C(e_{1t}, e_{2t})/\sqrt{V_1 V_2}$. Using the optimal weight λ^* leads to the inequality:

$$\text{MSFE}(f_{ct}) \leq \min\{\text{MSFE}(f_{1t}), \text{MSFE}(f_{2t})\}, \tag{2.34}$$

so that combination must be at least as good as the best individual forecasts. In the unlikely event that the forecasts are uncorrelated, $C(e_{1t}, e_{2t}) = 0$, and (2.33) is simply:

$$\lambda^* = \frac{V(e_{1t})}{V(e_{1t}) + V(e_{2t})}, \tag{2.35}$$

which has the natural interpretation that the weights only depend (inversely) on the sizes of the relative forecast-error variances, and that the larger $V(e_{1t})$ the smaller the weight $(1 - \lambda)$ of f_{1t}.

In practice, the weights can be calculated by replacing the population second-moment matrices in (2.33) by their sample counterparts:

$$\begin{aligned}\hat{\lambda} &= \frac{(1/T)\sum_{t=1}^{T} e_{1t}^2 - (1/T)\sum_{t=1}^{T} e_{1t}e_{2t}}{(1/T)\sum_{t=1}^{T} e_{1t}^2 + (1/T)\sum_{t=1}^{T} e_{2t}^2 - 2(1/T)\sum_{t=1}^{T} e_{1t}e_{2t}} \\ &= \frac{\sum_{t=1}^{T}(e_{1t} - e_{2t})e_{1t}}{\sum_{t=1}^{T}(e_{1t} - e_{2t})^2}.\end{aligned} \tag{2.36}$$

The second line is the OLS estimator of λ in equation (2.31), noting that equation (2.31) can be rearranged to give:

$$e_{1t} = \lambda(e_{1t} - e_{2t}) + e_{ct}. \tag{2.37}$$

Therefore, the optimal weight can be obtained by a simple OLS regression.[3] That being the case, it follows immediately that the hypothesis that f_{1t} forecast encompass f_{2t} (or f_{1t} is conditionally efficient) is simply the t-test of the null that $\lambda = 0$ in equation (2.37). This can be viewed as a one-sided test against the alternative that $\lambda > 0$, that is, that f_{2t} has a positive weight in the combination. This is more intuitive than a two-sided test with the alternative hypothesis that $\lambda \neq 0$, but see

Clements and Hendry (2004). Clearly, from (2.37) $\lambda = 0$ requires that $E(e_{1t}, e_{1t} - e_{2t}) = 0$. Then the forecast f_{2t} contains no useful information that is not already present in f_{1t}.

Harvey *et al*. (1998) provide some small-sample evidence on the size and power properties of the t-test that $\lambda = 0$ when the forecast errors are 'well-behaved', in the sense that they are normally distributed, and also when they are non-normal. Table 2.1 reports the results of a Monte Carlo[4] study similar to that undertaken by Harvey *et al*. (1998). We give the Monte Carlo estimates of the sizes of a number of tests of forecast encompassing, for two data generating processes, and for $T = \{8, 16, 32, 64, 128\}$. The first column of figures in the table are the sizes of tests when the forecast errors are normal. Samples of size T are generated for $\{e_{1t}, e_{2t}\}$ from:

$$
\begin{aligned}
e_{1t} &= \varepsilon_{1t}, \\
e_{2t} &= \varepsilon_{1t} + 0.5\varepsilon_{2t},
\end{aligned}
$$

where ε_{1t} and ε_{2t} are i.i.d. standard normal variables from a pseudo-random number generator. Note that:

$$E(e_{1t}, e_{1t} - e_{2t}) = E[\varepsilon_{1t}(\varepsilon_{1t} - (\varepsilon_{1t} + 0.5\varepsilon_{2t}))] = 0$$

because $E(\varepsilon_{1t}\varepsilon_{2t}) = 0$ by construction, so the simulated sample of forecast errors satisfy the relationship that forecast 1 encompasses forecast 2. Letting t_i be the value of a test statistic for the null that 1 forecast encompasses 2, calculated on the ith simulated sample of size T, then the size estimates reported in the table are calculated as:

$$100 \times \frac{1}{R} \sum_{i=1}^{R} 1(t_i > c_{0.05}),$$

where $1(\cdot)$ is the indicator function, equal to 1 when $t_i > c_{0.05}$ and 0 when $t_i \leq c_{0.05}$, and $c_{0.05}$ is the 5% one-sided (Student t or normal) critical value for $\lambda = 0$ versus $\lambda > 0$. R, the number of replications, is set to 40,000.

We calculate the following test statistics:

'Standard' The standard t-statistic for $\lambda = 0$ is compared to the standard normal.

R_1 The t-statistic calculated using a White-HCSE (from the Newey–West covariance matrix in (2.14) with $h = 1$), and compared to a Student t_{T-1} reference distribution.

Table 2.1 Monte Carlo estimates of sizes of tests of forecast encompassing

Test statistic	Normal errors	Student t errors
$T = 8$		
Standard	4.9	8.4
R_1	9.9	12.8
DM	8.1	7.2
MDM	4.2	3.2
SR_1	6.5	7.3
SR_2	5.5	6.9
$T = 16$		
Standard	4.9	9.8
R_1	7.6	11.0
DM	6.5	6.0
MDM	4.7	4.1
SR_1	5.0	5.9
SR_2	5.0	6.2
$T = 32$		
Standard	5.1	10.4
R_1	6.7	8.9
DM	5.9	5.3
MDM	5.0	4.3
SR_1	5.0	6.0
SR_2	5.1	6.1
$T = 64$		
Standard	5.1	11.5
R_1	6.1	7.8
DM	5.7	5.1
MDM	5.2	4.6
SR_1	4.9	6.0
SR_2	5.0	6.5
$T = 128$		
Standard	5.2	12.3
R_1	5.6	6.8
DM	5.4	5.0
MDM	5.2	4.8
SR_1	5.2	5.9
SR_2	5.2	6.2

Notes: The table gives the percentage rejection rates for tests of forecast encompassing. Rejection rates estimated from 40,000 replications. Tests are carried out at a 5% nominal level.

DM The Diebold–Mariano test for equal forecast accuracy applied to testing for forecast encompassing. Recall that in Section 2.3.1 the test for equal forecast accuracy (assuming squared-error loss) was based on testing whether $d_t = e_{1t}^2 - e_{2t}^2 = 0$. If instead we define $d_t = e_{1t}(e_{1t} - e_{2t})$, testing whether $d_t = 0$ is now a test of forecast encompassing. The resulting test statistic is compared to the standard normal distribution.

MDM Harvey *et al.* (1997) propose modifications to DM aimed at improving its small-sample performance. For $h = 1$, as here:

$$\text{MDM} = \sqrt{1 + \frac{1}{T}} \times DM$$

and MDM is compared to a Student t_{T-1} distribution.

SR Spearman's rank correlation test. This is a distribution free test that determines whether there is a monotonic relation between two variables, here e_{1t} and $(e_{1t} - e_{2t})$. It is applicable when, as here with $h = 1$, it is reasonable to assume that drawings of $\{e_{1t}, (e_{1t} - e_{2t})\}$ are independent. SR_1 is the one-sided rank correlation test against the alternative of positive correlation, and SR_2 is a two-sided test. All the other tests are one-sided tests, against the alternative of positive correlation, $\lambda > 0$.

The first column of the table indicates that R_1 and DM are over-sized for $n = 8$ and 16. Clearly, HCSEs (R_1) are not necessary, and the modification to DM (MDM) improves the performance of this statistic. Harvey *et al.* (1998) argue that forecast-error distributions are liable to be heavy-tailed if very large absolute errors are occasionally observed. They show analytically that in this case the standard t-test will be over-sized, and they analyse by Monte Carlo the usefulness of HCSEs and of the other tests of forecast encompassing described above. The second column of figures in the table records size estimates for heavy-tailed forecast errors. Following Harvey *et al.* (1998), forecast errors are generated from:

$$e_{it} = \frac{u_{it}}{\sqrt{(\chi^2_{\nu,t}/\nu)}},$$

where:

$$u_{1t} = \varepsilon_{1t}$$
$$u_{2t} = \varepsilon_{1t} + 0.5\varepsilon_{2t}$$

and ε_{1t}, ε_{2t} are i.i.d. $N(0, 1)$ variables, as before, and $\chi^2_{\nu,t}$ is a chi-squared random variable with $\nu = 5$ degrees of freedom. The results indicate that the standard tests becomes increasingly over-sized as T increases; R_1 is correctly sized for large samples, but exacerbates the problem for small samples, whilst the (M)DM and SR tests are reasonable throughout. Harvey *et al*. (1998) also consider the power properties of these statistics.

2.4 Testing model-based forecasts for predictive accuracy

In this section, we consider issues that arise in the evaluation of point forecasts that are explicitly model-based. If a set of forecasts are based on a model, then an approach to forecast evaluation presents itself that does not require the existence of rival forecasts (as in Section 2.3), namely, comparing the accuracy of the forecasts to what would have been expected based on the past fit of the model to the data. This idea underlies tests of predictive accuracy which compare an estimate of the forecast-error variance obtained from the past residuals with the actual mean-squared error of the forecast (see, *inter alia*, Chow (1960), Christ (1966) and Hendry (1974, 1979) for early developments). Such tests are briefly reviewed in Section 2.4.1. However, published forecasts of macro-economic variables based on large-scale macro-econometric models usually reflect in varying degree the properties of the model and the skills of the models' proprietors. Forecasts are rarely based on the model alone. Moreover, forecasters' adjustments tend to improve accuracy: see, for example, Marris (1954), Wallis *et al*. (1986, Table 4.8), Wallis *et al*. (1987, Figures 4.3 and 4.4) and Wallis and Whitley (1991). That being the case, tests of the predictive accuracy of the model-based forecasts may have little bearing on assessing the published forecasts. The usefulness of this approach may also be limited by the models on which the forecasts are based being unknown or not available for the purpose of evaluation.

A second aspect we consider in this section is the impact on tests of equal accuracy, and tests of forecast encompassing, of the forecasts and rival forecasts being model-based.

2.4.1 Tests of predictive accuracy

To begin with, suppose:

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \tag{2.38}$$

where $1, \ldots, T$ is now the estimation sample period and $T + 1$ is the period to be forecast. $\mathbf{x}_t$ is a p-dimensional vector of explanatory variables at period t, and let $\mathbf{X}$ be the $T \times p$ matrix of observations on the p explanatory variables for periods $1, \ldots, T$. We assume that the $\{\mathbf{x}_t\}$ are strongly exogenous stochastic regressors, ruling out lagged $\{y\}$'s. We also assume that the $\{\varepsilon_t\}$ are i.i.d. $N(0, \sigma_\varepsilon^2)$. The OLS estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is given by:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon},\end{aligned}$$

where $\boldsymbol{\varepsilon}' = [\varepsilon_1 \cdots \varepsilon_T]$. The 1-step error in forecasting the regressand at period $T + 1$ is:

$$\begin{aligned}e_{T+1|T} &= y_{T+1} - \mathbf{x}'_{T+1}\hat{\boldsymbol{\beta}} \\ &= \mathbf{x}'_{T+1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon_{T+1} \\ &= -\mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} + \varepsilon_{T+1}.\end{aligned}$$

Since $E(e_{T+1|T}) = 0$, the variance of $e_{T+1|T}$, $V(e_{T+1|T}) = E(e_{T+1|T}^2)$, and so:

$$\begin{aligned}V(e_{T+1|T}) &= \mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_{T+1} + \sigma_\varepsilon^2 \\ &= \sigma_\varepsilon^2\mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_{T+1} + \sigma_\varepsilon^2 \\ &\equiv \sigma_\varepsilon^2 f_{T+1}\end{aligned}$$

assuming that the process at time $T+1$ is the same as that at $1, \ldots, T$ (i.e. $y_{T+1} = x'_{T+1}\beta + \varepsilon_{T+1}$, and $E(\varepsilon_{T+1}^2) = \sigma_\varepsilon^2$). Because $e_{T+1|T}$ is a linear combination of normally distributed random variables, $e_{T+1|T} \sim N(0, \sigma_\varepsilon^2 f_{T+1})$, and hence:

$$\frac{e_{T+1|T}^2}{\sigma_\varepsilon^2 f_{T+1}} \sim \chi_{(1)}^2.$$

In practice σ_ε^2 will not be known. A standard textbook result is that:

$$\frac{(T-p)s^2}{\sigma_\varepsilon^2} \sim \chi_{(T-p)}^2.$$

Because the two chi-squared random variables are independent, dividing each by their degrees of freedom, and then the resulting statistics one by

the other, gives the Chow (1960) test statistic we referred to above:

$$Q = \frac{e_{T+1|T}^2}{s^2 f_{T+1}} \sim F_{T-p}^1 \tag{2.39}$$

a statistic with an F-distribution with $(1, T-p)$ degrees of freedom. When the explanatory variables contain lags of the dependent variable the justification for comparing Q to the F-distribution rests on asymptotic arguments, but Kiviet (1986) shows by simulation that it has good size properties, and compares favourably with other asymptotically equivalent statistics.

We can also calculate tests of predictive accuracy for h-step forecasts. To do so, we will explicitly consider forecasts for $\{y_t\}$ based on a time-series model, in contrast to (2.38). We consider the Box and Tiao (1976) test, proposed as a way of testing for parameter change at a particular point. $\{y_t\}$ is a scalar process given by the Wold representation:

$$y_t = \psi(L)\epsilon_t, \tag{2.40}$$

where $\epsilon_t \sim$ i.i.d. $N(0, \sigma_\epsilon^2)$, $\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + \cdots$, and $\psi_0 = 1$. Assuming $\psi(L)$ is invertible gives the AR representation:

$$\phi(L)y_t = \epsilon_t, \tag{2.41}$$

where $\phi(L) = \psi(L)^{-1}$. The actual value of the process at $T+h$ using (2.40) can be split into the sum of two sets of disturbances: those relating to the (present and) past relative to T, the date at which the forecast is made, and those relating to the future, that is:

$$y_{T+h} = \sum_{j=h}^{\infty} \psi_j \epsilon_{T+h-j} + \sum_{j=0}^{h-1} \psi_j \epsilon_{T+h-j}.$$

The MMSEP is:

$$y_{T+h|T} = \sum_{j=h}^{\infty} \psi_j \epsilon_{T+h-j},$$

because $E_T(\epsilon_s) = \epsilon_s$ for $s \leq T$, but $E_T(\epsilon_s) = 0$ for $s > T$. Thus when the parameters $\{\psi_j\}$ are known:

$$e_{T+h|T} = y_{T+h} - y_{T+h|T} = \sum_{j=0}^{h-1} \psi_j \epsilon_{T+h-j}. \tag{2.42}$$

The multi-period errors in forecasting $(y_{T+1}, \ldots, y_{T+h})$ conditional on period T are stacked in the vector $\mathbf{e}_h = (e_{T+1|T}, \ldots, e_{T+h|T})$, which can be related to the disturbances $\{\epsilon_t\}$ by:

$$\mathbf{e}_h = \boldsymbol{\psi} \boldsymbol{\epsilon}_h, \tag{2.43}$$

where $\boldsymbol{\epsilon}_h = (\epsilon_{T+1}, \ldots, \epsilon_{T+h})'$, and:

$$\boldsymbol{\psi} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \psi_1 & 1 & 0 & \cdots & 0 & 0 \\ \psi_2 & \psi_1 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ \psi_{h-2} & \psi_{h-3} & \cdots & \psi_1 & 1 & 0 \\ \psi_{h-1} & \psi_{h-2} & \cdots & \psi_2 & \psi_1 & 1 \end{pmatrix}.$$

Because $\boldsymbol{\psi}$ is non-singular, we can invert (2.43):

$$\boldsymbol{\epsilon}_h = \boldsymbol{\phi} \mathbf{e}_h, \tag{2.44}$$

where $\boldsymbol{\phi} = \boldsymbol{\psi}^{-1}$, so that (2.43) and (2.44) give the relationship between the forecast errors and the vector of disturbance terms.

The covariance matrix of the forecast errors is:

$$\boldsymbol{\Phi}_h = E(\mathbf{e}_h \mathbf{e}_h') = \boldsymbol{\psi} E(\boldsymbol{\epsilon}_h \boldsymbol{\epsilon}_h') \boldsymbol{\psi}' = \sigma_\epsilon^2 \boldsymbol{\psi} \boldsymbol{\psi}',$$

because $E(\boldsymbol{\epsilon}_h \boldsymbol{\epsilon}_h') = \sigma_\epsilon^2 \mathbf{I}_h$. Thus, if the model appropriate over $(1, \ldots, T)$ remains so over the forecast horizon $(T+1, \ldots, T+h)$, then:

$$Q = \mathbf{e}_h' \boldsymbol{\Phi}_h^{-1} \mathbf{e}_h \sim \chi^2_{(h)}. \tag{2.45}$$

Under the null, $\mathbf{e}_h \sim N_h(\mathbf{0}, \boldsymbol{\Phi}_h)$, so the distribution of Q follows from that of a quadratic form in normally distributed variables (see, e.g., Lütkepohl (1991, Proposition B.3, p. 481)).

Again following Box and Tiao, it is straightforward to show that we can replace the multi-step forecast errors e_h in (2.45) by 1-step ahead forecast errors made over the period $(e_{T+1|T}, \ldots, e_{T+h|T+h-1})$, noting that the latter are simply $(\epsilon_{T+1}, \ldots, \epsilon_{T+h})$, because $\boldsymbol{\Phi}_h^{-1} = \sigma_\epsilon^{-2}\boldsymbol{\psi}^{-1\prime}\boldsymbol{\psi}^{-1} = \sigma_\epsilon^{-2}\boldsymbol{\phi}'\boldsymbol{\phi}$. Thus:

$$Q = \mathbf{e}_h'\boldsymbol{\Phi}_h^{-1}\mathbf{e}_h = \frac{\mathbf{e}_h'\boldsymbol{\phi}'\boldsymbol{\phi}\mathbf{e}_h}{\sigma_\epsilon^2} = \frac{\boldsymbol{\epsilon}_h'\boldsymbol{\epsilon}_h}{\sigma_\epsilon^2} = \sigma_\epsilon^{-2}\sum_{j=1}^{h}\epsilon_{T+j}^2. \tag{2.46}$$

Since in practice we will require an estimate of σ_ϵ^2 to calculate Q, Box and Tiao suggest an approximate F-variant defined by:

$$\hat{Q} = \frac{\sum_{j=1}^{h}\epsilon_{T+j}^2}{hs^2} \sim F_{T-p}^h, \tag{2.47}$$

where p denotes the number of parameters estimated in the model when obtaining s^2. This test statistic is formed as the ratio of two chi-squared statistics exactly as for (2.39), where the h in the denominator appears as it is the degrees of freedom of the numerator statistic (given by (2.46)).

2.4.2 Tests of equal accuracy and encompassing when parameters are estimated

The asymptotic distributions of tests of predictive accuracy may be affected by parameter estimation error: see West (1996), West and McCracken (1998) and McCracken (2000). We will illustrate the general argument with the example of testing for unbiasedness taken from West and McCracken (2002, Section 14.3.1). For some tests of predictive ability, such as the Diebold–Mariano test of equal mean squared errors between two sets of forecasts (discussed in Section 2.3.1), the limiting distribution of the test statistic remains standard normal even in the presence of parameter estimation uncertainty. West and McCracken (2002) provide a readable account, setting out a general framework for inference about predictive ability, the conditions under which parameter estimation can be ignored (at least asymptotically), and ways of correcting test statistics.

For testing unbiasedness, West and McCracken (2002) consider the following set up. The model is:

$$y_t = \mathbf{X}_{t-1}'\boldsymbol{\beta}^* + \varepsilon_t,$$

where $\mathbf{X}'_{t-1} = (1, x_{t-1})$, and ε_t is i.i.d. $(0, \sigma^2)$. $\hat{\boldsymbol{\beta}} = (\sum_{s=1}^{R} \mathbf{X}_{s-1}\mathbf{X}'_{s-1})^{-1} \times (\sum_{s=1}^{R} \mathbf{X}_{s-1} y_s)$ is the OLS estimator of $\boldsymbol{\beta}^*$ on the sample $1, \ldots, R$. 1-step forecasts are given by $y_{t+1|t} = \mathbf{X}'_t \hat{\boldsymbol{\beta}}$ for $t = R, \ldots, R + P - 1$, with forecast errors $e_{t+1|t} = y_{t+1} - y_{t+1|t}$.

The test for unbiasedness is the t-statistic for the null that $\alpha = 0$ in the regression:

$$e_{t+1|t} = \alpha + \upsilon_{t+1}$$

for the P forecast errors. The standard t-statistic is:

$$\frac{P^{-1} \sum_{t=R}^{R+P-1} e_{t+1|t}}{\sqrt{s^2 \left(\sum_{t=R}^{R+P-1} 1^2 \right)^{-1}}} = \frac{P^{-1/2} \sum_{t=R}^{R+P-1} e_{t+1|t}}{\sqrt{\frac{1}{P-1} \sum_{t=R}^{R+P-1} (e_{t+1|t} - \bar{e})^2}} \tag{2.48}$$

from substituting $s^2 = (P-1)^{-1} \sum_{t=R}^{R+P-1} (e_{t+1|t} - \bar{e})^2$, and where $\bar{e}$ is the sample mean of the forecast errors. This statistic will have a limiting standard normal distribution if the numerator is asymptotically normal with limiting variance Ω, say, and the denominator converges in probability to $\Omega^{1/2}$.

We consider the numerator, noting that we can write $e_{t+1|t}$ as:

$$\begin{aligned} e_{t+1|t} &= y_{t+1} - y_{t+1|t} \\ &= \varepsilon_{t+1} + \mathbf{X}'_t(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) \\ &= \varepsilon_{t+1} - \mathbf{X}'_t \left(\sum_{s=1}^{R} \mathbf{X}_{s-1}\mathbf{X}'_{s-1} \right)^{-1} \left(\sum_{s=1}^{R} \mathbf{X}_{s-1}\varepsilon_s \right), \end{aligned} \tag{2.49}$$

where the third line follows from $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = (\sum_{s=1}^{R} \mathbf{X}_{s-1}\mathbf{X}'_{s-1})^{-1} (\sum_{s=1}^{R} \mathbf{X}_{s-1}\varepsilon_s)$. Substituting into the numerator of (2.48) gives:

$$\begin{aligned} &P^{-1/2} \sum_{t=R}^{R+P-1} e_{t+1|t} \\ &\quad = P^{-1/2} \sum_{t=R}^{R+P-1} \left(\varepsilon_{t+1} - \mathbf{X}'_t \left(\sum_{s=1}^{R} \mathbf{X}_{s-1}\mathbf{X}'_{s-1} \right)^{-1} \left(\sum_{s=1}^{R} \mathbf{X}_{s-1}\varepsilon_s \right) \right) \end{aligned}$$

$$
\begin{aligned}
&= P^{-1/2} \sum_{t=R}^{R+P-1} \varepsilon_{t+1} - P^{-1/2} \sum_{t=R}^{R+P-1} \mathbf{X}_t' \left(\sum_{s=1}^{R} \mathbf{X}_{s-1} \mathbf{X}_{s-1}' \right)^{-1} \left(\sum_{s=1}^{R} \mathbf{X}_{s-1} \varepsilon_s \right) \\
&= P^{-1/2} \sum_{t=R}^{R+P-1} \varepsilon_{t+1} - P^{-1} \sum_{t=R}^{R+P-1} \mathbf{X}_t' \left(R^{-1} \sum_{s=1}^{R} \mathbf{X}_{s-1} \mathbf{X}_{s-1}' \right)^{-1} \\
&\quad \times \left(\left(\frac{P}{R} \right)^{1/2} R^{-1/2} \sum_{s=1}^{R} \mathbf{X}_{s-1} \varepsilon_s \right).
\end{aligned}
$$

The third line scales the sums of random variables.

An assumption then has to be made about the relative rates at which R, the number of in-sample observations, and P, the number of out-of-sample predictions, get large. We assume that $P, R \to \infty$, and $P/R \to \pi < \infty$. In that case:

$$
\begin{aligned}
P^{-1/2} \sum_{t=R}^{R+P-1} e_{t+1|t} &= P^{-1/2} \sum_{t=R}^{R+P-1} \varepsilon_{t+1} - (E\mathbf{X}_t')(E\mathbf{X}_t \mathbf{X}_t')^{-1} \\
&\quad \times \left(\pi^{1/2} R^{-1/2} \sum_{s=1}^{R} \mathbf{X}_{s-1} \varepsilon_s \right) + o_p(1) \\
&= \left(1 : - \left(E\mathbf{X}_t' \right) \left(E\mathbf{X}_t \mathbf{X}_t' \right)^{-1} \right) \\
&\quad \times \begin{pmatrix} P^{-1/2} \sum_{t=R}^{R+P-1} \varepsilon_{t+1} \\ \pi^{1/2} R^{-1/2} \sum_{s=1}^{R} \mathbf{X}_{s-1} \varepsilon_s \end{pmatrix} + o_p(1)
\end{aligned}
$$

under general conditions.

Given the assumptions about the $\{\varepsilon\}$, the two components of the column vector are independent, and so:

$$
\begin{pmatrix} P^{-1/2} \sum_{t=R}^{R+P-1} \varepsilon_{t+1} \\ \pi^{1/2} R^{-1/2} \sum_{s=1}^{R} \mathbf{X}_{s-1} \varepsilon_s \end{pmatrix} \xrightarrow{d} N \left(0_{3\times 1}, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & \pi(E\mathbf{X}_t \mathbf{X}_t') \end{pmatrix} \right).
$$

Thus the numerator in (2.48) is a linear combination of two independent zero-mean normally distributed random variables, and so it is itself a

zero-mean normal random variable:

$$P^{-1/2} \sum_{t=R}^{R+P-1} e_{t+1|t} \xrightarrow{d} N(0, \Omega),$$

where:

$$\begin{aligned}\Omega &= \sigma^2 + \sigma^2 \pi (E\mathbf{X}_t')(E\mathbf{X}_t\mathbf{X}_t')^{-1}(E\mathbf{X}_t\mathbf{X}_t')(E\mathbf{X}_t\mathbf{X}_t')^{-1}(E\mathbf{X}_t') \\ &= \sigma^2(1 + \pi(E\mathbf{X}_t')(E\mathbf{X}_t\mathbf{X}_t')^{-1}(E\mathbf{X}_t')).\end{aligned}$$

One can show that the denominator of (2.48) converges in probability to σ^2 (see West and McCracken (2002, pp. 306–307) for details), so that the asymptotic distribution of the t-statistic is $N(0, V)$, where:

$$V = \frac{\Omega}{\sigma^2} = 1 + \pi(E\mathbf{X}_t')(E\mathbf{X}_t\mathbf{X}_t')^{-1}(EX_t').$$

Because $V > 1$, the asymptotic distribution of the test of unbiasedness when the null is true is not standard normal, but has a variance in excess of unity, so that using critical values taken from the standard normal will lead to the null being rejected too often – we will tend to falsely infer that forecasts are biased if we fail to take account of parameter estimation uncertainty. Notice that when there are a large number of in-sample observations relative to forecast period observations the problem will be less acute (π positive but close to zero) and for low values of π can perhaps be ignored. Intuitively, there are sufficient in-sample observations relative to periods being forecast that the estimation error will be small.

Some of the above expressions simplify considerably if we assume that the model simply consists of an intercept:

$$y_t = \beta^* + \varepsilon_t$$

corresponding to $X_{t-1} = (1)$. Then:

$$P^{-1/2} \sum_{t=R}^{R+P-1} e_{t+1|t} = (1 \quad -1) \begin{pmatrix} P^{-1/2} \sum_{t=R}^{R+P-1} \varepsilon_{t+1} \\ \pi^{1/2} R^{-1/2} \sum_{s=1}^{R} \varepsilon_s \end{pmatrix}$$

and:

$$\begin{pmatrix} P^{-1/2}\sum_{t=R}^{R+P-1}\varepsilon_{t+1} \\ \pi^{1/2}R^{-1/2}\sum_{s=1}^{R}\varepsilon_s \end{pmatrix} \xrightarrow{d} N\left(0_{2\times 1}, \sigma^2\begin{pmatrix} 1 & 0 \\ 0 & \pi \end{pmatrix}\right)$$

so $V(P^{-1/2}\sum_{t=R}^{R+P-1} e_{t+1|t}) = \sigma^2(1+\pi)$. This corresponds to the simple case of testing whether the mean of a sample of i.i.d. normal variables is an unbiased predictor of the mean of another i.i.d. sample from the same population.

In the absence of estimation error, or assuming $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ in (2.49), so that $e_{t+1|t} = \varepsilon_{t+1}$,

$$P^{-1/2}\sum_{t=R}^{R+P-1} e_{t+1|t} = P^{-1/2}\sum_{t=R}^{R+P-1}\varepsilon_{t+1}$$

and

$$P^{-1/2}\sum_{t=R}^{R+P-1}\varepsilon_{t+1} \xrightarrow{d} N(0,\sigma^2)$$

resulting in $V = 1$.

West (2001) shows that the forecast encompassing test based on $E(d_t) = 0$, where $d_t = e_{1t}(e_{1t} - e_{2t})$ (see Section 2.3.2) will tend to reject too often when forecasts are based on models with estimated parameters, and no allowance is made for this. As suggested by the example of testing for unbiasedness outlined above, the problem will be more acute the larger π. West concludes that the Harvey *et al.* (1998) modified forecast encompassing tests may perform reasonably well when parameters are estimated provided that either P is small ($P \leq 8$) or less than 10% of the total observations are assigned to the forecast period ($P < 0.1(P+R)$).

Finally, Clark and McCracken (2001) show that the asymptotic distributions of tests of equal forecast accuracy and encompassing will differ when the models are nested, rather than non-nested (as assumed hitherto). When the models are nested (and abstracting from parameter estimation uncertainty by assuming the models' parameters are known in advance) then the forecast errors are identical under the null, and the standard distribution theory does not apply. They tabulate critical values for a test statistic that is applicable in this case. The distribution depends

on π and the number of excess variables in the nesting model, as well as whether the parameter estimates are continuously updated (or estimated once on the data to R and kept fixed, as above).

2.5 Non-linear models and forecasting

We consider in this section non-linear models of the sort that characterize processes which switch between two or more regimes. Models with two or more regimes are a natural generalisation of 'one-regime' linear ARMA time-series models. Such models have been found to be a useful way of capturing business-cycle regimes in the empirical macroeconomics literature, as well as describing high and low volatility regimes in financial market returns variables. For linear ARMA models the best point forecast assuming squared-error loss is simply the conditional expectation, which is simple to calculate and has a closed analytical form. For one of the two types of regime-switching models we consider, namely the Markov-switching autoregessive models (MSAR), the conditional expectation has a closed form solution, whereas for the threshold models the conditional expectation needs to be found by simulation or numerical integration. As well as the issue of how to generate forecasts from non-linear models, we consider issues that arise in the evaluation of forecasts from such models.

In Section 2.5.1 we provide a simple proof that the conditional expectation is the MMSE predictor whether the model is linear or non-linear, to motivate the relevance of calculating the conditional expectation for the non-linear models discussed in this section. In Section 2.5.2 we show why the calculation of multi-step forecasts for a generic non-linear model may not be as straightforward as for linear models. Section 2.5.3 introduces the class of threshold models, and Section 2.5.4 MSAR. Section 2.5.5 considers issues that might be particularly relevant to the evaluation of forecasts from non-linear models.

2.5.1 The conditional expectation is the MMSE predictor

Provided the first two moments of the process exist, forecasts calculated as the conditional expectation $y_{t+h|t} = E(y_{t+h}|Y_t)$ are unbiased (where $Y_t = (y_1 \cdots y_t)$), and no other predictor conditional on Y_t alone has a smaller MSFE, $E[(y_{t+h} - y_{t+h|t})^2|Y_t]$. We sketch out a proof of this result for $h = 1$: see, for example, Clements and Hendry (1998, ch. 2) for $h > 1$.

Consider any alternative predictor that depends on the same information set, Y_t, say $\tilde{y}_{t+1|t} = g(y_{t+1}|Y_t)$ such that:

$$E(\tilde{y}_{t+1}|Y_t) = E[g(y_{t+1}|Y_t)] = E[y_{t+1}|Y_T],$$

so $\tilde{y}_{t+1|t}$ is also conditionally unbiased. Then:

$$\begin{aligned} E[(y_{t+1} - \tilde{y}_{t+1|t})^2|Y_t] &= E\{[(y_{t+1} - y_{t+1|t}) - (\tilde{y}_{t+1|t} - y_{t+1|t})]^2|Y_t\} \\ &= E[(y_{t+1} - y_{t+1|t})^2 + (\tilde{y}_{t+1|t} - y_{t+1|t})^2|Y_t] \\ &= E[(y_{t+1} - y_{t+1|t})^2|Y_t] + \upsilon \\ &\geq E[(y_{t+1} - y_{t+1|t})^2|Y_t], \end{aligned} \tag{2.50}$$

as $\upsilon \geq 0$, and the cross product:

$$E[(y_{t+1} - y_{t+1|t})(\tilde{y}_{t+1|t} - y_{t+1|t})|Y_t] = 0$$

by the unbiasedness of $y_{t+1|t}$ and the fact that both $y_{t+1|t}$ and $\tilde{y}_{t+1|t}$ are conditional on Y_t. Thus, $y_{t+1|t}$ has desirable properties. It is conditionally unbiased and no other unbiased predictor has a smaller variance.

For the AR(1) model given by (2.9) in Section 2.1.1, $y_t = \phi y_{t-1} + \upsilon_t$, we can write y_{t+h} as:

$$y_{t+h} = \phi^h y_t + \sum_{t=0}^{h-1} \phi^i \upsilon_{t+h-i}$$

(by substituting for $y_{t+h-1} = \phi y_{t+h-2} + \upsilon_{t+h-1}$ in $y_{t+h} = \phi y_{t+h-1} + \upsilon_{t+h}$ to give $y_{t+h} = \phi^2 y_{t+h-2} + \upsilon_{t+h} + \phi\upsilon_{t+h-1}$, and then substituting for y_{t+h-2}, and so on). The conditional expectation is:

$$y_{t+h|t} \equiv E(y_{t+h}|Y_t) = E\left[\left(\phi^h y_t + \sum_{t=0}^{h-1} \phi^i \upsilon_{t+h-i}\right)\middle| Y_t\right] = \phi^h y_t$$

because:

$$E\left(\sum_{t=0}^{h-1} \phi^i \upsilon_{t+h-i}|Y_t\right) = \sum_{t=0}^{h-1} \phi^i E(\upsilon_{t+h-i}|Y_t)$$

and $E(\upsilon_{t+s}|Y_t) = E(\upsilon_{t+s}) = 0$ for all $s > 0$.

2.5.2 Multi-step forecasts and non-linear models

To see why a closed-form solution for the conditional expectation of a non-linear model may not exist, consider:

$$y_t = g(y_{t-1}; \phi) + \varepsilon_t, \tag{2.51}$$

where $g(y, \phi)$ is a non-linear function. For example, $g(y, \phi) = \phi y^2$ gives a non-linear equation, while $g(y, \phi) = \phi y$ specifies an AR(1). As usual, $\{\varepsilon_t\}$ is assumed to be a zero-mean, i.i.d. random variable, $E(\varepsilon_t^2) = \sigma_\varepsilon^2$, with distribution function F (so that $\Pr(\varepsilon_t < \varepsilon) = F(\varepsilon)$). Assuming $g(\cdot)$ is known, the conditional expectation is simply:

$$y_{t+1|t} \equiv E[y_{t+1}|Y_t] = E[(g(y_t; \phi) + \varepsilon_{t+1})|Y_t] = g(y_t; \phi). \tag{2.52}$$

whatever the form of $g(\,)$. So 1-step forecasts from non-linear models are obtained in the same way as 1-step forecasts from linear models.

Consider now the 2-step forecast (suppressing the dependence of $g(\,)$ on ϕ for convenience):

$$\begin{aligned} y_{t+2|t} \equiv E(y_{t+2}|Y_t) &= E[(g(y_{t+1}) + \varepsilon_{t+2})|Y_t] \\ &= E[g(y_{t+1})|Y_t] + E(\varepsilon_{t+2}|Y_t) \\ &= E[g(y_{t+1})|Y_t]. \end{aligned}$$

Consider $E[g(y_{t+1})|Y_t]$.

$$\begin{aligned} E[g(y_{t+1})|Y_t] &= E[g(g(y_t) + \varepsilon_{t+1})|Y_t] \\ &= E[g(y_{t+1|t} + \varepsilon_{t+1})|Y_t], \end{aligned}$$

where the second line follows from using (2.52). The point to note is that for a non-linear function:

$$E[g(\cdot)] \neq g(E[\cdot])$$

so that:

$$E[g(y_{t+1|t} + \varepsilon_{t+1})|Y_t] \neq g[E(y_{t+1|t}|Y_t) + E(\varepsilon_{t+1}|Y_t)].$$

When $g(y)$ is linear (e.g., as for the AR(1) model above):

$$E[(\phi y_{t+1|t} + \phi\varepsilon_{t+1})|Y_t] = \phi y_{t+1|t},$$

but for $g(y) = \phi y^2$:

$$E[(\phi y_{t+1|t}^2 + \phi\varepsilon_{t+1}^2 + 2\phi y_{t+1|t}\varepsilon_{t+1})|Y_t] = g(y_{t+1|t}) + \phi\sigma_\varepsilon^2.$$

Only when $\sigma_\varepsilon^2 = 0$ will it be true that $y_{t+2|t} = g(y_{t+1|t})$. And then $\{y_t\}$ is non-stochastic.

Granger and Teräsvirta (1993, ch. 8) consider four alternative methods for forecasting 2-steps ahead, based on:

$$\begin{aligned} y_{t+2|t} &= E[g(y_{t+1})|Y_t] \\ &= E[g(y_{t+1|t} + \varepsilon_{t+1})]. \end{aligned} \tag{2.53}$$

Naive or skeleton method This method assumes that $E[g(\cdot)] = g(E[\cdot])$, such that $y_{t+2|t}^n = g(y_{t+1|t})$. The random variable $\{\varepsilon_{t+1}\}$ in (2.53) is effectively replaced by its mean value of zero. This method is generally not to be recommended.

Exact method Requires numerical integration to solve (2.53):

$$y_{t+2|t}^e = \int_{-\infty}^{\infty} g(y_{t+1|t} + z)dF(z). \tag{2.54}$$

Monte Carlo As an approximation to the exact method, we can instead average over $g(\cdot)$ evaluated at a number of randomly chosen values of $\{\varepsilon_{t+1}\}$. The random variables $\{z_j\}$ are drawn from the distribution F. Therefore, values of the random variable with a higher probability under F will be drawn more frequently than low probability values (e.g. values around zero compared to values less than -2 or greater than $+2$, in the case of a normal random variable). This performs the role played by $F(z)$ in (2.54).

$$y_{t+2|t}^{\text{mc}} = \frac{1}{R}\sum_{j=1}^{R} g(y_{t+1|t} + z_j).$$

Bootstrap Similar to Monte Carlo, except that the random variables $\{\hat{\varepsilon}_j\}$ are drawings from the model's estimated error terms, and may be preferred when F is unknown or cannot be easily sampled from.

$$y_{t+2|t}^{\text{bs}} = \frac{1}{B}\sum_{j=1}^{B} g(y_{t+1|t} + \hat{\varepsilon}_j).$$

As the number of replications R is increased the Monte Carlo method (and the bootstrap) will provide an increasingly good approximation to the exact method, and is no more difficult computationally to implement for longer horizon forecasts. By way of contrast, the exact method requires numerical evaluation of a double integral for $h=3$, and for longer forecasts soon becomes an unattractive method. For 3-steps ahead:

$$\begin{aligned} y_{t+3|t} &= E[g(y_{t+2|t} + \varepsilon_{t+2})] \\ &= E[g(g(y_{t+1|t} + \varepsilon_{t+1}) + \varepsilon_{t+2})] \\ &= E[g(g(g(y_t) + \varepsilon_{t+1}) + \varepsilon_{t+2})]. \end{aligned}$$

Simulation-based forecasting methods can also be fairly easily extended to look at the effects on forecast performance of parameter estimation uncertainty (the effects of having to estimate the model's parameters) and model uncertainty, whereby the model specification (e.g. an AR(1) versus an AR(2)) is not known but needs to be determined in some way, perhaps using a model selection criterion.

2.5.3 SETAR models and multi-period forecasts

The threshold autoregressive (TAR) model was first proposed by Tong (1978, 1983), Tong and Lim (1980) (see also Tong 1995a). At each point in time, y_t is determined by one of a small number of linear autoregressions. Which autoregression is in force depends upon the value of some past lag of the process relative to a threshold (or set of thresholds), or alternatively it may depend on the value of an extraneous variable. When the threshold variable is a lag of y_t, say, y_{t-d}, so that d is the length of the delay, then the model is 'self-exciting', giving rise to the acronym SETAR. When there are two regimes, then the process is in regime $i = 1$ at period t when $y_{t-d} \leq r$, and otherwise ($y_{t-d} > r$) in regime $i = 2$:

$$\begin{aligned} y_t &= \phi_0^{\{i\}} + \phi_1^{\{i\}} y_{t-1} + \cdots + \phi_p^{\{i\}} y_{t-p} + \epsilon_t^{\{i\}}, \\ \epsilon_t^{\{i\}} &\sim \text{i.i.d.}(0, \sigma^{2\{i\}}), \quad i = 1, 2, \end{aligned} \tag{2.55}$$

where the superscripts $\{i\}$ indicate parameters that may vary across regime. As written, the model allows the variance of the disturbances to depend upon the regime. Stationarity and ergodicity conditions are discussed in, for example, Tong (1995a).

If we assume for the moment that $\sigma^{2\{i\}} = \sigma^2$ (no regime-dependent heteroskedasticity) then in terms of the generic notation of Section 2.5.2 we can write the model as:

$$y_t = g\left(y_{t-1}; \phi^{\{1\}}, \phi^{\{2\}}, r\right) + \varepsilon_t$$

when $p = d = 1$ and with $g(y_{t-1}; \cdot)$ given by:

$$g(y_{t-1}; \phi^{\{1\}}, \phi^{\{2\}}, r) = [\phi^{\{1\}} + 1(y_{t-1} > r)(\phi^{\{2\}} - \phi^{\{1\}})]y_{t-1}. \tag{2.56}$$

In equation (2.56) $1(\cdot)$ is the indicator function, that is, $1(y_{t-1} > r) = 1$ when $y_{t-1} > r$ and $1(y_{t-1} > r) = 0$ when $y_{t-1} \leq r$.

The exact 1-step ahead point forecast defined by $y_{t+1|t} \equiv E(y_{t+1}|Y_t)$, where $Y_t = y_t, y_{t-1}, \ldots$, is given by:

$$y_{t+1|t} = E[(g(y_t) + \epsilon_{t+1})|Y_t] = g(y_t).$$

However, for 2-steps ahead:

$$y_{t+2|t} \equiv E(y_{t+2}|Y_t) = E[(g(y_{t+1}) + \epsilon_{t+2})|Y_t] = E[g(y_{t+1})|Y_t]. \tag{2.57}$$

From (2.56):

$$\begin{aligned} g(y_{t+1}; \cdot) &= [\phi^{\{1\}} + 1(y_{t+1} > r)(\phi^{\{2\}} - \phi^{\{1\}})]y_{t+1} \\ &= [\phi^{\{1\}} + 1(y_{t+1|t} + \varepsilon_{t+1} > r)(\phi^{\{2\}} - \phi^{\{1\}})](y_{t+1|t} + \varepsilon_{t+1}). \end{aligned} \tag{2.58}$$

The second line comes from replacing y_{t+1} by $y_{t+1|t} + \varepsilon_{t+1}$: the forecast value plus the 'forecast error'. The non-linearity in the forecast function arises from the presence of $\{\varepsilon_{t+1}\}$ in the indicator function and the conditional mean. Calculating the conditional expectation of (2.58) requires numerical integration or use of a simulation method, as described in Section 2.5.2.

The Monte Carlo method

Given its popularity, we outline the Monte Carlo method (denoted MC) for generating forecasts from a SETAR model. This simulation method can be applied as easily to complex models (high autoregressive lag orders, several regimes) as to the simple two-regime SETAR model with $p = d = 1$. For forecasting the $t + 1$ to $t + h$ observations conditional on Y_t, we draw a vector of i.i.d. variables from F, the distribution function of

the $\{\varepsilon_t\}$, which we label as $\{z_{2,j}, \ldots, z_{h,j}\}$, where j denotes the replication. For the jth replication, we solve the following equations:

$$y_{t+1|t} = [\phi^{\{1\}} + 1(y_t > r)(\phi^{\{2\}} - \phi^{\{1\}})]y_t$$

$$y^j_{t+2|t} = [\phi^{\{1\}} + 1(y_{t+1|t} + z_{2,j} > r)(\phi^{\{2\}} - \phi^{\{1\}})](y_{t+1|t} + z_{2,j})$$

$$y^j_{t+3|t} = [\phi^{\{1\}} + 1(y^j_{t+2|t} + z_{3,j} > r)(\phi^{\{2\}} - \phi^{\{1\}})](y^j_{t+2|t} + z_{3,j})$$

$$y^j_{t+4|t} = [\phi^{\{1\}} + 1(y^j_{t+3|t} + z_{4,j} > r)(\phi^{\{2\}} - \phi^{\{1\}})](y^j_{t+3|t} + z_{4,j})$$

and:

$$y^j_{t+h|t} = [\phi^{\{1\}} + 1(y^j_{t+h-1|t} + z_{h,j} > r)(\phi^{\{2t\}} - \phi^{\{1\}})](y^j_{t+h-1|t} + z_{h,j}).$$

Repeating for $j = 1, \ldots, R$ gives a sample of k-step ahead forecasts ($k = 1, \ldots, h$), $\{y^1_{t+k|t}, \ldots, y^R_{t+k|t}\}$, and averaging yields the MC estimator of the point forecast:

$$y^{\text{mc}}_{t+k|t} = \frac{1}{R}\sum_{j=1}^{R} y^j_{t+k|t}.$$

If there is regime-dependent heteroskedasticity, then the drawing of $\{z_{i,j}\}$, say, can be scaled to have a variance appropriate to the regime the process is in at period $t + i$, as determined by $y^j_{t+i-1|t}$.

Smooth transition threshold models

The SETAR model features an abrupt switch from one regime (linear autoregression) to another as y_{t-1} (with $d = 1$) crosses the threshold value, r. Smooth transition autoregessive (STAR) models allow for a more gradual adjustment, whereby the process is typically determined by some weighted average of the two regimes, and the relative importance of the two regimes in the average depends upon y_{t-1}. The two-regime STAR model can be written as:

$$y_t = [\phi^{\{1\}} + G(\gamma, c; y_{t-1})(\phi^{\{2\}} - \phi^{\{1\}})]y_{t-1} + \varepsilon_t,$$

which is equivalent to the SETAR process in (2.56), except that $1(y_{t-1} > r)$ is replaced by $G(\gamma, c; y_{t-1})$, where, for example:

$$G(\gamma, c; y_{t-1}) = (1 + \exp\{-\gamma(y_{t-1} - c)\})^{-1}, \quad \gamma > 0. \tag{2.59}$$

The transition function increases monotonically from zero to unity as y_{t-1} goes from minus to plus infinity. The smoothness parameter γ controls the slope of the transition function, that is, the speed with which the process moves between regimes as y_{t-1} varies. When $y_{t-1} = c$

the two regimes each receive an equal weight, $G(\gamma, c; c) = \frac{1}{2}$. As $\gamma \to \infty$ then (2.59) becomes a step function, $G(\gamma, c; y_{t-1}) \to 1(y_{t-1} > c)$. As $\gamma \to 0$ then $G(\gamma, c; y_{t-1}) \to \frac{1}{2}$, and the model becomes linear, with the AR parameter ϕ not depending on y_{t-1}. This model may be used to capture different dynamic behaviour in expansions and contractions, for example. A model with transition function given by (2.59) is often known by the acronym LSTAR, for logistic STAR model.

Any function which is bounded between zero and one is a suitable candidate for $G(\cdot)$. Another popular choice is the exponential function:

$$G(\gamma, c; y_{t-1}) = 1 - \exp\{-\gamma(y_{t-1} - c)^2\}$$

giving rise to the ESTAR model. $G(\cdot)$ is symmetric about 0 (attained when $y_{t-1} = c$), and approaches unity as $y_{t-1} \to \pm\infty$. The dynamics of the model are therefore similar as y_{t-1} departs from c in either direction, but differ in a band around c. Such a transition function be useful in describing bands of inactivity around an equilibrium in the presence of adjustment costs: see, for example, Anderson (1997). The specification, estimation and evaluation of smooth transition models is described in Teräsvirta and Anderson (1992) and Teräsvirta (1994), *inter alia*.

2.5.4 Markov-switching models

In MSAR processes the switch between regimes is determined by a stochastic process, whereby at each period t there is a constant probability of remaining in a given regime, say regime 1, which we denote by p_{11}, and therefore a probability of switching into the other regime (assuming a two-regime model) of $p_{12} = 1 - p_{11}$. Similarly, p_{22} is the probability of remaining in regime 2, and $p_{21} = 1 - p_{22}$ the probability of switching from regime 2 to 1. The stochastic process just described is a Markov process, because we assume that the transition probabilities depend only on the current state. More formally, the transition probabilities can be written as:

$$p_{ij} = \Pr(s_{t+1} = j | s_t = i), \quad \sum_{j=1}^{2} p_{ij} = 1 \ \ \forall i, j \in \{1, 2\},$$

where the unobservable states are given by $s_t = 1$, if the process is in state (regime) 1 at period t, and otherwise $s_t = 2$, indicating state or regime 2. The assumption of fixed transition probabilities p_{ij} has been relaxed by a number of authors.[5]

Whilst the states $\{s_t\}$ evolve independently of the observed values $\{y_t\}$, from the data $\{y_t\}_{t=1}^{T}$ we can infer the state the process was in at each period t. In the seminal paper by Hamilton (1989), y_t was the growth rate

of US GNP. The Hamilton model of the US business cycle fits an AR(4) to the quarterly percentage change in US real GNP from 1953 to 1984:

$$y_t - \mu(s_t) = \alpha_1(y_{t-1} - \mu(s_{t-1})) + \cdots + \alpha_4(y_{t-4} - \mu(s_{t-4})) + u_t, \qquad (2.60)$$

where $u_t \sim$ i.i.d. $N[(0, \sigma_u^2)]$. The conditional mean $\mu(s_t)$ switches between two states:

$$\mu(s_t) = \begin{cases} \mu_1 > 0, & \text{if } s_t = 1(\text{'expansion' or 'boom'}), \\ \mu_2 < 0, & \text{if } s_t = 2(\text{'contraction' or 'recession'}), \end{cases}$$

Maximum likelihood (ML) estimation of the MSAR model entails an iterative technique, based on an implementation of the expectation maximization (EM) algorithm proposed in Hamilton (1990). The EM algorithm of Dempster *et al*. (1977) is used because the observable time series depends on the s_t, which are unobservable stochastic variables.

Forecasting is straightforward. While the MMSEP is not linear, it can be derived analytically (contrast the threshold models in Section 2.5.3). For convenience, suppose the process is first order:

$$y_t - \mu(s_t) = \alpha(y_{t-1} - \mu(s_{t-1})) + u_t.$$

The 1-step ahead conditional expectation is:

$$y_{t+1|t} \equiv E(y_{t+1}|Y_t) = \hat{\mu}_{t+1|t} + \alpha(y_t - \hat{\mu}_{t|t}),$$

where $\hat{\mu}_{t+i|t} = E(\mu(s_{t+i})|Y_t)$. $\hat{\mu}_{t+1|t}$ is the forecast value of the mean in period $t+1$. The forecast of the mean is a weighted average of μ_1 and μ_2, where the weights are the predicted probabilities of the two regimes:

$$\hat{\mu}_{t+1|t} = \sum_{j=1}^{2} \mu_j \Pr(s_{t+1} = j|Y_t).$$

The predicted regime probabilities are given by:

$$\begin{aligned} \Pr(s_{t+1} = j|Y_t) &= \sum_{i=1}^{2} \Pr(s_{t+1} = j|s_t = i)\Pr(s_t = i|Y_t) \\ &= \sum_{i=1}^{2} p_{ij} \Pr(s_t = i|Y_t). \end{aligned}$$

They depend on the transition probabilities, $\{p_{ij}\}$, and the filtered probablilities of the regimes at period t, $\{\Pr(s_t = i \mid Y_t)\}$. $\hat{\mu}_{t|t}$ can be calculated from the filtered probabilities alone:

$$\hat{\mu}_{t|t} = \sum_{j=1}^{2} \mu_j \Pr(s_t = j|Y_t).$$

Multi-step forecasts can be built up from the recursion:

$$y_{t+h|t} = \hat{\mu}_{t+h|t} + \alpha(y_{t+h-1|t} - \hat{\mu}_{t+h-1|t})$$

with initial values $y_{t+s|t} = y_{t+s}$ for $s \leq 0$, and where the predicted regime probabilities are given by:

$$\Pr(s_{t+h} = j|Y_t) = \sum_{i=1}^{2} \Pr(s_{t+h} = j|s_t = i)\Pr(s_t = i|Y_t). \tag{2.61}$$

Collecting the transition probabilities in the matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

it is straightforward to show that $\Pr(s_{t+h} = j|s_t = i) = \{(P')^h\}_{j,i}$, that is, the probability of being in regime j, h periods after being in regime i, is given by the $\{j, i\}$ element of $(\mathbf{P}')^h$.

When $p_{ij} = p_j$, $i, j = 1, 2$, the regimes are unpredictable, where p_j is the unconditional probability of regime j (given by the relative frequency of occurrence of regime j). Then $\Pr(s_{t+h} = j|Y_t) = p_j$, $\hat{\mu}_{t+h|t} = \sum_{j=1}^{2} \mu_j p_j = \bar{\mu}$, say, and the h-step forecasts are given by:

$$\begin{aligned} y_{t+h|t} &= \bar{\mu} + \alpha(y_{t+h-1|t} - \bar{\mu}) \\ &= \bar{\mu} + \alpha^h(y_t - \bar{\mu}) \end{aligned}$$

matching the forecast function of a linear AR(1) model.

2.5.5 Evaluating non-linear model forecasts

A number of considerations arise when forecasts from non-linear models are involved.

First, it is often argued that non-linear models will be better in some states than others, for example, Tong (1995b, pp. 409–410, 'how well we can predict depends on where we are' and that there are 'windows of opportunity for substantial reduction in prediction errors'. If those

occasions which favour the non-linear model are relatively infrequent, then the good performance at those times may be diluted by averaging squared forecast errors over all periods. This will be especially misleading if those occasions happen to be times when the user particularly values accurate forecasts (e.g. at turning points in the economy). This has led to the practice of reporting MSFEs for specific regimes, where the regimes are determined by the model designation at the time the forecast is made.

Second, the value of non-linear models may not be apparent from empirical forecast comparison exercises if the 'non-linearity' fails to persist into the future (e.g. Granger and Teräsvirta (1993, p. 164).

Third, the forecast user may often be interested in how well the direction of change of a variable is forecast, and less interested in the magnitude of the forecast error *per se*. Tests of sign predictability were developed by Henriksson and Merton (1981), *inter alia*, and are sometimes known as market-timing tests. To see why, consider an investor who has to decide whether to buy or sell an asset. If the asset price rises tomorrow, he would wish to have bought today, and otherwise to have sold. A forecast of the sign will be valuable to the forecaster if it helps predict the actual sign: it will not be useful if the forecast sign is independent of the sign of the actual movement. Related ideas in the macroeconomic forecasting literature appear in, for example, Schnader and Stekler (1990), Stekler (1994) and Pesaran and Timmermann (1992). Henriksson and Merton (1981) show that the test of market timing is asymptotically equivalent to the standard chi-squared test of independence for the 2×2 contingency table. A chi-squared (3 degrees of freedom) test of independence between the actual and predicted directions is calculated as:

$$\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

based on:

$$\begin{array}{lll} & & \text{outcome} \\ & & \begin{array}{cc} \text{up} & \text{down} \end{array} \\ \text{forecast} & \begin{array}{c} \text{up} \\ \text{down} \end{array} & \begin{bmatrix} n_{uu} & n_{ud} \\ n_{du} & n_{dd} \end{bmatrix}, \end{array}$$

where the n's are the number of occurrences of the joint events. O_i is the observed number in cell i, where i is one of the four events, and E_i is the expected number assuming independendence. For example, letting cell

$i = 1$ refer to 'actual up & predicted up':

$$E_1 = n\left(\frac{n_{uu} + n_{du}}{n} \times \frac{n_{uu} + n_{ud}}{n}\right),$$

where $(n_{uu} + n_{du})/n$ is the probability of 'actual up' and the probability of 'forecast up' is $(n_{uu} + n_{ud})/n$.

Evaluating a model in terms of how well it forecasts the direction of change of a variable would appear to be particularly relevant for business-cycle models of output growth, such as the Hamilton (1989) MSAR model, or for regime-switching models more generally. For example, the threshold value for 2-regime SETAR models of post-War US output growth is often found to be close to zero (e.g., Potter (1995)) so that a correct prediction of the sign of output growth corresponds to correctly predicting the (model-designated) regime. The ability of the MSAR model to 'predict' regimes in-sample, that is, to give an assignation of observations to regimes that closely matched the NBER business-cycle chronology, was one reason for the popularity of the Hamilton (1989) model. To illustrate, Figure 2.1 plots the smoothed and

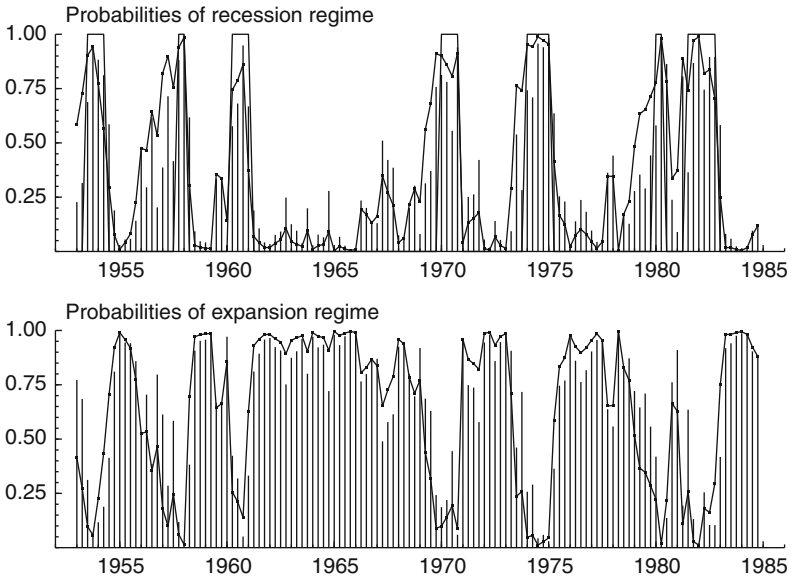


Figure 2.1 Filtered and smoothed regime-probability estimates for the Hamilton (1989) two-regime MSAR model of US output growth, 1953–84

filtered probabilities of the recessionary and expansionary regimes of the Hamilton (1989) model for the original sample period of 1953:1 to 1984:4 (for quarterly, seasonally-adjusted US GNP growth rates, where 1953:1 refers to the first quarter of 1953, etc.). The smoothed full-sample probabilities are the dotted lines, and the filtered probabilities the vertical bars. The filtered regime probabilities are calculated as $\Pr(s_t = i|Y_t) = \Pr(s_t = i|y_t, y_{t-1}, \ldots, y_0)$, $i = 1, 2$, and the smoothed probabilities as $\Pr(s_t = i|Y_T) = \Pr(s_t = i|y_T, \ldots, y_{t+1}, y_t, y_{t-1}, \ldots, y_0)$. That is, the latter use full-sample information, including observations known only after period t. The NBER business-cycle recessions are superimposed on the recession probability estimates in the top panel. It is clear that the MSAR model does a good job at determining which observations come from which regime. The time series of the quarterly growth rates is given in Figure 2.2.

We close this chapter by reporting on a study by Clements and Smith (1999) into the forecast performance of non-linear models, that addresses a number of these issues. Those authors undertake a Monte Carlo study to enable the multi-period forecast performance of non-linear models to be gauged with some precision, and also to ensure that the future realizations of the process have the same non-linear imprint as the past. Data are simulated from estimated models for variables such as exchange

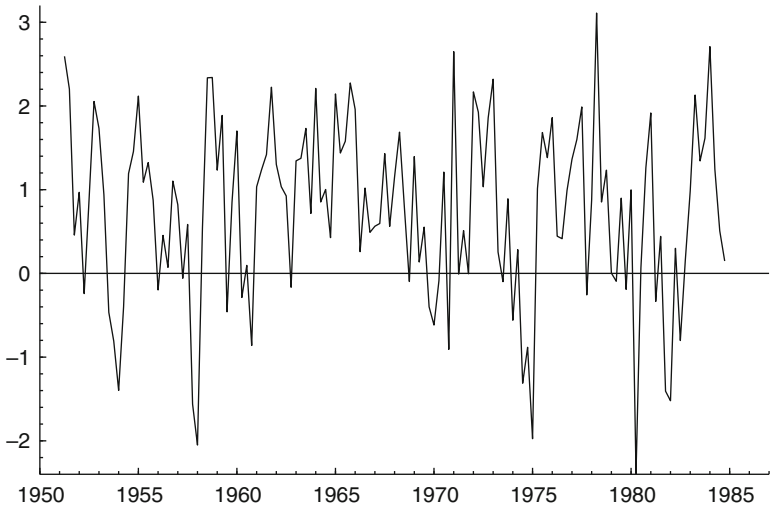


Figure 2.2 US quarterly GNP growth, 1951–84

rates, output growth, etc. so that the simulated data have properties that match empirical processes. The Monte Carlo approach also allows us to explore other aspects of forecasting with non-linear models which have occupied recent investigators, such as the dependence on regime at which the forecast is made, and the impact on forecast performance of parameter estimation and model uncertainty. Here we assume that the SETAR model has the same specification as the SETAR DGP and that the values of the parameters are known. Forecasts for the SETAR model are calculated by the Monte Carlo method (as described in Section 2.5.2) and compared to those of a linear AR model. Thus, for each of $R = 1000$ replications a realization $\{y_t\}_{t=1}^{T+h}$ is generated from the SETAR DGP by replacing the disturbances by normal random variates. Because of the difficulty of deducing the values of the AR model parameters, the AR models are estimated on $\{y_t\}_{t=1}^{T}$, where T is large ($T = 500$) to ensure estimation error for the AR is small. The choice of lag order for the AR model is done pragmatically: we choose the AR process that gives the best forecasts. The AR model is used to forecast the observations $\{y\}_{T+1}^{T+H}$, and the resulting forecast errors are then stored. The SETAR model forecasts are obtained by averaging over an additional $R_f = 500$ realizations of $\{y\}_{T+1}^{T+H}$ for each of the R replications. These realizations are generated from drawings of the errors from the normal distribution with appropriate regime-specific error variances.

Here we report only the results for US GNP. Tiao and Tsay (1994) compare the empirical forecast performance of an AR(2) and a two-regime SETAR model for real US quarterly GNP growth. They find a maximum gain to the SETAR of only 6%, and that at 3-steps ahead. However, dividing up the forecast errors into two groups depending upon the regime at the forecast origin, and then assessing forecast accuracy for each regime separately, the SETAR records gains of up to 15% in the first regime. Because a clear majority of the data points (around three quarters) fall in the second, expansionary regime, the linear model will largely be determined by these points and will match the second-regime of the SETAR model. Thus the forecast performance of the two models is broadly similar for data points in the second regime. However, data points in the first regime of the SETAR model are characterized by different dynamics, so it is here that the SETAR model can gain relative to the linear model.

We analyse a SETAR model similar to that estimated by Potter (1995), who estimates a SETAR(2; 5, 5) but with the third and fourth lags restricted to zero under both regimes. The delay lag $d = 2$, and the model is in the expansionary regime when $y_{t-2} > 0$ (where y_t is the difference of the log of quarterly US GNP) and otherwise in the contractionary phase. The

model we use is the same except that the zero values of the coefficients on the third and fourth lags are not imposed (so that the model corresponds to Potter 1995, table 2, p. 113). A summary of the results is given in Table 2.2.

Comparing the SETAR model to an AR(2), we find a gain of around 17% at 1-step rising to 22% at 2-steps with little to choose between the two thereafter. These 1 and 2-step horizon gains are significant judged by the Diebold–Mariano test. Conditional on being in the Lower (recessionary) regime, the gains at 1 and 2 steps are of the order of 35%. This mirrors the empirical finding of improved forecast accuracy (relative to the linear model) when the economy happens to be in the lower regime. Conditional on the upper regime, the gain is only around 8% at $h = 1$ relative to an AR(2), rising to 16% at $h = 2$.

We have cast the SETAR model in the best possible light, by sidestepping the inherent uncertainty in the selection and estimation of the number of regimes, the delay lag, the threshold value, the autoregressive lag order(s) and coefficient values in each regime. Accepting that each of these will have to be determined empirically in practice may severely limit the usefulness of the non-linear model. For example, estimating the SETAR model autoregressive coefficients (everything else assumed known) reduces the unconditional gain to 5% at $h = 1$ and a maximum

Table 2.2 An evaluation of SETAR and AR models of US GNP on simulated data

	MSFE: AR/SETAR					
Horizon, *h*	**Unconditional**		**Lower regime**		**Upper regime**	
	MSFE ratio	***p*-value**	**MSFE ratio**	***p*-value**	**MSFE ratio**	***p*-value**
1	1.167	0.000	1.351	0.000	1.084	0.000
2	1.219	0.000	1.346	0.000	1.167	0.000
3	1.009	0.399	1.048	0.181	1.018	0.204
5	1.033	0.103	1.046	0.121	1.018	0.157
10	1.012	0.154	1.012	0.124	1.016	0.060

Notes: The *p*-values are of the null of equal MSFEs based on the Diebold and Mariano (1995) test, given by (2.28):

$$\frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \overset{app}{\sim} N(0,1),$$

where a uniform lag window is assumed:

$$\hat{V}(\bar{d}) \approx \tfrac{1}{T}(\hat{\gamma}_0 + 2\textstyle\sum_{i=1}^{h-1} \hat{\gamma}_i).$$

of 11% at $h = 2$. The gain relative to an AR(2) is now only 13% at $h = 1$ conditional upon the lower regime.

2.6 Summary

This chapter serves as a general introduction to the now well-established methods of evaluating point forecasts, as well as reviewing more recent developments. A point forecast is a statement that the rate of inflation next year will be x%, for example. Methods of evaluating sequences of point forecasts and forecast errors in terms of the properties of unbiasedness and efficiency are described, as are the reasons for choosing the conditional expectation as the point forecast.

We also describe general tests of whether differences in the accuracy of rival sets of point forecasts can be attributed to sampling variability, as opposed to indicating that one set of forecasts is more accurate than another. These tests are general in that they can be applied for a wide variety of loss functions or measures of forecast accuracy. In addition, tests of forecast encompassing consider whether a sequence of forecasts from one model (or source) provides useful information even though these forecasts may be less accurate than the forecasts from another model. That is, whether a combination of the two sets of forecasts may be more accurate than the best individual set of forecasts.

We then consider a number of issues that arise when the forecasts are model-based. The discussion of testing for unbiasedness, equal predictive ability and forecast encompassing up to this point applies equally well to survey-based forecasts. Tests of predictive accuracy are described, which essentially compare an estimate of the forecast-error variance, based on the past fit of the model to the observed squared forecast errors. We also describe the impact of parameter estimation uncertainty on some of the tests and sketch the implications for testing for unbiasedness.

Finally, we discuss the calculation of the conditional expectation for commonly used examples of non-linear time-series models, as well as a number of issues specific to the evaluation of forecasts from such models.

3
Volatility Forecasts

3.1 Introduction

Forecasting the conditional variance of a process is primarily of interest if the conditional variance is changing over time.[1] For a large number of financial time-series, as well as some macroeconomic time-series (such as inflation), time-varying conditional variances are an important feature. The autoregressive conditional heteroskedasticity (ARCH) model of Engle (1982), and its generalizations,[2] have become almost indispensable in the modelling of financial series. ARCH models are capable of capturing variances that change (giving rise to clusterings of large (small) changes in the series), as well as other features typical of many financial series, such as thick-tailed unconditional distributions. As an example, Figure 3.1 plots monthly observations on three-month US Treasury Bill interest rates and ten-year Treasury bond interest rates (taken from the Federal Reserve of St Louis database, www.stls.frb.org/fred) and the first differences of these series. The clustering of large and small changes is clearly evident.

Figure 3.2 plots the estimated densities of the changes in the two interest rates, with matched Gaussian densities, illustrating the fatter tails, and the QQ plots[3] against Gaussian densities (with the same means and variances) confirm the departures from normality of the unconditional distributions.

Non-linear ARCH models are also capable of capturing the empirical finding that negative shocks or 'bad news' affect volatility differently from positive shocks. ARCH implies that there may be more uncertainty surrounding the point prediction at some times than at others. Thus, prediction or confidence intervals will vary not just with the forecast horizon, but the whole forecast profile as the horizon lengthens may

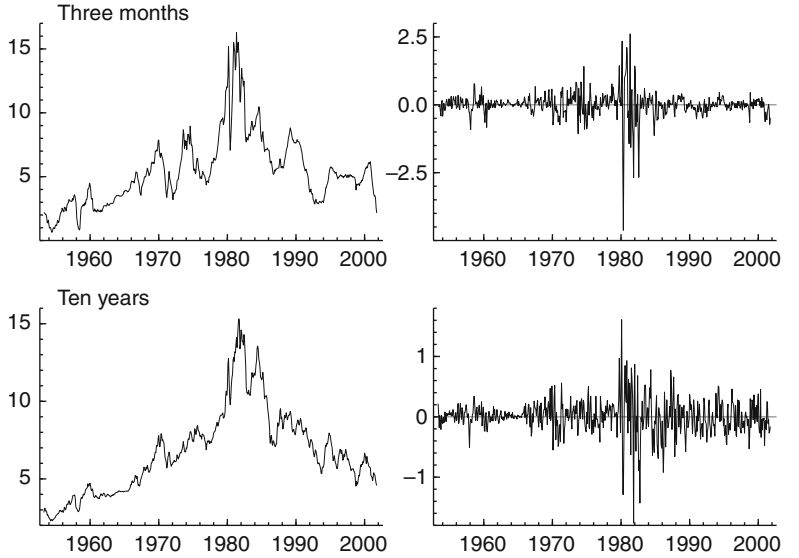


Figure 3.1 Three-month and ten-year monthly interest rates and interest rate changes

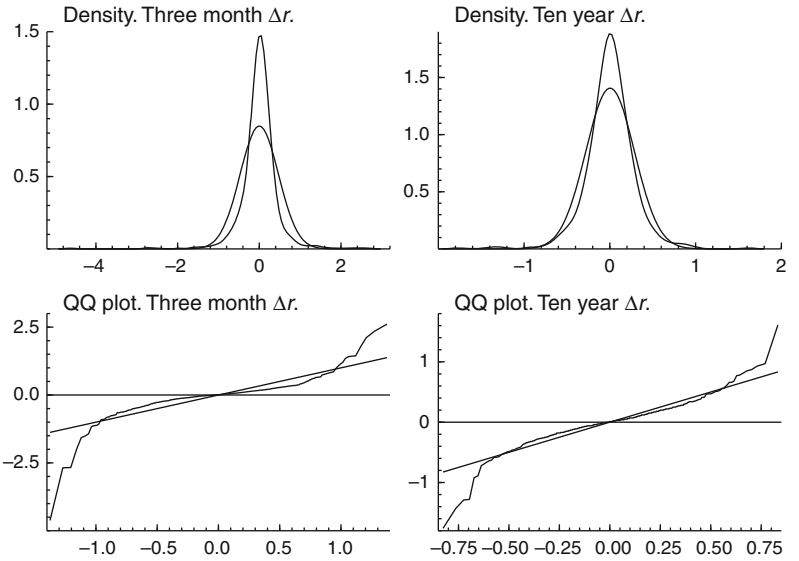


Figure 3.2 Densities and QQ plots of the series of interest rate changes

depend upon y_t, the value of the process at the time the forecast is made, or $y_t, y_{t-1}, \ldots$, more generally. In finance theory modelling the conditional variance of a process has an importance and scope that goes beyond the calculation of more accurate prediction intervals. Forecasts of the conditional variance of the returns on a risky asset are a key ingredient in theories of asset pricing: see, for example, Engle and Bollerslev (1986). In such theories a representative agent is assumed to allocate wealth between the risky asset, and a risk-free asset, where their utility depends positively on the mean return achieved next period but negatively on the variance of next period's return.

In Section 3.2 we begin by looking at the implications of time-varying conditional variances for point forecasts, without specifying a model for the way in which the conditional variance changes. That is, we are not interested in modelling or forecasting the conditional variance at this stage, but only in the extent to which a changing variance might temper the conclusions we reached in Chapter 2 regarding what constitutes the 'best' point forecast. Assuming squared-error loss, time-varying conditional variances are shown not to affect the optimal point forecast. Section 3.3 shows that for more general loss functions the optimal point forecast may depend on the forecast variance. We will then consider the popular ARCH and related approaches to modelling, and therefore forecasting, conditional variance, as well as forecasts based on exponential smoothing: Section 3.4. Assessing the adequacy of conditional variance forecasts raises a number of difficulties because the 'actual' conditional variance is not observed. The problems that arise with traditional evaluation criteria, such as realization-forecast regressions and mean squared forecast error (MSFE), form the subject matter of Section 3.5. Some alternative proposals for evaluating volatility forecasts are discussed in Section 3.6.

3.2 Changing conditional-variances and optimal point forecasts

Our discussion of point forecasting has been based on evaluating the optimal forecast, which equates to the conditional expectation under squared-error loss. Typically the conditional expectation depends on the past of the process, $y_{t+h|t} = g(Y_t)$, and for linear models with known parameters has a simple form, say:

$$y_{t+h|t} = E(y_{t+h} \mid Y_t) = \phi^h y_t$$

for an AR(1), $y_t = \phi y_{t-1} + \varepsilon_t$, where ε_t is i.i.d.$(0, \sigma_\varepsilon^2)$, say. So whilst $E(y_{t+h} \mid Y_t)$ depends on y_t (or Y_t more generally), $V(y_{t+h} \mid Y_t)$ will depend on h but not Y_t. In the case of the AR(1), for example:

$$y_{t+h} = \phi^h y_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}$$

so that:

$$\begin{aligned} V(y_{t+h} \mid Y_t) &\equiv E[(y_{t+h} - E(y_{t+h}))^2 \mid y_t] \\ &= E\left[\left(\sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i} \right)^2 \Big| Y_t \right] \\ &= \sum_{i=0}^{h-1} \phi^{2i} E(\varepsilon_{t+h-i}^2 \mid Y_t) \\ &= \sigma_\varepsilon^2 \frac{1 - \phi^{2h}}{1 - \phi^2}. \end{aligned} \tag{3.1}$$

The third line follows because independence of the $\{\varepsilon_t\}$ implies they are serially uncorrelated, so that the covariance terms from the expected square of the sum in the second line are all zero. The 'identically' distributed part of the assumption on the disturbances gives $E(\varepsilon_s^2) = \sigma_\varepsilon^2$ for all s.

Engle (1982) argued that the failure to allow for the possibility that $V(y_{t+h} \mid Y_t)$ depends on Y_t may be a serious shortcoming: better forecasts of variance (and prediction intervals) may be attainable by modelling the variance as a function of past values, Y_t, rather than assuming the forecast variances (for a horizon h) are all constant, and do not depend on Y_t. Just as one typically allows that $E(y_{t+h} \mid Y_t)$ depends on Y_t, in what has become a seminal paper, Engle proposed a class of ARCH models that allow $V(y_{t+h} \mid Y_t)$ to depend on Y_t.

One way of making the the conditional variance of y_{t+1} depend on Y_t is to specify a form of dependence of this type for the disturbance terms, ε_t. We replace the assumption that ε_t is i.i.d.$(0, \sigma^2)$ with the assumption that the $\{\varepsilon_t\}$ remain serially uncorrelated but are no longer independent, so allowing for dependence in higher-order moments. Specifically:

$$E(\varepsilon_t^2 \mid Y_{t-1}) = h_t,$$

where h_t is a non-negative function of Y_{t-1}, $h_t = h(Y_{t-1})$. That is, the conditional variance of the disturbances depends on the past of Y_t (alternatively, ε_t, ε_{t-1},...). Modelling the conditional variance as a function of the history of the series being analysed matches the univariate time-series approach to modelling the conditional mean.[4] It is usual to define $h_t(Y_{t-1})$ such that

$$\varepsilon_t = z_t\sqrt{h_t},$$

where z_t is i.i.d. $N(0, 1)$.

It is then straightforward to establish the following properties of $\{\varepsilon_t\}$.

1. $E(\varepsilon_t) = 0$

2. $E(\varepsilon_t \mid Y_{t-1}) = 0$

3. $E(\varepsilon_t^2 \mid Y_{t-1}) = h_t$

4. $E(\varepsilon_t^2) = E(h_t)$

(1) follows from $E(\varepsilon_t) = E(z_t\sqrt{h_t}) = E(z_t)E(\sqrt{h_t}) = 0$ because z_t is an innovation on Y_{t-1}, and $h_t = h(Y_{t-1})$. (2) follows from $E(\varepsilon_t|Y_{t-1}) = E(z_t\sqrt{h_t} \mid Y_{t-1}) = \sqrt{h_t}E(z_t \mid Y_{t-1}) = 0$. For (3), $E(\varepsilon_t^2 \mid Y_{t-1}) = E(z_t^2 h_t \mid Y_{t-1}) = h_t E(z_t^2|Y_{t-1}) = h_t$. And finally, (4) from $E(\varepsilon_t^2) = E(E(\varepsilon_t^2 \mid Y_{t-1})) = E(h_t)$. When $E(h_t) = \sigma^2$, that is, the unconditional variance is constant, the $\{\varepsilon_t\}$ process is weakly stationary, because the first- and second-moments of the process are time invariant.

These properties immediately allow us to show that the conditional expectation is unaffected by the $\{\varepsilon_t\}$ exhibiting time-varying conditional variances. For the AR(1) with $\varepsilon_t = z_t\sqrt{h_t}$:

$$E(y_{t+h} \mid Y_t) = E\left[\left(\phi^h y_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}\right) \middle| Y_t\right] = \phi^h y_t$$

because $E[\sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i} \mid Y_t] = \sum_{i=0}^{h-1} \phi^i E(\varepsilon_{t+h-i} \mid Y_t) = 0$, using property (2). We can also show that the conditional expectation remains the MMSEP, assuming squared error loss, by noting that the proof in Section 2.5.1 is unaffected by allowing for dependence in the higher moments of $\{\varepsilon_t\}$.

Dependence in the higher moments of $\{\varepsilon_t\}$ will affect forecast variances of y_{t+h}. From (3.1):

$$\begin{aligned} V(y_{t+h} \mid Y_t) &= E\left[\left(\sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}\right)^2 \Big| y_t\right] \\ &= \sum_{i=0}^{h-1} \phi^{2i} E(\varepsilon_{t+h-i}^2 \mid y_t) \\ &= \phi^{2(h-1)} E(\varepsilon_{t+1}^2 \mid y_t) + \sum_{i=0}^{h-2} \phi^{2i} E(\varepsilon_{t+h-i}^2 \mid y_t), \end{aligned} \tag{3.2}$$

where from property (3) $E(\varepsilon_{t+1}^2 \mid y_t) = h_{t+1} = h(y_t)$. For $h = 1$ the second term in the third line of (3.2) vanishes, so $V(y_{t+1} \mid y_t) = h_{t+1} = h(y_t)$, and the forecast variance depends on y_t. For $h > 1$, the second term is relevant. We have not as yet specified a model that allows one to deduce $E(\varepsilon_{t+s}^2 \mid y_t)$ for $s > 1$, but such terms will also depend on Y_t in general. Because the forecast variability in y_{t+h} depends on y_t, confidence intervals (described in detail in Chapter 4) will depend on the observation at which the forecast was made (in contrast to confidence intervals calculated from (3.1)).

3.3 Time-varying conditional variances and asymmetric loss

Squared-error loss supposes that the cost function depends only on $|e_{t+h|t}|$, that is, the absolute magnitude of the forecast error, such that under and over-predictions of the same magnitude (i.e. positive and negative forecast errors of the same size) attract the same penalty. It is intuitively clear that it then makes sense to aim to make a zero forecast error on average: positive and negative forecast errors will then be realized in equal number due to the unpredictable random variation in y_{t+i} relative to $y_{t+i|t+i-1}$. In the AR(1) with known parameters, the forecast errors will be given by $e_{t+i|t+i-1} = \varepsilon_{t+i}$. However, if positive errors are penalized more heavily than negative errors, it will be optimal (in the sense of minimizing the expected loss) to aim to make a negative expected forecast error, so that the stream of realized errors, $\{e_{t+i|t+i-1}\}$, will be predominantly negative. Moreover, the greater the variablity of the process, the larger the variability of the forecast errors. Assuming that large positive forecast errors are penalized proportionately more heavily than

small positive errors, it will be optimal to aim to make a larger expected negative error on average to guard against the costly large positive errors (which are the more likely the more variable the process).

These ideas are formalized by Granger (1969). Although the conditional mean predictor is not optimal for asymmetric loss, for Gaussian processes a simple fixed adjustment to the conditional mean yields the optimal predictor, where the adjustment depends only on the form of the loss function, and the forecast variance. An immediate implication of the conditional expectation being non-optimal is that biased predictions are consistent with rational behaviour: see Zellner (1986). Christoffersen and Diebold (1997) generalize the results in Granger (1969) to processes which are conditionally Gaussian. They show that if the forecast variance is time varying, the adjustment to the conditional mean that yields the optimal predictor will not be constant. As a consequence, for an asymmetric loss function time variation in the variance of the process will affect the optimal point predictions.

One of the most popular asymmetric loss functions is the 'linex' loss function of Varian (1975). This is commonly used as the optimal predictor that can be solved for analytically. Following Christoffersen and Diebold (1997), we illustrate these ideas with linex loss:

$$C(e_{t+h|t}) = b[\exp(ae_{t+h|t}) - ae_{t+h|t} - 1], \quad a \neq 0, \;\; b \geq 0.$$

For $a > 0$, the loss function is approximately *lin*ear for $e_{t+h|t} < 0$ ('over-predictions'), and *ex*ponential for $e_{t+h|t} > 0$, ('under-predictions'). For small a, the loss function is approximately quadratic:

$$C(e_{t+h|t}) \simeq \frac{ba^2}{2} e_{t+h|t}^2$$

from the first two terms of the Taylor-series expansion.[5]

The optimal predictor h-steps ahead solves:

$$\arg\min_{\hat{y}_{T+h}} E_t[b(\exp(ae_{t+h|t}) - ae_{t+h|t} - 1)], \tag{3.3}$$

where $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h}$, $\hat{y}_{t+h}$ is the optimal predictor, and is assumed to have the form $\hat{y}_{t+h} = y_{t+h|t} + \alpha_{t+h}$, where the process is conditionally Gaussian:

$$y_{t+h} \mid Y_t \sim N(y_{t+h|t}, \sigma_{t+h|t}^2),$$

and $y_{t+h|t}$ is the conditional expectation and $\sigma^2_{t+h|t}$ is the conditional variance (given by (3.2), for example). Substituting for $e_{t+h|t} = y_{t+h} - y_{t+h|t} - \alpha_{t+h}$ in (3.3) gives:

$$\arg\min_{\alpha_{t+h}} E_t[b(\exp(a(y_{t+h} - y_{t+h|t} - \alpha_{t+h})) - a(y_{t+h} - y_{t+h|t} - \alpha_{t+h}) - 1)]$$

and using the result that:

$$E_t[\exp(ay_{t+h})] = \exp\left(ay_{t+h|t} + \frac{a^2\sigma^2_{t+h|t}}{2}\right),$$

and:

$$E_t(y_{t+h}) = y_{t+h|t}$$

gives:

$$\arg\min_{\alpha_{T+h}} b\left[\exp\left(\frac{a^2\sigma^2_{t+h|t}}{2} - a\alpha_{t+h}\right) + a\alpha_{t+h} - 1\right]. \tag{3.4}$$

The first-order condition is satisfied by:

$$\alpha_{t+h} = \frac{a}{2}\sigma^2_{t+h|t}$$

so that the optimal predictor becomes:

$$\tilde{y}_{t+h|t} = y_{t+h|t} + \frac{a}{2}\sigma^2_{t+h|t}. \tag{3.5}$$

To interpret (3.5), assume that $a > 0$ and not close to zero. Costs to under-prediction are weighted more heavily than those to over-prediction. The optimal predictor then exceeds the conditional expectation, so that the conditionally expected error will on average be negative, that is, there will be a tendency to over-predict. The greater the conditionally expected variation ($\sigma^2_{t+h|t}$) around the conditional expectation the greater the tendency to over-predict. As the degree of asymmetry lessens ($a \to 0$) so the optimal predictor approaches the conditional expectation.

3.4 Models of conditional variance

Recall from Section 3.2 that we are defining h_t as the conditional variance of $\{\varepsilon_t\}$:

$$E(\varepsilon_t^2 \mid Y_{t-1}) = h_t$$

such that:

$$\varepsilon_t = z_t\sqrt{h_t}, \tag{3.6}$$

where z_t is i.i.d. $N(0, 1)$. Hitherto we have thought of $\{\varepsilon_t\}$ as one component of $\{y_t\}$. For example, in the AR(1) model with ε_t given by (3.6) we have:

$$y_t = \phi y_{t-1} + \varepsilon_t, \qquad \varepsilon_t = z_t\sqrt{h_t},$$

which jointly models the conditional mean and conditional variance of $\{y_t\}$. For many financial returns series, especially those sampled at a high frequency, there is little if any dependence in the mean of the series, at least compared to the magnitude of the standard deviation of $\{\varepsilon_t\}$, and the mean can be taken to be zero. In that case, what we have termed $\{\varepsilon_t\}$ is itself the returns series. For example, if P_t is the price of an asset at the close of period t:

$$\varepsilon_t = R_t \equiv \ln P_t - \ln P_{t-1} = \ln\left(\frac{P_t - P_{t-1}}{P_{t-1}} + 1\right) \simeq \frac{P_t - P_{t-1}}{P_{t-1}}.$$

When ε_t is the return R_t, it is apparent from (3.6) that $\varepsilon_{t+1} \mid Y_t \sim N(0, h_{t+1})$, given the assumption that $\{z_t\}$ is normal. That is, the one step-ahead conditional distribution of returns is normal. When instead the conditional mean is non-zero, for example, interpreting $\{y_t\}$ as the returns series, $y_{t+1} \mid Y_t \sim N(\phi y_t, h_{t+1})$ so that the nature of the one-step conditional distribution is unchanged.

3.4.1 ARCH models

The ARCH model of Engle (1982) models h_t as a linear function of the squares of the past shocks. Restricting the function to the last shock only, ε_{t-1}, gives the ARCH(1):

$$h_t = \omega + \alpha\varepsilon_{t-1}^2. \tag{3.7}$$

Because a variance must be non-negative, restrictions on the values that the parameters can take are required. For the ARCH(1), we have $\omega, \alpha > 0$. For more complicated models the restrictions are more involved.

From (3.7) a large value of ε_{t-1} (of either sign) will give rise to a large h_t and, *ceteris paribus*, a large value of ε_t (of either sign), as z_t in (3.6) is scaled by $\sqrt{h_t}$. This is the volatility clustering phenomenon that gives rise to dependence in the squares $\{\varepsilon_t^2\}$ even though the $\{\varepsilon_t\}$ are serially uncorrelated. The ARCH model can also be written as an AR(1) in $\{\varepsilon_t^2\}$ to underline this point. Letting $h_t = \varepsilon_t^2 - \nu_t$, substituting in (3.7) gives:

$$\begin{aligned}\varepsilon_t^2 - \nu_t &= \omega + \alpha\varepsilon_{t-1}^2 \\ \varepsilon_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \nu_t.\end{aligned}$$

From $\nu_t = \varepsilon_t^2 - h_t = h_t(z_t^2 - 1)$, $E[\nu_t \mid Y_{t-1}] = h_t E[(z_t^2 - 1) \mid Y_{t-1}] = 0$, so that the disturbance term $\{\nu_t\}$ in the AR(1) model is uncorrelated with the regressor, as required. From the AR(1) representation the condition for covariance stationarity of $\{\varepsilon_t^2\}$ is $|\alpha| < 1$,[6] which given that $\alpha \geq 0$ requires that $1 > \alpha \geq 0$. When this condition holds:

$$E(\varepsilon_t^2) = \omega + \alpha E(\varepsilon_{t-1}^2)$$

and so:

$$E(\varepsilon_t^2) = \frac{\omega}{1-\alpha}.$$

Because $V(\varepsilon_t) = E(\varepsilon_t^2)$, the condition on α, that $1 > \alpha \geq 0$, implies that the unconditional variance is homoskedastic.

Taking the fourth moment of $\varepsilon_t = z_t\sqrt{h_t}$:

$$E(\varepsilon_t^4) = E(z_t^4 h_t^2) = E(z_t^4)E(h_t^2).$$

From $E(h_t^2) = V(h_t) + E(h_t)^2$, and $V(h_t) \geq 0$, it follows that $E(h_t^2) \geq E(h_t)^2$, so that:

$$E(\varepsilon_t^4) = E(z_t^4)E(h_t^2) \geq E(z_t^4)E(h_t)^2 = E(z_t^4)E(\varepsilon_t^2)^2.$$

As $z_t \sim N(0,1)$, the kurtosis $E(z_t^4) = 3$. Defining $\sigma^2 = E(\varepsilon_t^2)$, it then follows that $E(\varepsilon_t^4) > 3\sigma^4$, so that the ARCH process $\{\varepsilon_t\}$ exhibits excess kurtosis, that is, it has thicker tails than the normal. Thus the ARCH process is able to capture volatility clustering and thick tails.

From the AR(1) representation for $\{\varepsilon_t^2\}$ it follows immediately that the autocorrelation function is given by:

$$\frac{\mathrm{Cov}(\varepsilon_t^2 \varepsilon_{t-k}^2)}{\sqrt{V(\varepsilon_t^2)}\sqrt{V(\varepsilon_{t-k}^2)}} = \alpha^k.$$

This implies that the correlations will become small quite quickly in k, especially if α is relatively small. However, a number of squared return series exhibit autocorrelations which appear to decay less rapidly in k than implied by the AR(1). A less rapid decay will result from specifying a higher-order autoregression for $\{\varepsilon_t^2\}$. This can be achieved by replacing the ARCH(1) by an ARCH(p), $p > 1$:

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2. \tag{3.8}$$

Substituting $h_t = \varepsilon_t^2 - \nu_t$ as in the case of the ARCH(1), we obtain:

$$\varepsilon_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \nu_t.$$

Assuming the stationarity of the process (the roots of $(1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_p z^p) = 0$ lie outside the unit circle) the unconditional variance exists and is given by:

$$E(\varepsilon_t^2) = \frac{\omega}{1 - \alpha_1 - \cdots - \alpha_p}.$$

An ARCH model is in many ways a very natural way of obtaining a forecast of h_{t+1}. The 1-period ahead forecast of volatility from the ARCH model, $E(h_{t+1} \mid Y_t)$, is given directly from the model as:

$$h_{t+1} = \omega + \alpha_1 \varepsilon_t^2 + \alpha_2 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p+1}^2.$$

The forecast of volatility in $t + 1$ is a weighted sum of the ε's (which we now interpret directly as returns) in the previous p periods, plus the fixed amount ω. If we were instead to assume equal weights, $\alpha_i = 1/p$ for $i = 1, \ldots, p$, and set $\omega = 0$, we can forecast h_{t+1} as:

$$h_{t+1,a} = \frac{1}{p} \sum_{s=0}^{p-1} \varepsilon_{t-s}^2,$$

whereby the forecast of volatility in $t+1$ is a simple average of the p most recently observed squared returns. When $p = 1$ one of the disadvantages of the simple-average approach is most easily seen. Then $h_{t+1,a} = \varepsilon_t^2$, which will also be the forecast value of volatility j-periods ahead, $j > 0$. High values of ε_t^2 will result in forecasts that volatility remains at that high level for all j. By way of contrast, the ARCH(1) forecasts are empirically more plausible. Substituting for ω from $\sigma^2 \equiv E(\varepsilon_t^2) = \omega/(1-\alpha)$ in (3.7) gives:

$$h_t = \sigma^2(1-\alpha) + \alpha\varepsilon_{t-1}^2,$$

so that for period $t+j$ we have:

$$h_{t+j} = \alpha(\varepsilon_{t+j-1}^2 - \sigma^2).$$

To calculate the forecasts, take the conditional expectation of h_{t+j} based on information up to period t:

$$\begin{aligned} h_{t+j|t} - \sigma^2 \equiv E(h_{t+j} \mid Y_t) - \sigma^2 &= \alpha(E(h_{t+j-1}z_{t+j-1}^2 \mid Y_t) - \sigma^2), \\ &= \alpha(h_{t+j-1|t} - \sigma^2), \end{aligned}$$

where we substitute $\varepsilon_{t+j-1}^2 = h_{t+j-1}z_{t+j-1}^2$ and use $E(\varepsilon_{t+j-1}^2 \mid Y_t) = E(h_{t+j-1} \mid Y_t)\, E(z_{t+j-1}^2 \mid Y_t) = E(h_{t+j-1} \mid Y_t) = h_{t+j-1|t}$ for $j > 2$. By repeated backward substitution:

$$\begin{aligned} h_{t+j|t} - \sigma^2 &= \alpha^{j-1}(h_{t+1|t} - \sigma^2) \\ &= \alpha^j(\varepsilon_t^2 - \sigma^2), \end{aligned}$$

noting that $h_{t+1|t} = h_{t+1}$. The forecasts of $h_{t+j|t} \to \sigma^2$ as j gets large whatever the initial deviation of ε_t^2 from σ^2, given the process is stationary ($|\alpha| < 1$). The same is true of the general ARCH process of order p. By combining the expressions for h_t and σ^2 as for the ARCH(1), we obtain:

$$h_{t+1} - \sigma^2 = \alpha_1(\varepsilon_t^2 - \sigma^2) + \alpha_2(\varepsilon_{t-1}^2 - \sigma^2) + \cdots + \alpha_p(\varepsilon_{t-p+1}^2 - \sigma^2).$$

with σ^2 the (unconditional) mean of h_{t+1}. Forecasts of stationary processes revert to the mean as the horizon increases (see, e.g., Clements and Hendry 1998, ch. 4).

3.4.2 Estimation

The method of estimation of ARCH models is not central to our primary concerns of forecasting and forecast evaluation, but the general principles involved will be mentioned briefly. At the level of generality of our treatment, the discussion applies equally well to the generalizations discussed in subsequent sections.

Maximum likelihood (ML) estimation involves choosing values of the unknown parameters, θ (ω and α_i, $i = 1, \ldots, p$, for the conditional variance, but more generally values of the parameters of the conditional mean as well: ϕ for the AR(1)) to make the observed sample as 'likely as possible', in the sense of maximizing the assumed joint density function of the whole sample of data with respect to the parameters. Given the assumption that $\varepsilon_t = z_t\sqrt{h_t}$, and $z_t \sim$ i.i.d. $N(0,1)$, $\varepsilon_t \mid Y_t \sim$ i.i.d. $N(0, h_t)$. Then the conditional log likelihood, of observing ε_t is:

$$\begin{aligned}\ln l_t(\theta) &= \ln\left[\frac{1}{\sqrt{2\pi h_t(\theta)}}\exp\left(-\frac{\varepsilon_t^2}{2h_t(\theta)}\right)\right] \\ &= -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln h_t(\theta) - \frac{\varepsilon_t^2}{2h_t(\theta)}.\end{aligned}$$

For simplicity, we interpret ε_t as the return series and assume the ε_t are zero mean, thus ignoring conditional mean parameters. Given independence, the log likelihood of the whole sample $(\varepsilon_1, \ldots, \varepsilon_T)$ is the sum of the $\ln l_t(\theta)$, $\sum_{t=1}^{T} \ln l_t(\theta)$. The ML estimators are the values of $\theta = (\omega, \alpha_1, \ldots, \alpha_p)$ that maximize $\sum_{t=1}^{T} \ln l_t(\theta)$:

$$\hat{\theta} = \arg\max_{\theta}\left(\sum_{t=1}^{T} \ln l_t(\theta)\right).$$

Assuming the normality assumption for z_t is correct, and that the conditional mean and variance are correctly specified, $\hat{\theta}$ are consistent, asymptotically normal, and have the smallest variance of all consistent estimators.

The assumption that $\{z_t\}$ is normal may be inappropriate for some returns series. The normality assumption implies that the conditional distribution of returns is normal, whereas actual returns data may be better characterized by conditional distributions with fatter tails than the normal. A natural solution is to replace the assumption that $\{z_t\}$ is normal with the assumption that the $\{z_t\}$ come from a Student's t distribution, say. One could then obtain the ML estimators by constructing

the likelihood function assuming that $\varepsilon_t/\sqrt{h_t}$ have a Student's t distribution with particular degrees of freedom. Because that particular choice of distribution function could also be incorrect, rather than attempting to obtain the ML estimators, it is common to construct the quasi ML (QML) estimators. QML is the name given to ML estimation assuming normality of $\{z_t\}$ (i.e., using a Gaussian likelihood function) even when this assumption is false. It can be shown that under general conditions QML estimators are consistent and asymptotically normal, so long as the conditional mean and variance functions are correctly specified, although they will not be asymptotically efficient (ML using the correct distribution will be more precise): see, for example, Bollerslev and Wooldridge (1992).

3.4.3 GARCH models

High-order ARCH models are not very common, in part because of the difficulty of checking and imposing non-negativity and stationarity conditions in estimation, but also because related models, Generalized ARCH (GARCH: Bollerslev 1986) have been found to offer a useful description of the volatility patterns in the returns of a wide range of underlying price series. The simplest GARCH model includes h_{t-1} as an explanatory variable:

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}. \tag{3.9}$$

By substituting for h_{t-1}:

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta(\omega + \alpha\varepsilon_{t-2}^2 + \beta h_{t-2})$$

and for h_{t-2}, and so on:

$$\begin{aligned} h_t &= \omega \sum_{i=0}^{\infty} \beta^i + \alpha \sum_{i=1}^{\infty} \beta^{i-1} \varepsilon_{t-i}^2 \\ &= \frac{\omega}{1-\beta} + \alpha \sum_{i=1}^{\infty} \beta^{i-1} \varepsilon_{t-i}^2 \end{aligned}$$

assuming $\beta < 1$. Thus, the GARCH model is an infinite order ARCH, albeit that the coefficients on the $\{\varepsilon_{t-i}^2\}$ are constrained and depend on only two parameters, α, β.

We have shown that the ARCH(1) implies an AR(1) for $\{\varepsilon_t^2\}$. We can show that the GARCH(1, 1) (one lagged ε_t^2, one lagged h_t) implies an

ARMA(1, 1) for $\{\varepsilon_t^2\}$. Letting $h_t = \varepsilon_t^2 - \nu_t$, substitution into (3.9) gives:

$$\begin{aligned}\varepsilon_t^2 - \nu_t &= \omega + \alpha\varepsilon_{t-1}^2 + \beta(\varepsilon_{t-1}^2 - \nu_{t-1}) \\ \varepsilon_t^2 &= \omega + (\alpha + \beta)\varepsilon_{t-1}^2 + \nu_t - \beta\nu_{t-1} \end{aligned} \tag{3.10}$$

an ARMA(1, 1) for ε_t^2 with ν_t as the disturbance. The stationarity of the process depends on the AR polynomial, that is, whether the root of $1 - (\alpha + \beta)z = 0$ lies outside the unit circle. From $z = 1/(\alpha + \beta)$, and $\alpha, \beta > 0$, $z > 1$ if and only if $(\alpha + \beta) < 1$.

When $\alpha + \beta < 1$, the unconditional variance of the GARCH(1, 1) can be obtained by taking expectations of (3.10) as:

$$E(\varepsilon_t^2) = \omega + (\alpha + \beta)E(\varepsilon_{t-1}^2)$$

so that:

$$\sigma^2 = E(\varepsilon_t^2) = \frac{\omega}{1 - (\alpha + \beta)}.$$

Combining the equations for h_t and σ^2 for the GARCH(1, 1) gives:

$$h_t - \sigma^2 = \alpha(\varepsilon_{t-1}^2 - \sigma^2) + \beta(h_{t-1} - \sigma^2), \tag{3.11}$$

which has the interpretation that the conditional variance will exceed the long-run (or unconditional) variance if last period's squared returns exceed the long-run variance and/or if last periods conditional variance exceeds the unconditional.

Just as ARCH forecasts are related to simple averaging of recent squared returns, GARCH model forecasts can be related to exponential smoothing of squared returns, that is an exponentially weighted moving average (EWMA) of squared returns. The EWMA formula for forecasting $t + 1$ based on Y_t is:

$$\begin{aligned} h_{t+1,\text{ew}} &= \frac{1}{\sum_{s=0}^{\infty}\lambda^s}(\varepsilon_t^2 + \lambda\varepsilon_{t-1}^2 + \lambda^2\varepsilon_{t-2}^2 + \cdots) \\ &= (1 - \lambda)\sum_{s=0}^{\infty}\lambda^s\varepsilon_{t-s}^2, \end{aligned} \tag{3.12}$$

where $\lambda \in (0, 1)$. The weights sum to one, but unlike simple averaging, the largest weight is accorded to the most recent squared return, $(1 - \lambda)$,

and thereafter the weights decline exponentially, $\lambda(1-\lambda)$, $\lambda^2(1-\lambda), \ldots$. We can take the first term out of the summation to give:

$$\begin{aligned} h_{t+1,\text{ew}} &= (1-\lambda)\varepsilon_t^2 + \sum_{s=1}^{\infty} \lambda^s \varepsilon_{t-s}^2 \\ &= (1-\lambda)\varepsilon_t^2 + \lambda \sum_{s=0}^{\infty} \lambda^s \varepsilon_{t-1-s}^2 \\ &= (1-\lambda)\varepsilon_t^2 + \lambda h_{t,\text{ew}} \end{aligned} \tag{3.13}$$

or $h_{t+1,\text{ew}} = \varepsilon_t^2 + \lambda(h_{t,\text{ew}} - \varepsilon_t^2)$. The forecast of volatility is equal to the squared return plus a positive (negative) amount if the current-period volatility exceeds (is less than) the squared return. From (3.13) exponential smoothing corresponds to a restricted GARCH(1, 1) model with $\omega = 0$ and $\alpha + \beta = (1-\lambda) + \lambda = 1$.

Figure 3.3 presents a graphical summary of the results of estimating a GARCH(1, 1) model for Δr_t, the change in the three-month treasury bill rate plotted in Figure 3.1. The model contains four autoregressive lags,

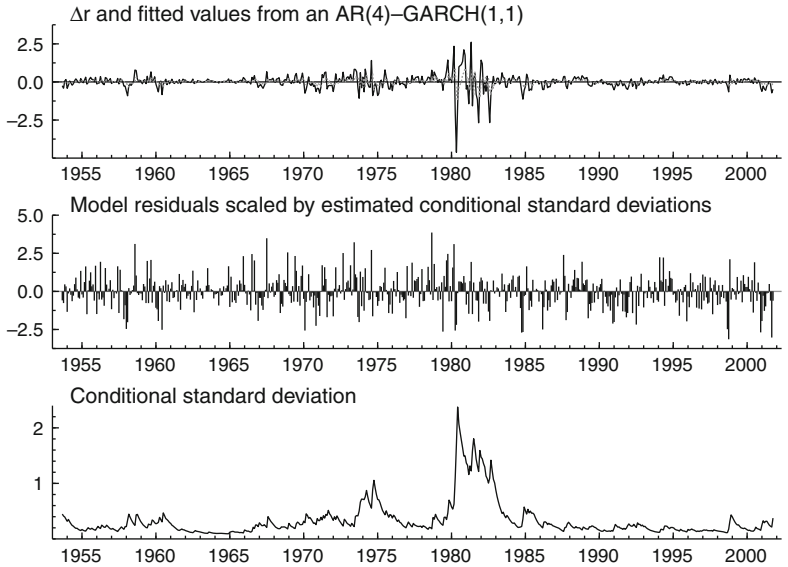


Figure 3.3 Graphical output for an AR(4)–GARCH(1, 1) for the three-month Δr_t

to capture the dependence in the conditional mean. Freely estimated, $\alpha + \beta$ would exceed unity, and so is restricted to 1 in estimation. The conditional variance h_t is high around 1975, and especially in the early 1980s (bottom panel). The middle panel of Figure 3.3 plots the model residuals scaled by $\sqrt{h_t}$: these should be approximately $N(0,1)$ if the GARCH model adequately accounts for the volatility in Δr_t.

3.4.4 GARCH model forecasts

The expected value of the conditional variance next period (one-step ahead forecast) from GARCH models is given directly by the model. $E(h_{t+1} \mid Y_t) = h_{t+1}$, for the GARCH model (3.11) written for $t+1$:

$$h_{t+1} - \sigma^2 = \alpha(\varepsilon_t^2 - \sigma^2) + \beta(h_t - \sigma^2). \tag{3.14}$$

Multi-period forecasts can be calculated as follows. First, write (3.14) for h_{t+j}:

$$\begin{aligned} h_{t+j} - \sigma^2 &= \alpha(\varepsilon_{t+j-1}^2 - \sigma^2) + \beta(h_{t+j-1} - \sigma^2) \\ &= \alpha(h_{t+j-1}z_{t+j-1}^2 - \sigma^2) + \beta(h_{t+j-1} - \sigma^2), \end{aligned}$$

where the second line substitutes $\varepsilon_{t+j-1}^2 = h_{t+j-1}z_{t+j-1}^2$. Taking conditional expectations:

$$\begin{aligned} h_{t+j|t} - \sigma^2 \equiv E(h_{t+j} \mid Y_t) - \sigma^2 &= \alpha(E(h_{t+j-1}z_{t+j-1}^2 \mid Y_t) - \sigma^2) \\ &+ \beta(E(h_{t+j-1} \mid Y_t) - \sigma^2) = (\alpha + \beta)(E(h_{t+j-1} \mid Y_t) - \sigma^2) \end{aligned}$$

using $E(h_{t+j-1}z_{t+j-1}^2 \mid Y_t) = E(h_{t+j-1} \mid Y_t)$ for $j > 2$ (and $E(h_{t+j-1} \mid Y_t) = h_{t+1}$ for $j = 2$). Using the recursive formula relating forecasts of $t+j$ and $t+j-1$, we obtain ($j > 0$):

$$\begin{aligned} h_{t+j|t} - \sigma^2 &= (\alpha + \beta)^{j-1}(h_{t+1} - \sigma^2) \\ &= (\alpha + \beta)^{j-1}[\alpha(\varepsilon_t^2 - \sigma^2) + \beta(h_t - \sigma^2)]. \end{aligned} \tag{3.15}$$

This derivation assumes the existence of σ^2, that is, that $\alpha + \beta < 1$. Therefore $h_{t+j|t} \to \sigma^2$ as $j \to \infty$. $\alpha + \beta$ measures the persistence of the effect of the current shock z_t (more precisely, $\varepsilon_t^2 = h_t z_t^2$) on the volatility forecasts, but assuming stationarity, the forecasts revert (more or less quickly) to σ^2. The exponential smoother extrapolates forward $h_{t+1,\text{ew}}$ for all future periods, and will over-(under-) state volatility in the medium/long term if there is mean-reversion and $h_{t+1,\text{ew}}$ is greater (smaller) than σ^2.

3.4.5 IGARCH

Empirically, $\alpha + \beta$ is often found to be close to 1. $\alpha + \beta = 1$ gives rise to the Integrated GARCH (IGARCH). Some authors such as Lamoureux and Lastrapes (1990) have argued that the IGARCH model may arise through the neglect of structural breaks in GARCH models. They consider standard GARCH models, and GARCH models which allow for structural change through the introduction of a number of dummy variables, for a number of daily stock return series, and use a bootstrap to test between the two models.[7]

Figures 3.4 and 3.5 present forecasts using the AR-GARCH model of Δr_t for three-month Treasury Bills discussed in previous sections. The first is based on the model estimated up to 1991:10, and depicts 1 to 120 step ahead forecasts (1991:10 is the tenth month of 1991, etc.). The second presents 1 to 240 step ahead forecasts from 1981:10. Because $\alpha + \beta = 1$ in both cases, the multi-step sequences of forecasts of h_t increase in the forecast horizon, leading to widening error bars. But notice the difference in scales of the two graphs – starting forecasting from the highly

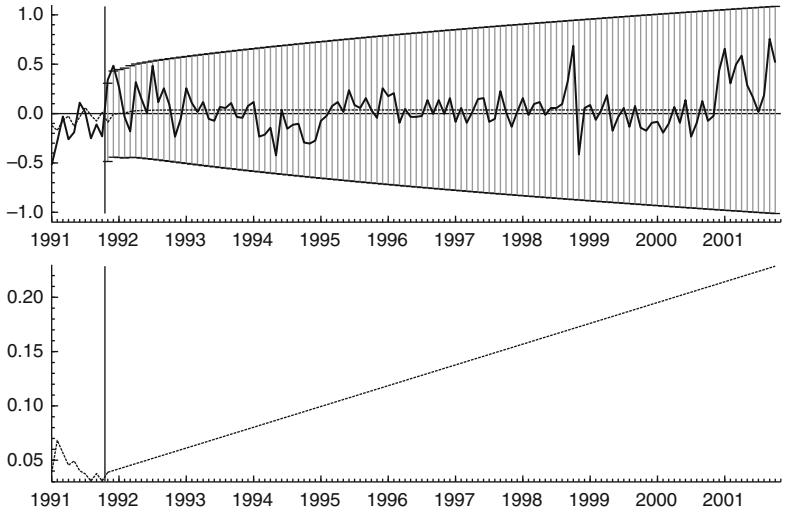


Figure 3.4 Forecasts from an AR(4)–GARCH(1, 1) for Δr_t, with $\alpha + \beta = 1$, starting in 1991:10

Note: The top panel displays the actual and forecast values with 95% error bars. The bottom panel shows multi-step forecasts of h_t.

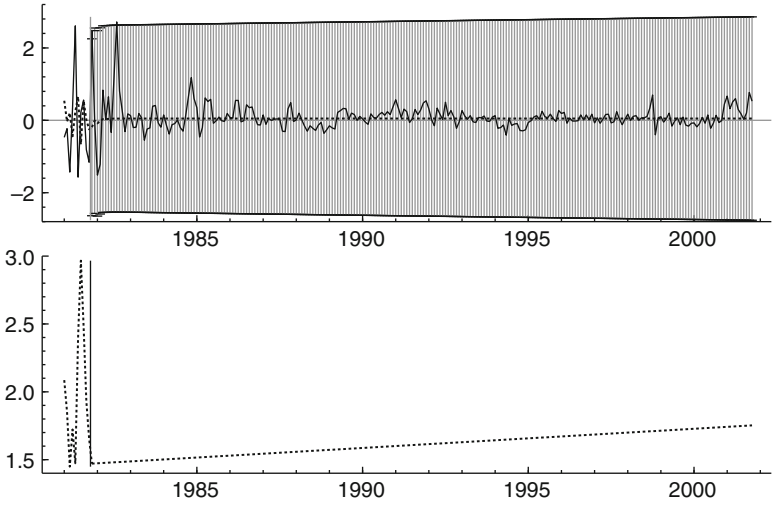


Figure 3.5 Forecasts from an AR(4)–GARCH(1, 1) for Δr_t, with $\alpha + \beta = 1$, starting in 1981:10

Note: The top panel displays the actual and forecast values with 95% error bars. The bottom panel shows multi-step forecasts of h_t.

volatile 1980s (Figure 3.5) generates much wider error bands throughout the period.

There are numerous other extensions, including Fractionally Integrated GARCH (FIGARCH) and ARCH in mean, which are discussed in the references given at the beginning of Section 3.1.

3.4.6 Non-linear GARCH

In addition to the above models of conditional variance, non-linear (G)ARCH models have been developed, primarily to allow positive and negative shocks to have different impacts on the conditional variance of the following observations. In the models considered so far, such as the GARCH(1, 1) of (3.9), because the past disturbances enter as squares the sign of the shock has no impact on the way shock is transmitted and affects the volatility of subsequent shocks. Black (1976) argued that by raising the 'leverage of a firm' (the firm's debt to equity ratio) negative shocks may increase the volatility of returns on the firm's equity. We will give only a brief review of a few non-linear GARCH models to present the key aspects of generating and evaluating forecasts of volatility made with such models.

The Glosten *et al*. (1993) GJR–GARCH model is perhaps the most intuitive non-linear generalization of (3.9). The coefficient on ε_{t-1}^2 depends on whether the shock is positive or negative, so in place of $\alpha\varepsilon_{t-1}^2$ we have $\alpha\varepsilon_{t-1}^2 1_{(\varepsilon_{t-1}<0)} + \gamma\varepsilon_{t-1}^2 1_{(\varepsilon_{t-1}>0)}$, giving:

$$h_t = \omega + \alpha\varepsilon_{t-1}^2(1 - 1_{(\varepsilon_{t-1}>0)}) + \gamma\varepsilon_{t-1}^2 1_{(\varepsilon_{t-1}>0)} + \beta h_{t-1}. \tag{3.16}$$

A 'leverage effect' is operative if $\alpha > \gamma$ and absent if $\alpha = \beta$, in which case the model reverts to the linear GARCH. A natural generalization of the 'abrupt switch' in 'regimes' from using the indicator function is instead to allow for a smooth transition, parallelling the discussion in Section 2.5.3 (in the context of modelling the conditional mean rather than the conditional variance). For example, the indicator function $1_{(\cdot)}$ in (3.16) can be replaced by:

$$G(\varepsilon_{t-1}) = \frac{1}{1 + \exp(-\theta\varepsilon_{t-1})}, \quad \theta > 0.$$

When $\varepsilon_{t-1} = 0$, $G(\varepsilon_{t-1}) = \frac{1}{2}$, and the coefficient on ε_{t-1}^2 in the equation for h_t is $(\alpha+\gamma)/2$. For $\varepsilon_{t-1} < 0$, the value of $G(\varepsilon_{t-1}) < \frac{1}{2}$, putting more weight on 'α'. For large θ the smooth transition becomes sharper and in the limit the function is the indicator function. Both the GJR–GARCH and smooth transition version could be generalized to allow the intercept ω and the coefficient of h_{t-1}, β, to switch between regimes (either abruptly, or with some degree of smoothness). These models apply the threshold switching ideas discussed in Section 2.5.3 to modelling h_t. One could equally suppose the regimes for h_t are determined by an unobservable Markov process (as assumed in modelling the conditional distributions of output growth in Section 2.5.4). For these extensions, including allowing for leverage effects, see Hamilton and Susmel (1994) and Hamilton and Lin (1996).

Forecasting the conditional variance using non-linear GARCH models

In Section 2.5.2 we showed that multi-step ahead forecasts from non-linear models of the conditional mean of a time series in general require numerical or simulation methods. As an example, we considered a two-step ahead forecast for a SETAR model. This required calculating the conditional expectation of the expression given in (2.58), repeated here for convenience:

$$g(y_{t+1}; \cdot) = [\phi^{\{1\}} + 1(y_{t+1|t} + \varepsilon_{t+1} > r)(\phi^{\{2\}} - \phi^{\{1\}})](y_{t+1|t} + \varepsilon_{t+1}).$$

The problem arises because $E_t[(1(y_{t+1|t} + \varepsilon_{t+1} > r))\varepsilon_{t+1}] \neq E_t[1(y_{t+1|t} + \varepsilon_{t+1} > r)] \times E_t(\varepsilon_{t+1})$ – the value of the indicator function is more likely to equal unity for positive values of ε_{t+1}. The similarity of threshold-type non-linear GARCH models, such as (3.16), to the SETAR models of conditional mean might suggest similar difficulties in obtaining multi-period forecasts. But this is not the case, at least for the threshold models. As for the standard GARCH, the 1-step ahead forecast of volatility from GJR–GARCH is given directly by the model as:

$$h_{t+1} = \omega + \alpha\varepsilon_t^2(1 - 1_{(\varepsilon_t>0)}) + \gamma\varepsilon_t^2 1_{(\varepsilon_t>0)} + \beta h_t.$$

The 2-step ahead forecast is defined by $h_{t+2|t} \equiv E(h_{t+2} \mid Y_t)$:

$$E(h_{t+2} \mid Y_t) = \omega + \alpha E_t(\varepsilon_{t+1}^2(1 - 1_{(\varepsilon_{t+1}>0)})) + \gamma E_t(\varepsilon_{t+1}^2 1_{(\varepsilon_{t+1}>0)}) + \beta E_t(h_{t+1}). \tag{3.17}$$

We can obtain a simple form for $h_{t+2|t}$ when $\{\varepsilon_t\}$ has a symmetric distribution, because in that case:

$$E_t(\varepsilon_{t+1}^2 1_{(\varepsilon_{t+1}>0)}) = E_t(\varepsilon_{t+1}^2)E_t(1_{(\varepsilon_{t+1}>0)}) = h_{t+1} \times \frac{1}{2}.$$

The expectation of the product is the product of the expectations because ε_{t+1}^2 and $1_{(\varepsilon_{t+1}>0)}$ are uncorrelated. When ε_{t+1} is symmetric about zero:

$$E_t(1_{(\varepsilon_{t+1}>0)}) = 1 \times \Pr(\varepsilon_{t+1} > 0) + 0 \times (1 - \Pr(\varepsilon_{t+1} > 0)) = \frac{1}{2}$$

given that $\Pr(\varepsilon_{t+1} > 0) = \frac{1}{2}$. Thus (3.17) becomes:

$$E(h_{t+2} \mid Y_t) = \omega + ((\alpha + \gamma)/2 + \beta)h_{t+1}.$$

More generally, a j-step forecast can be computed recursively from:

$$\begin{aligned} h_{t+j|t} &= \omega + \alpha h_{t+j-1|t}E_t(1 - 1_{(\varepsilon_{t+j-1}>0)}) + \gamma h_{t+j-1|t}E_t(1_{(\varepsilon_{t+j-1}>0)}) + \beta h_{t+j-1|t} \\ &= \omega + ((\alpha + \gamma)/2 + \beta)h_{t+j-1|t}, \end{aligned}$$

which by backward substitution gives:

$$\begin{aligned} h_{t+j|t} &= \omega + ((\alpha+\gamma)/2+\beta)h_{t+j-1|t} \\ &= \omega + ((\alpha+\gamma)/2+\beta)(\omega + ((\alpha+\gamma)/2+\beta)h_{t+j-2|t}) \\ &\vdots \\ &= \omega \sum_{i=0}^{j-2}((\alpha+\gamma)/2+\beta)^i + ((\alpha+\gamma)/2+\beta)^{j-1}h_{t+1}. \end{aligned}$$

Noting that $\omega \sum_{i=0}^{j-2}((\alpha+\gamma)/2+\beta)^i = \omega[1-((\alpha+\gamma)/2+\beta)^{j-1}][1-((\alpha+\gamma)/2+\beta)]^{-1}$, for $|((\alpha+\gamma)/2+\beta)| < 1$, and letting $\sigma^2 = \omega[1-((\alpha+\gamma)/2+\beta)]^{-1}$, we obtain an expression for $h_{t+j|t}$ from the GJR–GARCH model which is directly comparable to (3.15) for the linear GARCH model:

$$h_{t+j|t} - \sigma^2 = ((\alpha+\gamma)/2+\beta)^{j-1}(h_{t+1}-\sigma^2).$$

3.4.7 GARCH and forecasts of the conditional mean

In Section 3.2 we derived an expression for the conditional variance of an AR(1), $y_t = \phi y_{t-1} + \varepsilon_t$, allowing for dependence in $\{\varepsilon_t^2\}$, but not specifying the form this might take – see equation (3.2). One can write down expressions for general AR processes with general GARCH processes for the $\{\varepsilon_t\}$, although the key intuition is evident from an AR(1) with a GARCH(1, 1) for the disturbance. Because $V(y_{t+j} \mid Y_t) = V(e_{t+j|t} \mid Y_t)$, where $e_{t+j|t} = y_{t+j} - y_{t+j|t}$, expressions for the conditional variance of y_{t+j} are also informative about the effects of (G)ARCH on the variance of the conditional mean predictor about y_{t+j}, so could in principle be used to derive confidence intervals about $y_{t+j|t}$.

From (3.2) for the AR(1):

$$\begin{aligned} V(y_{t+j} \mid Y_t) &= \sum_{i=0}^{j-1} \phi^{2i} E(\varepsilon_{t+j-i}^2 \mid y_t) \\ &= \sum_{i=0}^{j-1} \phi^{2i} E(z_{t+j-i}^2 \mid y_t) h_{t+j-i|t} \\ &= \sum_{i=0}^{j-1} \phi^{2i} h_{t+j-i|t}. \end{aligned}$$

Substituting for $h_{t+s|t}$ from (3.15):

$$V(y_{t+j} \mid Y_t) = \sum_{i=0}^{j-1} \phi^{2i}(\sigma^2 + (\alpha+\beta)^{j-i-1}(h_{t+1} - \sigma^2))$$
$$= \sigma^2 \frac{1-\phi^{2j}}{1-\phi^2} + (h_{t+1} - \sigma^2) \sum_{i=0}^{j-1} \phi^{2i}(\alpha+\beta)^{j-i-1}.$$

The first term on the second line is the forecast variance for a process with i.i.d.$(0, \sigma^2)$ disturbances. This converges monotonically to $V(y_t) = \sigma^2/(1-\phi^2)$ as j gets large. The second term is due to the GARCH(1, 1) dependence in the disturbances. It can be positive or negative, depending on the sign of $|h_{t+1} - \sigma^2|$, and can be larger or smaller for j than $j-1$. Although $V(y_{t+j} \mid Y_t)$ converges to $V(y_t)$, it need not do so in a monotonic fashion because of the influence of the second term, so that, for example, there may be less uncertainty associated with medium-term forecasts compared with short-term forecasts.

3.5 Evaluation of volatility forecasts

We have derived expressions for multi-step forecasts of volatility from GARCH models. The next step is to evaluate these forecasts. It is natural to consider the bias and variance of these forecasts. Letting $e_{t+j|t}$ be the error in the j-step ahead forecast of the conditional variance at $t+j$, based on t (i.e., defining $e_{t+j|t}$ as for conditional mean prediction), we have:

$$e_{t+j|t} = h_{t+j} - h_{t+j|t}$$

putting to one side for the moment the fact that the actual conditional variance is unobserved. We assume that the GARCH model is correctly specified, so that $\varepsilon_t = z_t\sqrt{h_t}$, and $h_t = \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}$, with $z_t \sim N(0,1)$, say. It then follows from the definition of $h_{t+j|t}$ that the errors in the volatility forecasts are conditionally (and as a consequence unconditionally) unbiased, $E_t(e_{t+j|t}) = E_t(h_{t+j} - h_{t+j|t}) = 0$.

Substituting for h_{t+j} and $h_{t+j|t}$ we obtain:

$$e_{t+j|t} = \omega + \alpha\varepsilon_{t+j-1}^2 + \beta h_{t+j-1} - E_t(\omega + \alpha\varepsilon_{t+j-1}^2 + \beta h_{t+j-1})$$
$$= \beta e_{t+j-1|t} + \alpha(\varepsilon_{t+j-1}^2 - h_{t+j-1|t})$$

$$
\begin{aligned}
&= \beta e_{t+j-1|t} + \alpha(\varepsilon_{t+j-1}^2 - h_{t+j-1} + h_{t+j-1} - h_{t+j-1|t}) \\
&= \beta e_{t+j-1|t} + \alpha(v_{t+j-1} + e_{t+j-1|t}) \\
&= (\alpha + \beta) e_{t+j-1|t} + \alpha v_{t+j-1},
\end{aligned}
$$

where the fourth line uses $h_t = \varepsilon_t^2 - v_t$. Repeated backward substitution results in:

$$
\begin{aligned}
e_{t+j|t} &= (\alpha + \beta) e_{t+j-1|t} + \alpha v_{t+j-1} \\
&= \alpha v_{t+j-1} + (\alpha + \beta)[(\alpha + \beta) e_{t+j-2|t} + \alpha v_{t+j-2}] \\
&\vdots \\
&= \alpha \sum_{i=0}^{j-2} (\alpha + \beta)^i v_{t+j-1-i} + (\alpha + \beta)^{j-1} e_{t+1|t} \\
&= \alpha \sum_{i=0}^{j-2} (\alpha + \beta)^i v_{t+j-1-i}
\end{aligned}
$$

because $e_{t+1|t} = 0$. It is then a simple task to calculate the variance of the forecast error. Note $V(e_{t+j|t} \mid Y_t) = E(e_{t+j|t}^2 \mid Y_t)$, so that:

$$
E(e_{t+j|t}^2 \mid Y_t) = \alpha^2 \sum_{i=0}^{j-2} (\alpha + \beta)^{2i} E(v_{t+j-1-i}^2 \mid Y_t),
$$

where the cross-products are zero because the $\{v_t\}$ are serially uncorrelated. From $v_t = h_t(z_t^2 - 1)$:

$$
\begin{aligned}
E(v_{t+j-1-i}^2 \mid Y_t) &= E(z_{t+j-1-i}^2 - 1)^2 E(h_{t+j-1-i}^2 \mid Y_t) \\
&= (E(z_t^4) - 1) E(h_{t+j-1-i}^2 \mid Y_t)
\end{aligned}
$$

so:

$$
E(e_{t+j|t}^2 \mid Y_t) = \alpha^2 (E(z_t^4) - 1) \sum_{i=0}^{j-2} (\alpha + \beta)^{2i} E(h_{t+j-1-i}^2 \mid Y_t). \tag{3.18}
$$

In principle, if the model for volatility were correctly specified, then we could use (3.18) to construct confidence intervals for the volatility forecasts, and we would expect these intervals to contain the actual level

of volatility the number of times indicated by the nominal coverage rate. In practice, such a calculation is complicated by the fact that $h_{t+j} \mid Y_t$ is non-normal. Even given the mean and variance of the distribution, $h_{t+j|t}$ and $V(h_{t+j} \mid Y_t)(= E(e^2_{t+j|t} \mid Y_t))$, non-normality warns against calculating 95% confidence intervals as $h_{t+j|t} \pm 1.96\sqrt{E(e^2_{t+j|t} \mid Y_t)}$, for example. This would at best be an approximate 95% interval.

The main problem that arises in an attempt to evaluate the quality of volatility forecasts is that volatility is unobserved. This means that even if we were to calculate confidence intervals based on (3.18) for $h_{t+j|t}$, there is no readily available series for actual volatility to allow one to compare the actual coverage rate to the nominal. Furthermore, parallelling the point forecast evaluation literature, two ways of evaluating forecasts of volatility are 'realization-forecast' regressions (e.g., equation (2.2)) and comparisons based on MSFEs (e.g., equation (2.26)). But both require a 'proxy' or measure for actual volatility, h_{t+j}. For example, the realization-forecast regression takes the form:

$$h_{t+j} = \alpha + \beta h_{t+j|t} + v_{t+j}, \tag{3.19}$$

where, for example, $t = 1, \ldots, T$ and $j > 0$ is the forecast horizon, and comparisons of rival volatility forecasts are based on:

$$\frac{1}{T}\sum_{t=1}^{T}(h_{t+j} - h_{1,t+j|t})^2$$

and:

$$\frac{1}{T}\sum_{t=1}^{T}(h_{t+j} - h_{2,t+j|t})^2$$

for sequences of volatility forecasts $(h_{1,1+j|1}, \ldots, h_{1,T+j|T})$ and $(h_{2,1+j|1}, \ldots, h_{2,T+j|T})$. Suppose that $j = 1$. For both types of evaluation it is commonplace to proxy $\{h_{t+1}\}$ by $\{\varepsilon^2_{t+1}\}$, the squared returns (or squared shocks). The rationale for using the squared return is that, if the model is correctly specified (i.e., $\varepsilon_t = z_t\sqrt{h_t}$, with h_t correctly specified and z_t having its assumed distribution):

$$E_t(\varepsilon^2_{t+1}) = E_t(z^2_{t+1}h_{t+1}) = h_{t+1},$$

that is, the realized squared return is an unbiased proxy for actual volatility.

Replacing h_{t+1} by ε_{t+1}^2, we obtain the return-volatility forecast regression:

$$\varepsilon_{t+1}^2 = \alpha + \beta h_{t+1|t} + \nu_{t+1}. \tag{3.20}$$

Because $h_{t+1|t} = h_{t+1}$, and the expected value of the dependent variable is h_{t+1}, the population values of α and β are $\alpha = 0$ and $\beta = 1$, respectively. In practice the dependent variable $h_{t+1|t}$ will be generated from a model with estimated parameters (e.g., $\hat{\omega}$, $\hat{\alpha}$ and $\hat{\beta}$ in the case of the GARCH(1, 1)) and is likely to measure h_{t+1} with error, for example, $h_{t+1|t} = h_{t+1} + \upsilon_{t+1}$. Assuming that $E(\upsilon_{t+1}h_{t+1}) = E(\upsilon_{t+1}\nu_{t+1}) = 0$ as in the standard textbook treatment of measurement error in an explanatory variable, we obtain:

$$\hat{\beta} = 1 - \frac{V(\upsilon_t)}{V(\upsilon_t) + V(h_t)} \tag{3.21}$$

when $\beta = 1$. The ordinary least squares (OLS) estimator $\hat{\beta}$ will be downward biased to the extent given by the second term in (3.21). For this reason, tests based on the values of α and β may be misleading, and a number of papers have instead focused on the R^2 of this regression. The R^2 measures the proportion of the variability in squared returns explained by the volatility forecasts of the particular model under scrutiny. A number of studies have found low values of R^2, particularly for daily data, suggesting that standard GARCH volatility models account for little of the variability of *ex post* squared returns, and consequently that such models produce poor volatility forecasts that are of limited practical value.[8] Anderson and Bollerslev (1998) counter by showing that low R^2's are not symptomatic of inaccurate forecasts of volatility. The important point is that whilst squared returns are an unbiased proxy of actual volatility they will typically be a very noisy proxy, especially for high-frequency data, such as daily data. The variance of $\{\varepsilon_{t+1}^2\}$ is:

$$\begin{aligned} V(\varepsilon_{t+1}^2) &= E[(\varepsilon_{t+1}^2 - E(\varepsilon_{t+1}^2))^2] \\ &= E[h_{t+1}^2(z_{t+1}^2 - 1)^2] \\ &= h_{t+1}^2(\kappa - 1), \end{aligned}$$

where the second line uses $\varepsilon_{t+1}^2 = h_{t+1} z_{t+1}^2$ and $E(\varepsilon_{t+1}^2) = h_{t+1}$, and the third $E(z_{t+1}^2 - 1)^2 = E(z_{t+1}^4) - 2E(z_{t+1}^2) + 1 = \kappa - 1$ (recall $\{z_t\}$ is standard normal, and $\kappa \equiv E(z_{t+1}^4) = 3$). Thus the variance of the proxy is at least twice[9] the square of the actual volatility. The noisiness of the proxy implies a low R^2 even for forecasts from the conditional volatility model that actually generated the returns.

Anderson and Bollerslev (1998) illustrate by deriving the population value of the R^2 for a returns-volatility forecast regression when the GARCH(1, 1) volatility model is true. They obtain:

$$R^2 = \frac{\alpha^2}{(1 - \beta^2 - 2\alpha\beta)} \tag{3.22}$$

assuming the existence of a finite unconditional fourth moment. The latter requires that $\kappa\alpha^2 + \beta^2 + 2\alpha\beta < 1$, so that substituting $\alpha^2 < (1 - \beta^2 - 2\alpha\beta)/\kappa$ into the numerator of (3.22), we obtain the inequality that the population $R^2 < \kappa^{-1}$ for a GARCH(1, 1) with a finite unconditional fourth moment. So R^2 is bounded from above by $\frac{1}{3}$ when z_t is normally distributed, and an even tighter bound is operative when z_t is Student t.

For a GARCH(1, 1) model estimated on daily percentage returns of the Deutschemark-US dollar spot rate, 1 October 1987 to 30 September 1992, they obtain $\hat{\alpha} = 0.068$, and $\hat{\beta} = 0.898$. Plugging these values into the above formula gives a population R^2 of 0.064: this is the value of the R^2 that we would calculate from (3.20) (barring sampling variability) if the actual sample of daily percentage exchange rate returns were generated by a GARCH(1, 1) with α and β given by the values of the OLS estimates. The value of R^2 for the the regression given by (3.20) for GARCH model volatility forecasts of 1 October 1992 to 30 September 1993 was 0.047, similar in magnitude to the population value under correct specification, and to other estimates in the literature which have led to a loss of confidence in volatility forecasts.

These comments also apply to less formal comparisons of MSFEs of rival models where actual volatility is proxied by *ex post* returns. If one has a number of volatility models, including GARCH and non-linear GARCH models, the high values of the MSFEs for all the models might tend to hide any (small by comparison) differences in MSFEs between models, leading to the conclusion that none of the models offer a good description of the volatility in the series.

3.6 Recent developments in the evaluation of volatility forecasts

In the previous section we argued that assessing the quality of volatility forecasts by using *ex post* squared returns as a proxy for actual volatility might be misleading. *Ex post* squared returns are likely to be a very poor proxy for volatility, especially for high-frequency data. In this section we discuss a number of alternative proxies. We postpone until Chapters 4 and 5 an assessment of volatility forecasts based on the quality of derived interval and density forecasts, concentrating here on the volatility forecasts themselves. Christoffersen and Diebold (2000) consider whether volatility is forecastable, but their techniques are closely related to interval evaluation techniques and so will be discussed in Chapter 4.[10]

3.6.1 Realized volatility

Anderson and Bollerslev (1998) suggest replacing the unobserved volatility series h_{t+1} in (3.19) by an estimate based on *ex post* squared returns sampled more frequently than the underlying unit of observation on which the volatility forecasts are based. They take the underlying unit of observation to be a day, as in the daily exchange rate forecasts of the Deutschemark-US dollar example reported in the previous section. Thus, instead of replacing h_{t+1} by ε^2_{t+1}, the dependent variable in (3.19) becomes:

$$\sum_{j=1}^{m} \varepsilon^2_{t+1,j}$$

the sum of the corresponding squared intra-period returns, where $(\varepsilon^2_{t+1,1}, \ldots, \varepsilon^2_{t+1,m})$ are the m intraday squared returns for day $t+1$. When $m = 1$ there is one intraday observation, that is, one observation per day, and $\varepsilon^2_{t+1} \equiv \sum_{j=1}^{1} \varepsilon^2_{t+1,j}$. For $m = 1$ we reported Anderson and Bollerslev's finding of an $R^2 = 0.047$ for the Deutschemark-US dollar returns-volatility forecast regression. For $m = 24$, corresponding to proxying daily volatility by summing the squares of hourly returns, they report an $R^2 = 0.331$. For $m = 288$ (five-minute intervals), the R^2 is close to $\frac{1}{2}$ at 0.479. Estimates of volatility obtained in this way are often referred to as realized volatility, and when used in place of same-frequency squared returns give a very different picture regarding the predictability of the volatility process using standard GARCH models.

A major drawback is of course that data recorded at a higher frequency than the underlying unit of observation might not be available in any

given instance: only recently has it become possible to obtain intradaily data. Moreover, even when data are available, it is not clear how to choose m. Setting m as high as possible would appear to be warranted on the basis of the theoretical results reported by Anderson and Bollerslev, but such a strategy would neglect potentially distortionary effects from market microstructure effects and infrequent trading.

Andersen *et al*. (2003) set out a framework for modelling and forecasting realized volatility directly, rather than viewing realized volatility solely as a proxy for actual volatility for forecast evaluation purposes. They illustrate with estimates of daily volatility obtained by summing 30-minute returns for US $, Yen and DM exchange rates. Forecasts from VAR models of the daily realized volatilities are found to fare well compared to alternative forecasts (e.g., from GARCH models fitted to daily returns). Their evaluation techniques include comparing R^2 from realization-forecast regressions where the dependent variable is observed realized volatility. A number of authors, such as Blair *et al*. (2001), have shown that adding realized volatility as an explanatory variable in GARCH models may result in improved volatility forecasts, and Koopman *et al*. (2004) find that realized volatility models produce more accurate volatility forecasts than conditional variance models of the S&P 100 daily returns. The evidence suggests that the usefulness of realized volatility estimates goes beyond the evaluation of volatility forecasts. Time-series of daily realized volatility can be used to produce superior volatility forecasts, either by direct modelling, or as an input to standard GARCH models.

3.6.2 Intraday range

For daily data, an important source of intraday information is the daily high ($P_{t+1}^{\{\text{High}\}}$) and low ($P_{t+1}^{\{\text{Low}\}}$), from which the daily intraday range is simply $\{Rg_{t+1} = P_{t+1}^{\{\text{High}\}} - P_{t+1}^{\{\text{Low}\}}\}$. The intraday range has been suggested as a proxy for daily volatility, and has the advantage of often being readily available for daily equity price series. If we assume that the price follows a geometric Brownian motion, as in the recent finance literature, then the extreme value estimator of daily volatility due to Parkinson (1980) is given by:

$$\hat{h}_{t+1} = \frac{(\ln P_{t+1}^{\{\text{High}\}} - \ln P_{t+1}^{\{\text{Low}\}})^2}{4 \ln 2}.$$

See Parkinson (1980) for details, and Garman and Klass (1980).

Anderson and Bollerslev (1998, p. 898) reference a number of studies using the range, and report some calculations which indicate that the intraday range is on par with two or three hour intraday returns as an estimator of underlying volatility. Finally, Bollen and Inder (2002) suggest an approach to the estimation of daily realized volatility that allows for heteroskedasticity and time-varying autocorrelations in intraday returns. The Anderson and Bollerslev (1998) estimator of daily realized volatility that sums intraday squared returns requires that these returns are uncorrelated, but this may not be the case due to market microstructure effects. Bollen and Inder (2002) show that their estimator compares favourably with the other methods we have reviewed, such as extreme value estimators and summing squared returns. Bollen and Inder (2002) also include in their comparisons the 'simple' volatility estimator:

$$\sqrt{\widehat{h_{t+1}}} = \frac{|\varepsilon_{t+1}|}{\sqrt{2/\pi}}.$$

This estimator is unbiased by construction if returns are normally distributed, $\varepsilon_{t+1} \equiv \ln P_{t+1} - \ln P_t \sim N(0, h_{t+1})$, but inefficient in that only one observation is used in its construction.

3.6.3 Utility-based measures and trading rules

West *et al*. (1993) evaluate the forecast performance of standard models of the volatility of weekly exchange rates using a utility-based metric. Specifically, they assume an investor with a mean-variance utility function, and calculate the expected values of making investment decisions using the different volatility models. Maddala and Li (1996, section 5) critically discuss the literature on the use of trading rules to evaluate volatility models of stock market and exchange rate data.

3.7 Summary

This chapter considers models of volatility or conditional variance, and the evaluation of forecasts of the conditional variance. Modelling and forecasting the conditional variance of a process is potentially important and feasible when the second-moment is non-constant but varies in a systematic fashion that is amenable to modelling. A large number of financial time series and some macroeconomic time series exhibit time-varying conditional variances capable of being represented

by members of the autoregressive conditional heteroskedasticity class of models begun by Engle (1982).

We show that for asymmetric loss functions time-varying conditional variances will affect the optimal point forecasts, as well as the degree of uncertainty around the point forecast. With 'standard' squared-error loss functions, the optimal point forecast will be unaffected by the time-varying nature of the conditional variance of the process, but the degree of uncertainty will depend on when the forecast is made.

We describe the popular ARCH and GARCH models of conditional variance, and compare the forecasts from these models to those obtained by averaging past squared returns to bring out their distinctive features from a forecasting perspective. The evaluation of conditional variance forecasts is hampered by the absence of an 'observable' volatility series. We review the critiques of using squared returns as a proxy for volatility in traditional evaluation criteria, such as realization-forecast regressions and MSFE comparisons, and discuss the use of realized volatility and related measures.

4
Interval Forecasts

4.1 Introduction

In this chapter, we consider the evaluation of interval forecasts (also commonly referred to as prediction intervals). An interest in interval forecasts recognizes that the traditional emphasis on point estimates neglects any measure or assessment of the uncertainty surrounding the point forecast, or the 'confidence' that the forecaster has in the prediction. Point forecasts are sometimes provided with simple summary statistics about the forecaster's historical track record, such as *ex post* root mean squared errors calculated for past forecasts, as a tacit admission that in most practical settings the likely range of outcomes will influence the usefulness of the forecast. Unfortunately the magnitude or variability of past forecast errors may offer little guidance to the uncertainties attached to current forecasts, if the conditional variance of the process is changing over time, as in the volatility models of Chapter 3.

An interval forecast is a prediction about the range in which the outcome will occur (with a pre-assigned probability), and as such is a formal way of conveying *ex ante* forecast uncertainty. In the next chapter, we will consider forecast densities or probability distributions which provide a complete description of the probabilities that the forecaster attaches to all possible values or ranges of values of the outcome variable.

In addition to conveying uncertainty around a point forecast, agents involved in financial markets place a great deal of emphasis on interval forecasts, as they form the basis of the popularly employed Value-at-Risk (VaR) analysis.[1] For horizons in excess of a trading day, the long-run solvency of the institution is the appropriate risk to be hedged. By contrast, intraday horizons are important to traders seeking to manage

the risk of the trading desk. An empirical example is provided of the evaluation of intraday interval forecasts.

Section 4.2 considers the calculation of interval forecasts for linear autoregressions with independent disturbances. The assumption of independence rules out autoregressive conditional heteroskedasticity (ARCH)-effects, and would appear to be unduly restrictive. Nevertheless, many of the considerations that arise are equally pertinent when there is ARCH. We run through the Box–Jenkins approach and bootstrap techniques. We describe a setting where we have data from $t = 1$ to T, and wish to construct an interval forecast for period $T + 1$ (1-step) or $T + j, j > 1$ more generally. We look at whether the actual coverage is close to the nominal coverage for intervals calculated in various ways, and report on some of the results of a Monte Carlo study undertaken by Clements and Taylor (2001). A bootstrap procedure for ARCH processes due to Pascual *et al*. (2000) is briefly reviewed.

In Section 4.3, we define desirable properties of interval forecasts when explicit recognition is made that the conditional variance of the series may be changing over time. A close match between actual and nominal coverage rates is then necessary but not sufficient. To capture the potential importance of ARCH-type effects, the evaluation focuses on sequences of interval forecasts for periods during which the conditional variability of the series is changing. Section 4.4 outlines a procedure for testing whether a sequence of interval forecasts are conditionally efficient, as defined in Sections 4.3, and 4.5 considers regression-based tests of conditional efficiency, drawing on Christoffersen (1998), Engle and Manganelli (1999) and Clements and Taylor (2003). Section 4.6 presents a suggestion by Granger *et al*. (1989) for calculating interval forecasts when there is ARCH. The empirical illustration in Section 4.7 explores the usefulness of some of the tests of interval forecasts for hourly returns data, based on FTSE100 index futures. This example also illustrates the Generalized ARCH (GARCH) models discussed in Chapter 3.

4.2 Calculating interval forecasts

In this section, we follow the mainstream literature by considering the calculation of interval forecasts for a linear AR(p). The specification of the model is assumed to be known, but we allow that the parameters of the model are unknown. In practice, 'model uncertainty' might be at least as important as parameter estimation uncertainty: see, for example, Chatfield (1993, 1995) and Draper (1995).

The Box–Jenkins method (Box and Jenkins 1976, henceforth BJ) is the simplest approach, and will give interval forecasts with the exact coverage rates in certain circumstances. Consider the AR(p):

$$y_t = \delta + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{4.1}$$

where the $\{\varepsilon_t\}$ sequence is i.i.d. with distribution function F_ε, $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma_\varepsilon^2 < \infty$. If $\{\delta, \phi_1, \ldots, \phi_p, \sigma_\varepsilon^2\}$ are known, and F_ε is the normal distribution function, then the 1-step forecast density at period $T+1$ given information available up to period T is $Y_{T+1} \mid I_T = N(y_{T+1|T}, \sigma_\varepsilon^2)$, where $I_T = (y_T, \ldots, y_{T-p+1})$ and $y_{T+1|T} = E(Y_{T+1} \mid I_T) = \delta + \phi_1 y_T + \cdots + \phi_p y_{T-p+1}$. The standard BJ interval forecast with a nominal coverage rate of $(1-\alpha) \times 100\%$ is given by:

$$\{y_{T+1|T} + z_{\alpha/2}\sigma_\varepsilon, \quad y_{T+1|T} + z_{1-(\alpha/2)}\sigma_\varepsilon\}, \tag{4.2}$$

where z_γ is the γ quantile of the standard normal, that is, $\gamma = \Phi(z_\gamma)$, where Φ denotes the standard normal distribution function. Typical values of α might be 0.05, 0.1 and 0.2, to give intervals with coverage rates of 95, 90 and 80%, respectively. Notice that $z_{1-(\alpha/2)} = -z_{\alpha/2}$ so that we can write the interval as:

$$\{y_{T+1|T} \pm z_{\alpha/2}\sigma_\varepsilon\}.$$

By construction:

$$\Pr(Y_{T+1} \in \{y_{T+1|T} \pm z_{\alpha/2}\sigma_\varepsilon\}) = 1 - \alpha$$

because:

$$\Pr\left(z_{\alpha/2} < \frac{Y_{T+1} - y_{T+1|T}}{\sigma_\varepsilon} < z_{1-(\alpha/2)}\right) = \Pr\left(z_{\alpha/2} < \frac{\varepsilon_{T+1}}{\sigma_\varepsilon} < z_{1-(\alpha/2)}\right)$$
$$= 1 - 2\Phi(z_{\alpha/2}) = 1 - \alpha.$$

Multi-step intervals can be derived similarly, so that a k-step ahead interval is given by:

$$\{y_{T+k|T} \pm z_{\alpha/2}\sigma_{\varepsilon,k}\},$$

where $y_{T+k|T} = E(Y_{T+k} \mid I_T)$, and $\sigma_{\varepsilon,k} = \sqrt{V(Y_{T+k} \mid I_T)}$.

We have assumed that the parameters are known, and that disturbances have a normal distribution. The BJ approach can be used when neither of these assumptions hold. For example, the unknown parameters can be estimated, and the population values of $\{\delta, \phi_1, \ldots, \phi_p, \sigma_\varepsilon^2\}$ used to construct $y_{T+1|T}$ and σ_ε can simply be replaced by the estimates. Ignoring the estimation uncertainty – by treating random variables as if they were the population values – is likely to result in intervals that are 'too narrow' with actual coverage less than the nominal. Similarly, the normality of the disturbances can be assumed to construct the intervals, but is likely to lead to poor intervals to the extent that the normal distribution provides a poor approximation to the true distribution.

One method of improving on BJ is to estimate the density of the prediction errors by bootstrap methods (BS), for example, Findley (1986), Stine (1987) and Masarotto (1994). A second approach is to directly estimate the distribution of Y_{T+k} conditional on $Y_T, \ldots, Y_{T-p+1}$, circumventing the need to calculate the prediction error distribution. In this tradition, Thombs and Schucany (1990) propose a BS method that uses bootstrap replicates generated backward in time to ensure that the probability distribution for future values of the process is conditional on the history of the process actually observed: see also Breidt *et al.* (1995), McCullough (1994) and Kim (2001, 2003). Pascual *et al.* (2001) propose a method that does not require re-sampling through the backward representation of the process, and Clements and Taylor (2001) ally to this the bootstrap bias-correction of Kilian (1998b). The rest of this section reviews these developments.

4.2.1 Bootstrap the forecasts

A standard bootstrap procedure is as follows. Estimate (4.1) for $t = p+1, \ldots, T$, to give $\{\hat{\delta}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\sigma}_\varepsilon^2\}$ and residuals $\{\hat{\varepsilon}_t\}$. Denote the empirical distribution function (EDF) of the residuals by $F_{\hat{\varepsilon}}$.[2] Obtain $\{\varepsilon^*_{b,t}\}$, for $t = T+1, \ldots, T+k$, by sampling from $F_{\hat{\varepsilon}}$, and then generate a bootstrap sample recursively from:

$$y^*_{b,t} = \hat{\delta} + \hat{\phi}_1 y^*_{b,t-1} + \cdots + \hat{\phi}_p y^*_{b,t-p} + \varepsilon^*_{b,t} \tag{4.3}$$

for $t = T+1, \ldots, T+k$ and where $y^*_{b,T+s} = y_{T+s}$ for $s \leq 0$. Repeat B times, so we have a bootstrap resample $\{\mathbf{y}^*_b\}$ for each $b = 1, \ldots, B$ (where $\mathbf{y}^*_b = (y^*_{b,T+1}, \ldots, y^*_{b,T+k})'$). The EDF of $\{\mathbf{y}^*_b\}$ is the bootstrap estimate of the unknown forecast distribution conditional on the parameter estimates $\{\hat{\delta}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\sigma}_\varepsilon^2\}$ and the observed values $\{y_T, y_{T-1}, \ldots, y_{T-p+1}\}$.

The Efron percentile method uses the quantiles of the empirical distribution to define the upper and lower critical points of the interval forecast. That is, if we let F_k^* be the EDF of the $\{y_{b,T+k}^*, b = 1, \ldots, B\}$, then formally the interval is given by:

$$[F_k^{*-1}(\alpha/2), F_k^{*-1}(1-\alpha/2)], \tag{4.4}$$

where for an interval with 90% coverage and $B = 100$ replications, $\alpha = 0.1$, and $F_k^{*-1}(0.05) = L_k = y_{T+k}^{*(5)}$, $F_k^{*-1}(0.95) = U_k = y_{T+k}^{*(95)}$. That is, U_k and L_k are set to the given 5th and 95th largest values of $\{y_{b,T+k}^*, b = 1, \ldots, B\}$. Hall (1988, p. 933, 937–8) suggests an alternative approach, often referred to as 'Hall's percentile interval', which is used in the Monte Carlo of Section 4.2.5. This interval is given by:

$$[\hat{y}_{T+k} - t_{1-\alpha/2},\ \hat{y}_{T+k} - t_{\alpha/2}], \tag{4.5}$$

where $\hat{y}_{T+k}$ is the k-step ahead forecast based on the sample parameter estimates and $y_T, \ldots, y_{T-p+1}$, that is:

$$\hat{y}_{T+k} = \hat{\delta} + \hat{\phi}_1 \hat{y}_{T+k-1} + \cdots + \hat{\phi}_p \hat{y}_{T+k-p}, \tag{4.6}$$

where $\hat{y}_{T+s} = y_{T+s}$ for $s \leq 0$, and $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are the quantiles of the bootstrap distribution of $Y_{T+k}^* - \hat{y}_{T+k}$.[3] Clements and Taylor (2001) discuss these and other bootstrap intervals, and see also Davison and Hinkley (1997) and Kilian (1998a).

4.2.2 Allowing estimation uncertainty

The bootstrap explained above dispenses with the assumption of normality but ignores the variability in the parameter estimates – estimates are treated as if they were the population values. To allow for estimation uncertainty, a bootstrap sample of T observations is generated by sampling from $F_{\hat{\varepsilon}}$ to obtain $\{\varepsilon_{b,t}^*\}$ for $t = p+1, \ldots, T$, and by recursive calculation of:

$$y_{b,t}^* = \hat{\delta} + \hat{\phi}_1 \hat{y}_{b,t-1}^* + \cdots + \hat{\phi}_p y_{b,t-p}^* + \varepsilon_{b,t}^* \tag{4.7}$$

for $t = p+1$ to T given the observed initial values $\{y_1, \ldots, y_p\}$ (so $y_{b,t}^* = y_t$ for $t = 1, \ldots, p$). This gives a sample $\{y_{b,p+1}^*, \ldots, y_{b,T}^*\}$. Create B such samples. On each of these estimate the model (4.1) to give parameter

estimates $\boldsymbol{\theta}_b^{*\prime} = (\hat{\delta}_b, \hat{\phi}_{1,b}, \ldots, \hat{\phi}_{p,b})$, and then solve recursively:

$$y_{b,t}^* = \hat{\delta}_b + \hat{\phi}_{1,b} y_{b,t-1}^* + \cdots + \hat{\phi}_{p,b} y_{b,t-p}^* + \varepsilon_{b,t}^* \quad t = T+1, \ldots, T+k, \quad (4.8)$$

where as before $\{\varepsilon_{b,T+1}^*, \ldots, \varepsilon_{b,T+k}^*\}$ are drawn from $F_{\hat{\varepsilon}}$. The empirical distribution of $\{\mathbf{y}_b^*\}$ (for a given k) is the bootstrap estimate of the unknown k-step ahead forecast distribution. The effect of estimation uncertainty has now been incorporated, because on each of the bootstrap replicates the 'past' data is re-drawn, and the forecasts are calculated from a model with parameters estimated on the simulated data. However, the prediction distributions no longer condition on the observed realization of the process $\{y_{T-p+1}, \ldots, y_T\}$ (but only on $y_1, \ldots, y_p$).

4.2.3 Conditional intervals and estimation uncertainty

In some cases, such as calculating confidence intervals for impulse responses, unconditional distributions are of interest, but an interval forecast is a probability statement for the future values of the process given the history of the process that was actually observed. Thombs and Schucany (1990) suggest a way of obtaining forecast intervals (that allow for estimation uncertainty) based on backcasting, but a simpler proposal of Pascual *et al.* (2001) is to use (4.8) but with:

$$y_{b,T+s}^* = y_{T+s} \quad \text{for } s \leq 0, \quad (4.9)$$

so that $\{y_{b,T-p+1}^*, \ldots, y_{b,T}^*\}$ are replaced by $\{y_{T-p+1}, \ldots, y_T\}$ for all b.

4.2.4 Bias-correcting the parameter estimates

The bootstrap replicates are generated from the original parameter estimates $\hat{\boldsymbol{\theta}}' = (\hat{\delta}, \hat{\phi}_1, \ldots, \hat{\phi}_p)$. However, when the sample size is small and there are roots close to the non-stationary region, the estimators will be subject to small-sample biases, which might result in intervals with poor coverage. Let $\boldsymbol{\theta}' = (\delta, \phi_1, \ldots, \phi_p)$, and let $\boldsymbol{\Psi}$ denote the bias in estimating $\boldsymbol{\theta}$, $\boldsymbol{\Psi} = E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. An estimate of this term can be obtained by a bootstrap. That is, calculate $\hat{\boldsymbol{\Psi}} = \bar{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}$, where $\bar{\boldsymbol{\theta}}^* = B^{-1} \sum_{b=1}^{B} \boldsymbol{\theta}_b^*$, and the $\boldsymbol{\theta}_b^*$ are the estimates of the parameters of (4.1) for each of the B bootstrap replicates generated using $\hat{\boldsymbol{\theta}}$ and $F_{\hat{\varepsilon}}$.

For models with roots close to the unit circle, care is required in ensuring that the bias-correction does not change the number of unit roots in $|\hat{\phi}(z)| = 0$ (where $\hat{\phi}(L) = 1 - \hat{\phi}_1 L - \cdots - \hat{\phi}_p L^p$). This is considered below: for now suppose $\hat{\boldsymbol{\theta}}$ falls within the stationarity region, as does the

bias-corrected estimate $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\Psi}}$. Then $\tilde{\boldsymbol{\theta}}$ simply replaces $\hat{\boldsymbol{\theta}}$ in (4.7) to generate the B bootstrap samples. The parameter estimates obtained on these samples denoted by θ_b^* are then bias-corrected using $\hat{\boldsymbol{\Psi}}$, $\tilde{\boldsymbol{\theta}}_b^* = \boldsymbol{\theta}_b^* - \hat{\boldsymbol{\Psi}}$. In principle, one could estimate specific bias-correction factors for each $\boldsymbol{\theta}_b^*$.

In performing the bias-corrections, we wish to ensure that the correction does not cause stationary forecasts to become non-stationary. Kilian (1998b) suggests the following. Let $m(\hat{\boldsymbol{\theta}})$ denote the modulus of the largest root of the characteristic equation formed using the autoregressive parameters in $\hat{\boldsymbol{\theta}}$. If $m(\hat{\boldsymbol{\theta}}) \geq 1$ (non-stationary), then $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ (and similarly, $\tilde{\boldsymbol{\theta}}_b^* = \boldsymbol{\theta}_b^*$ if $m(\boldsymbol{\theta}_b^*) \geq 1$) so the coefficients $\hat{\boldsymbol{\theta}}$ (or $\boldsymbol{\theta}_b^*$) are not adjusted when the roots of the associated characteristic equations are non-stationary. Thus, the bias-adjustment scheme does not rule out non-stationary forecasts, but merely prevents these from occurring as a result of the bias-adjustment scheme. If $m(\hat{\boldsymbol{\theta}}) < 1$, construct $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\Psi}}$. If $m(\tilde{\boldsymbol{\theta}}) \geq 1$, let $\hat{\boldsymbol{\Psi}}_1 = \hat{\boldsymbol{\Psi}}$, set $\tau_1 = 1$ and define $\hat{\boldsymbol{\Psi}}_{i+1} = \tau_i \hat{\boldsymbol{\Psi}}_i$ and $\tau_{i+1} = \tau_i - 0.01$, for $i = 1, 2, \ldots$. Then iterate on $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\Psi}}_i$, $i = 1, 2, \ldots$ until $m(\hat{\boldsymbol{\theta}}_i) < 1$ and $m(\hat{\boldsymbol{\theta}}_{i-1}) > 1$. Setting $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_i$ imposes the largest bias-correction possible subject to the largest root remaining stationary. The $\boldsymbol{\theta}_b^*$ are similarly treated, to give $\tilde{\boldsymbol{\theta}}_b^*$. This scheme lacks a theoretical basis but seems reasonable for stationary processes.

Having obtained the bias-corrected $\tilde{\boldsymbol{\theta}}_b^*$, we generate $y_{b,T+k}^*$, $k = 1, \ldots, K$ recursively by plugging $\tilde{\boldsymbol{\theta}}_b^*$ into (4.8) and making random draws from $F_{\hat{\varepsilon}}$, but conditioning on the 'observed' data. Thus we have a sample of B sets of future values of the series, $\{y_{b,T+k}^*, \ldots, y_{b,T+k}^*\}$. For each k, we calculate intervals with the desired coverage.

4.2.5 Monte Carlo evaluation: step-by-step guide

In summary, a Monte Carlo evaluation of the bootstrapped interval forecasts calculated with bias-corrected parameter estimates proceeds as follows:

Step 1. Simulate a series $\{y_1, \ldots, y_T\}$ of length T from a data generation process (DGP) (such as (4.1)) with an error distribution F_ε. Simulate R continuations of the series of length K, the maximum forecast horizon. Use the true parameter values, draw from the true error distribution, F_ε, and condition on $\{y_{T-p+1}, \ldots, y_T\}$ (for a p-order process). These continuations are 'possible future realizations', and will be used to estimate the coverage of the intervals.

Step 2. Estimate $\hat{\boldsymbol{\theta}}$ and $F_{\hat{\varepsilon}}$ on $\{y_1, \ldots, y_T\}$. Perform a bootstrap to bias-correct $\hat{\boldsymbol{\theta}}$: this gives $\tilde{\boldsymbol{\theta}}$.

Step 3. Simulate B bootstrap replicates of length T using $\tilde{\boldsymbol{\theta}}$ and drawings from $F_{\hat{\varepsilon}}$. On each of these, estimate (4.1) to give $\boldsymbol{\theta}_b^*$. Use the estimate of the bias from Step 2 to bias-correct these estimates to give $\tilde{\boldsymbol{\theta}}_b^*$.

Step 4. Conditional on $\{y_{T-p+1}, \ldots, y_T\}$, for each $\tilde{\boldsymbol{\theta}}_b^*$ (and drawings from $F_{\hat{\varepsilon}}$), generate a continuation of the series. This gives $\{y_{1,T+k}^*, \ldots, y_{B,T+k}^*\}$, $k = 1, \ldots, K$. Let L_k^* and U_k^* denote the interval endpoints. Then, $\lambda_k^* = U_k^* - L_k^*$ is the length of the interval, and the coverage $\beta_k^* = R^{-1} \sum_{r=1}^{R} 1(L_k^* \leq y_{T+k}^r \leq U_k^*)$, where r indexes the rth element of the R continuations simulated at Step 1. We also record the proportion above and below the interval: $\beta_{k,a}^* = R^{-1} \sum_{r=1}^{R} 1(y_{T+k}^r > U_k^*)$ and $\beta_{k,b}^* = R^{-1} \sum_{r=1}^{R} 1(y_{T+k}^r < L_k^*)$.

Step 5. Repeat Steps 1–4 M times, and index the values obtained on each iteration by m, $m = 1, \ldots, M$.

Step 6. Calculate the Monte Carlo estimates of length and coverage, and the variability of these estimates, as:

$$\begin{aligned}
\bar{\beta}_k^* &= \frac{1}{M} \sum_{m=1}^{M} \beta_{k,m}^* \\
\mathrm{SE}(\bar{\beta}_k^*) &= \sqrt{\frac{1}{M}} \sqrt{\sum_{m=1}^{M} (\beta_{k,m}^* - \bar{\beta}_k^*)^2 / (M-1)} \\
\bar{\lambda}_k^* &= \frac{1}{M} \sum_{m=1}^{M} \lambda_{k,m}^* \\
\mathrm{SE}(\bar{\lambda}_k^*) &= \sqrt{\frac{1}{M}} \sqrt{\sum_{m=1}^{M} (\lambda_{k,m}^* - \bar{\lambda}_k^*)^2 / (M-1)}
\end{aligned}$$

and $\bar{\beta}_{k,a}^*$ and $\bar{\beta}_{k,b}^*$ are calculated in the obvious way.

The data generating process is taken to be an AR(2) model with normal errors.[4]

$$y_t = 1.75 y_{t-1} - 0.76 y_{t-2} + \varepsilon_t.$$

The AR parameters were chosen to give roots close to the stationarity boundary to mirror the properties of actual economic time series. We consider two estimation sizes $T = \{25, 50\}$, a maximum forecast horizon $K = 10$ and a nominal coverage level ($c = 1 - \alpha$) of 95%.

The results for the AR(2) DGP with normal disturbances and a $c = 95\%$ coverage level are given in Table 4.1. The actual coverage levels of the

Table 4.1 AR(2) simulation results (normal errors and $c = 95\%$)

	k	$\bar{\lambda}_k^*$	SE(λ_k^*)	$\bar{\beta}_k^*$	SE(β_k^*)	$\bar{\beta}_{k,b}^*$	$\bar{\beta}_{k,a}^*$
$T = 25$							
BJ	1	3.87	0.02	91.49	0.20	4.24	4.27
BS		5.26	0.06	92.70	0.24	3.80	3.49
BJ	3	10.98	0.07	84.81	0.43	7.61	7.58
BS		18.36	0.26	89.34	0.44	5.60	5.06
BJ	5	16.65	0.15	77.90	0.60	11.10	11.00
BS		31.79	0.48	86.38	0.58	7.27	6.35
BJ	10	23.96	0.33	65.66	0.73	17.17	17.17
BS		62.77	1.13	83.08	0.74	8.89	8.03
$T = 50$							
BJ	1	3.89	0.01	93.35	0.11	3.27	3.38
BS		4.40	0.02	94.25	0.13	2.92	2.83
BJ	10	28.24	0.25	79.02	0.53	10.21	10.76
BS		53.33	0.64	90.22	0.48	5.04	4.73

Notes: This table presents the results of a Monte Carlo comparison of two methods of constructing prediction intervals – BJ and BS methods. The mean length of the interval ($\bar{\lambda}_k^*$), its associated standard error, the mean coverage level ($\bar{\beta}_k^*$), its associated standard error, and the mean proportions below ($\bar{\beta}_{k,b}^*$) and above ($\bar{\beta}_{k,a}^*$) are reported. Coverage rates are given as percentages. For the initial bootstrap (Step 2) and the bootstrap at Step 3 we carry out $B = 499$ replications. The actual coverage is calculated from $R = 1000$ continuations, and the Monte Carlo evaluation is based on $M = 1000$ replications. We use 'Hall's percentile' method to construct the Bootstrap (BS) intervals.

standard BJ intervals fall increasingly below the nominal for both sample sizes as the horizon lengthens, matching Thombs and Schucany (1990, table 1, p. 491) for example. A feature of the results is how much closer actual coverages are to the nominal for the bootstrapped intervals (BS) with bias-correction.

4.2.6 Bootstrapping ARCH processes

Pascual *et al.* (2000) consider the GARCH(1, 1) model:

$$\varepsilon_t = z_t \sqrt{h_t},$$
$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1},$$

where z_t is i.i.d. (0,1), and in the absence of any conditional mean dynamics, $\{\varepsilon_t\}$ is the return. As we noted in Section 3.4.7, the non-normality of the conditional returns distribution at more than 1-step ahead, even when $z_t \sim N(0,1)$, cautions against the use of the BJ interval given by (4.2). Pascual *et al.* (2000) suggest a way of bootstrapping both the conditional distribution of returns $\{\varepsilon_{T+j|T}\}$ and the conditional distribution of the volatility of returns $\{h_{T+j|T}\}$, and they establish the asymptotic validity of their procedures and investigate the small-sample performance via Monte Carlo. The bootstrap is similar in spirit to that for the linear AR model set out above. The bootstrap consists of the following steps.

Step 1. Given $\{y_1, \ldots, y_T\}$ from the GARCH(1, 1) DGP, and given that we wish to obtain the conditional distributions of ε_{T+j} and h_{T+j}, we begin by estimating a GARCH (1,1) by QML (see Section 3.4.2) to give $(\hat{\omega}, \hat{\alpha}, \hat{\beta})$ and:

$$\hat{h}_t = \hat{\omega} + \hat{\alpha}\varepsilon_{t-1}^2 + \hat{\beta}\hat{h}_{t-1}, \quad t = 2, \ldots, T.$$

The residuals are given by $\hat{z}_t = \hat{\varepsilon}_t / \sqrt{\hat{h}_t}$, and we note their empirical distribution function by $F_{\hat{z}}$.

Step 2. In order to incorporate the effects of parameter estimation uncertainty, simulate a bootstrap replicate of length T using $(\hat{\omega}, \hat{\alpha}, \hat{\beta})$ and drawings from $F_{\hat{z}}$:

$$\hat{h}_t^* = \hat{\omega} + \hat{\alpha}\varepsilon_{t-1}^{*2} + \hat{\beta}\hat{h}^*{}_{t-1}$$
$$\varepsilon_t^* = z_t^* \sqrt{\hat{h}_t^*}, \quad t = 2, \ldots, T.$$

$\hat{h}_1^* = \hat{h}_1$ and the $\{z_t^*\}$ are the draws from $F_{\hat{z}}$.

Step 3. Estimate the GARCH(1, 1) on this bootstrap replicate to give $(\hat{\omega}^*, \hat{\alpha}^*, \hat{\beta}^*)$ and then generate forecasts of future values using the recursions:

$$\hat{h}_{T+k}^* = \hat{\omega}^* + \hat{\alpha}^* \varepsilon_{T+k-1}^{*2} + \hat{\beta}^* \hat{h}^*{}_{T+k-1}$$
$$\varepsilon_{T+k}^* = z_{T+k}^* \sqrt{\hat{h}_{T+k}^*}, \quad k = 1, 2, \ldots,$$

$\varepsilon_T^* = \varepsilon_T$, and:

$$\hat{h}_T^* = \frac{\hat{\omega}^*}{1-\hat{\alpha}^*-\hat{\beta}^*} + \hat{\alpha}^* \sum_{j=0}^{T-2} \hat{\beta}^{*j} \left(\varepsilon_{T-j-1}^2 - \frac{\hat{\omega}^*}{1-\hat{\alpha}^*-\hat{\beta}^*} \right)$$

(by backward substitution on (3.11), for example). The $\hat{h}_{T+1}^*$ only vary across bootstrap replications because of the parameter estimation uncertainty – this is a consequence of next period's volatility being known given the GARCH model.

We run through the above steps B times to obtain the bootstrap distribution functions $\{\varepsilon_{b,T+k}^*, b = 1, \ldots, B\}$ and $\{\hat{h}_{b,T+k}^*, b = 1, \ldots, B\}$ from which the prediction intervals can be calculated. The Monte Carlo study reported by Pascual *et al*. (2000) indicates that when the $\{z_t\}$ are non-normal the BJ intervals can be poor, while the bootstrapped intervals perform well. Further, provided $\{z_t\}$ is symmetric allowing for parameter estimation uncertainty (as above) is not important for the returns intervals, but does matter for the conditional volatility intervals.

4.3 Desirable properties of interval forecasts

An interval forecast can be written as the triple $\{L_{t|t-1}, U_{t|t-1}, p\}$, where L denotes the lower level, and U the upper level of the interval and p is the nominal coverage rate. The subscripts attached to L and U indicate a 1-step forecast of period t, made at $t-1$. The nominal coverage p is to be interpreted as follows. If we were to observe n realizations of the process in period t, then the expected proportion lying in the interval $(L_{t|t-1}, U_{t|t-1})$ is equal to p, that is:

$$\Pr[Y_t \in (L_{t|t-1}(p), U_{t|t-1}(p))] = p, \tag{4.10}$$

where Y_t denotes the random variable. Only a single realization is observed for each t,[5] but as (4.10) holds for each t, we can compare the nominal coverage rate p with the actual coverage rate derived from a sequence of intervals and realizations for $t = 1, \ldots, n$.

From the sequence of intervals and realizations, a 'hit' sequence is derived as:

$$I_t = \begin{cases} 1, & \text{if } y_t \in (L_{t|t-1}(p), U_{t|t-1}(p)), \\ 0, & \text{otherwise,} \end{cases} \tag{4.11}$$

where each interval forecast is classified as a success (contains the realization) or a failure (realization falls outside). If the interval forecasts are 'well specified', in the sense that the probability that I_t takes the value 1 is equal to the nominal coverage rate p, then $E(I_t) = (1 \times p) + [0 \times (1 - p)] = p$.[6] We can estimate $E(I_t)$ from the sample mean of $\{I_t\}_{t=1}^{n}$, $(1/n)\sum_{t=1}^{n} I_t$, and testing whether this is significantly different from p is a test of 'unconditional' coverage, set out formally below. This is a test of whether there are the right number of 'hits' and 'misses' (or '1's and '0's) with no regard to whether there are discernible patterns in these occurrences.

One could imagine a situation where misses are clustered together. For example, if the process being forecast is subject to ARCH-type volatility clustering, then unless the interval forecasts are wider in turbulent periods compared to relatively tranquil times there are likely to be too many misses in the volatile periods. Also, the chances of I_t equalling zero (indicating a miss) should not be related to variables known at $t-1$, in addition to not depending on forecastable changes in volatility. These ideas underpin the notion of conditional efficiency. Christoffersen (1998) defines a set of *ex ante* interval forecasts as being efficient with respect to the information set (denoted Ω_{t-1}) if the conditional expectation of I_t equals p, that is $E(I_t|\Omega_{t-1}) = p$. If one restricts the information set to past values of the indicator function, $\Omega_t = \{I_t, I_{t-1,} \ldots\}$, then this is equivalent to saying that $\{I_t\}$ is i.i.d Bernoulli with parameter p.

Restricting the information set to lagged values of the indicator variable gives rise to the tests of conditional efficiency discussed in the next section.

4.4 Tests for conditional efficiency

Conditional efficiency requires the sequence $\{I_t\}$ is Bernouilli (p) and i.i.d. Below we consider the first part of this joint hypothesis. In Section 4.4.2 we consider the second part.

4.4.1 Unbiasedness

The test for correct unconditional coverage is sometimes termed a test for unbiasedness. It is a test of whether the actual coverage equals the nominal coverage (ignoring possible patterns in the hits and misses). The null is $E(I_t) = p$ versus $E(I_t) \neq p$. For a hit probability of π, the likelihood of the data is:[7]

$$L(\pi; I_1, I_2, \ldots, I_n) = (1 - \pi)^{n_0}\pi^{n_1}, \quad \pi \in \Pi = [0, 1], \tag{4.12}$$

where $n_1 = \sum_{j=1}^{n} I_j$ and $n_0 = n - n_1$. The likelihood ratio statistic is:

$$LR = -2 \ln \left(\frac{L(p; I_1, I_2, \ldots, I_n)}{L(\hat{\pi}; I_1, I_2, \ldots, I_n)} \right)$$
$$= -2 \left[n_0 \ln \left(\frac{1-p}{1-\hat{\pi}} \right) + n_1 \ln \frac{p}{\hat{\pi}} \right] \sim \chi_1^2$$

that is, the ratio of the likelihood under the null hypothesis to the likelihood evaluated under the maximum likelihood estimate (MLE) $\hat{\pi}$ of $\pi \in \Pi$. $\hat{\pi}$ solves:

$$\frac{\partial L}{\partial \pi} = n_1 \pi^{(n_1 - 1)} (1-\pi)^{n_0} - n_0 (1-\pi)^{(n_0 - 1)} \pi^{n_1} = 0$$

to give $\hat{\pi} = n_1/n$, the sample proportion of hits. Tests for correct unconditional coverage, or bias, can also be found in Granger *et al.* (1989), Baillie and Bollerslev (1992) and McNees (1995).

4.4.2 Independence

Christoffersen (1998) suggests testing for independence by modelling the indicator function as a binary first-order Markov chain with transition probability matrix:

$$\mathbf{\Pi}_1 = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}. \tag{4.13}$$

That is, we suppose that at any point in time there is a fixed probability that the process will record a hit or a miss next period, where these probabilities depend only on whether a hit, or a miss, was recorded this period. π_{10} is the probability of a miss ('0') if a hit ('1') was recorded in the current period, etc., so in general:

$$\pi_{ij} = \Pr(I_t = j | I_{t-1} = i).$$

A pth order Chain is characterized by I_t depending not just on I_{t-1}, but also $\{I_{t-2}, \ldots, I_{t-p}\}$. More generally, one might suppose that the probabilities of the transitions are different at different times, and/or might be affected by extraneous variables, so a very particular and simple form of dependence in the $\{I_t\}$ is being permitted here. Note that $\pi_{00} + \pi_{01} = 1$

and $\pi_{10} + \pi_{11} = 1$, so that we can write (4.13) as:

$$\mathbf{\Pi}_1 = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}. \tag{4.14}$$

Under independence, the state the process is in at t conveys no information on the relative likelihood of it being in one state as opposed to another at $t + 1$. This is achieved by restricting the transition probabilities by setting $\pi_{ij} = \pi_j$, $i, j = 0, 1$ where $\pi_j = \Pr(I_t = j)$ (the unconditional probability of being in state j). So under the null of independence:

$$\mathbf{\Pi}_2 = \begin{bmatrix} 1 - \pi_1 & \pi_1 \\ 1 - \pi_1 & \pi_1 \end{bmatrix}. \tag{4.15}$$

The π_{ij} and π_i, $i, j = 1, 2$, are estimated by their sample frequencies, and the LR test is based on the unrestricted likelihood:[8]

$$L(\hat{\mathbf{\Pi}}_1) = (1 - \hat{\pi}_{01})^{n_{00}} \hat{\pi}_{01}^{n_{01}} (1 - \hat{\pi}_{11})^{n_{10}} \hat{\pi}_{11}^{n_{11}} \tag{4.16}$$

relative to that with (4.15) imposed

$$L(\hat{\mathbf{\Pi}}_2) = (1 - \hat{\pi}_1)^{(n_{00}+n_{10})} \hat{\pi}_1^{(n_{01}+n_{11})}, \tag{4.17}$$

where n_{ij} is the number of times state i is followed by state j.

The usual LR test statistic has (asymptotically) a χ^2 distribution with one degree of freedom under the null hypothesis of independently distributed indicator function values. This test will be unaffected by any divergence of the actual (unconditional) coverage from the nominal. Thus, a sensible strategy is to combine the tests of independence[9] and correct unconditional coverage to obtain a test of correct conditional coverage.

Granger *et al.* (1989, note c to table 1, p. 91) offer an alternative approach to testing for independence. They use a 'contingency table' approach, based on whether the number of occurrences of (say) zeros followed by zeros is consistent with there being no association between the occurrence of a zero in one period, and the occurrence of a zero in the following period.

Finally, Christoffersen and Diebold (2000) suggest using tests of correct conditional coverage as a (model-free) means of assessing whether volatility is forecastable. If volatility is forecastable, then a fixed-width

interval for $t = 1, \ldots, n$, with an arbitrarily chosen coverage of p, should not be conditionally correct. If the conditional variance of the process varies over time, we have argued that fixed-width intervals will result in clusterings of misses at high volatility times. The non-unit eigenvalue of the matrix of transition probabilities (4.14) is proposed as a natural measure of the degree of forecastability in volatility. $\lambda = \pi_{11} - \pi_{01}$ solves:

$$\left| \lambda I_2 - \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix} \right| = 0$$

and is also the first-order autocorrelation coefficient of $\{I_t\}$.

4.5 Regression-based tests of conditional efficiency

Tests of conditional efficiency that restrict the information set to lagged values of the indicator variable, as described in Section 4.4, may in some cases lack power. It almost goes without saying that extending the information set is desirable – Clements and Taylor (2003) argue that this is especially true for intradaily returns, and this is the focus of the empirical example in Section 4.7. Engle and Manganelli (1999) give a general argument as to why tests based on lagged indicator values only are necessary but not sufficient to assess the performance of interval forecasts. Suppose we have a sequence of interval forecasts of which a proportion p are 'very large' and $(1-p)$ are of close to zero length. Moreover, these two classes of interval are independently distributed: the probability of the interval being 'large' is p irrespective of the type of interval previously observed. This sequence satisfies conditional efficiency on the restricted information set, by construction. But at any point in the sequence, the probability that the interval contains the actual is virtually zero or unity. This parallels a similar argument made by Nordhaus (1987, p. 673) in the context of assessing the efficiency of 'fixed-event' forecasts, suggesting such tests have little bearing on 'accuracy', *viz*.

> A baboon could generate a series of weakly efficient forecasts by simply wiring himself to a random-number generator, but such a series of forecasts would be completely useless.

The Engle–Manganelli sequence of interval forecasts could be rejected if the information set was extended to include, for example, the length of the current interval, since this would be highly (perfectly) correlated with

$\{I_t\}$. The Markov chain approach can be adapted to incorporate explanatory variables by positing logit models for the transition probabilities, for example, but a more attractive and conceptually simpler approach is the regression-based approach of for example, Christoffersen (1998, pp. 849–850) and Engle and Manganelli (1999).

The regression-based test assesses whether $E(I_t|\Omega_{t-1}) = p$ in:

$$I_t = \alpha_0 + \sum_{s=1}^{S} \alpha_s I_{t-s} + \boldsymbol{\beta}'\mathbf{W}_{t-1} + \epsilon_t, \quad t = S+1, S+2, \ldots, n, \tag{4.18}$$

where the information set now allows up to S lags of the indicator variable as well as a general vector of variables $\mathbf{W}_{t-1}$. The hypothesis of conditional efficiency is thus more general than that $\{I_t\}$ is i.i.d. Bernoulli (p). The test of the independence part of the conditional efficiency hypothesis is based on testing $\boldsymbol{\Phi} = \mathbf{0}$, where $\boldsymbol{\Phi} = (\alpha_1, \ldots, \alpha_S, \boldsymbol{\beta}')'$. Non-zero α_i suggest the $\{I_t\}$ sequence is serially correlated, and the $\boldsymbol{\beta} \neq \mathbf{0}$ suggests that misses are associated with the values of variables which were known when the forecasts were made. A test for correct conditional coverage is a joint test of $(\boldsymbol{\Phi}, \alpha_0) = (\mathbf{0},\ p)$, which implies $E(I_t \mid \Omega_{t-1}) = p$.

The binary nature of the dependent variable suggests fitting a regression model to a logistic transformation of the dependent variable (Clements and Taylor 2003). In terms of (4.18), the logit model is:

$$\Pr(I_t = 1) = \Lambda(\alpha_0, \boldsymbol{\Phi}; \mathbf{x}_t), \quad t = S+1, \ldots, n, \tag{4.19}$$

where:

$$\Lambda(\alpha_0, \boldsymbol{\Phi}; \mathbf{x}_t) = e^{\alpha_0 + \boldsymbol{\Phi}'\mathbf{x}_t}/(1 + e^{\alpha_0 + \boldsymbol{\Phi}'\mathbf{x}_t})$$

and $\mathbf{x}_t = (I_{t-1}, \ldots, I_{t-S}\ \ \mathbf{W}_{t-1})'$. The tests for independence and correct conditional coverage can be performed by using LR tests. In the case of the latter test, the restricted regression involves setting $\boldsymbol{\Phi} = \mathbf{0}$ and $\alpha_0 = \ln(p/(1-p))$, whereas the former requires only that $\boldsymbol{\Phi} = \mathbf{0}$.

4.6 Interval forecast construction and ARCH

The calculation of interval forecasts for AR models with independent errors can be extended in a straightforward way to accommodate ARCH-type error processes. Consider the AR(p) with $\varepsilon_t \sim N(0, h_t)$, where h_t

follows a volatility model as discussed in Chapter 3. The standard BJ 1-step interval is:

$$\{y_{T+1|T} + z_{\alpha/2}\sqrt{h_{T+1}}, \quad y_{T+1|T} + z_{1-\alpha/2}\sqrt{h_{T+1}}\}, \tag{4.20}$$

where recall that $z_\gamma = \Phi(\gamma)$ (compare (4.2)). Even if $\{\varepsilon_t\}$ is normal the assumption of normality is not valid for more than 1-step ahead forecasts. But even for 1-step forecasts Granger *et al*. (1989) question the wisdom of invoking the normality assumption (or the Student t assumption) for calculating intervals (or quantiles of the distribution more generally). Their suggestion is to consider in addition to (4.20) intervals constructed as:

$$\{y_{T+1|T} + \hat{Q}_{\alpha/2}\sqrt{h_{T+1}}, \quad y_{T+1|T} + \hat{Q}_{1-\alpha/2}\sqrt{h_{T+1}}\}, \tag{4.21}$$

for a nominal $(1 - \alpha) \times 100\%$ coverage, where $\hat{Q}_\gamma$ is the empirical γ-percentile of the standardized residuals $\hat{\epsilon}_t/\sqrt{h_t}$. When the standardized residuals are approximately normal, $\hat{Q}(\gamma) \simeq \Phi(\gamma) = z_\gamma$ and the two intervals are the same. When the standardized residuals have fatter tails than the standard normal (i.e., excess kurtosis) then $\hat{Q}(\alpha/2) < z_{\alpha/2}$ and $\hat{Q}(1 - \alpha/2) > z_{1-\alpha/2}$ so that the intervals based on the empirical distribution will be wider than those that assume normality.

In the empirical illustration of Section 4.7 we evaluate intervals constructed from (4.21) using the tests outlined in Sections 4.4 and 4.5. Constructing an interval forecast can be interpreted as estimating two quantiles (say, the 5% and 95% for $\alpha = 0.1$). Granger *et al*. (1989) suggest using quantile regression to combine different quantile estimates, and to test for unbiasedness.[10]

For example, let $Z_t = y_{t|t-1} + \hat{Q}_p\sqrt{h_t}$ be an estimate of the pth quantile (in this case derived from the empirical distribution of the standardized residuals). Suppose we have $t = 1, \ldots, n$, and wish to test whether the $\{Z_t\}$ are well-specified estimates of the conditional p quantile, versus, for example, that the quantiles should be $\{\delta + \beta Z_t\}$. We could undertake the quantile regression (see Koenker and Bassett 1978, 1982):

$$\min_{\delta,\beta} \frac{1}{n}\sum_{t=1}^{n} |y_t - (\delta + \beta Z_t)|[p \times 1_{(y_t \geq (\delta+\beta Z_t))} + (1 - p) \times 1_{(y_t < (\delta+\beta Z_t))}]. \tag{4.22}$$

If the $\{Z_t\}$ are unbiased, we would expect to find $(\delta, \beta) = (0, 1)$. Notice that a separate regression needs to be undertaken for each p. To test for, or estimate, a combination of quantile estimators, $\{\delta + \beta Z_t\}$ would be substituted by, for example $\{\delta + \beta_1 Z_t + \beta_2 \tilde{Z}_t\}$ in (4.22), where $\tilde{Z}_t$ is a rival estimator of the same quantile.

Some intuition into the form of (4.22) can be gained by considering $p = 0.5$:

$$\min_{\delta,\beta} \frac{1}{n} \sum_{t=1}^{n} |y_t - (\delta + \beta Z_t)| \times 0.5.$$

The estimators so defined are the mean absolute deviation estimators.

4.7 Empirical illustration

The empirical illustration is taken from Clements and Taylor (2003) and is based on hourly futures returns in the FTSE100 index futures market. A number of specific issues arise for intradaily data, and these are discussed in the following section before embarking on the illustration proper.

4.7.1 Interval forecasts and intradaily data

A common feature of financial markets is that the volatility of returns varies in a systematic pattern during each day: 'Most high-frequency asset returns exhibit seasonal volatility patterns', to quote Bollerslev and Ghysels (1996, p. 139), and see also Baillie and Bollerslev (1989) and Gallant *et al*. (1992), *inter alia*. With specific reference to the FTSE100 index futures market, Tse (1999) finds that volatility is high during the opening and closing of the floor trading periods, when investors have the greatest desire to re-balance their portfolios. As a consequence, an obvious choice of variable to be included in $\{\mathbf{W}_{t-1}\}$ in (4.18) is a set of hourly dummies, to check that the interval forecasts adequately reflect the changing volatilities in returns during the course of the day. Thus, $\boldsymbol{\beta}'\mathbf{W}_{t-1}$ in (4.18) includes $\sum_{s=1}^{S-1} \mu_s D_{s,t}$, where $D_{s,t} = 1$ when $t = (N-1)S + s$, and $D_{s,t} = 0$ otherwise. Here s indexes the hour in the day, of which there are a total of S, $s = 1, 2, \ldots, S$, and $N = 1, 2, \ldots, T/S$ is the number of days. The inclusion of the $\{D_{s,t}\}$ variables in Ω_t allows us to test against the specific alternative that the forecast intervals are not adequately capturing recurrent periodic effects, and are motivated by the nature of financial data. Note that the $\{D_{s,t}\}$ are known at $t-1$ because they are deterministic.

The first-order Markov chain test is unlikely to have power to detect the more complex dependence structures that may be present in intradaily data – it will fare best when the dependency is between adjacent observations. When there are periodic patterns in volatility, such that periods of high volatility occur every S observations (say), then we need tests with power to detects such patterns in the misses. One possibility is to increase the order of the Markov chain. For example, in other areas of empirical macroeconomic research, such as the analysis of business cycle turning points, second-order chains are routinely used (see, e.g., McQueen and Thorley (1993)). However, specifying higher-order chains may be problematic when the order of periodicity is unknown and may be large, leading to imprecise estimates and poor tests. A better idea is to specify a 'periodic lag' in a first-order chain when the periodicity of the data is apparent. That is, to calculate the transition probabilities based on the periodicity of the underlying data. For S-periodic data we would calculate $\pi_{ij,S} = \Pr(I_t = j | I_{t-S} = i)$ and $n_{ij,S} = \#\{I_t = j, I_{t-S} = i\}$. The $n_{ij,S}$ and $\hat{\pi}_{ij,S}$ can be plugged directly into (4.16) and (4.17) and the resulting statistic remains asymptotically χ^2 with 1 degree of freedom. Christoffersen's test is seen to emerge as a special case with $S = 1$.

4.7.2 Properties of futures returns data

FTSE100 index futures returns and trading volume were obtained from the *LIFFE Sterling Products Tick Data CD*. We consider the returns to the 'nearest' futures contract over the one year sample period commencing 2 January 1998 and ending 29 December 1998. Hourly returns are constructed based on prices observed at 9, 10 a.m. through to 4 p.m., giving seven observations per day, and with 239 full trading days during the sample period, a total of 1673 observations.

The intraday volatility patterns are displayed in Panel A of Figure 4.1. Volatility is highest during the first and penultimate hour of the floor trading period, consistent with the findings of Tse (1999) on UK data and Werner and Kleidon (1996) for the US. Also plotted in Figure 4.1 is the intraday mean of trading volume. Comparing these plots reveals a clear correspondence between the intraday patterns in volatility and volume. These observations motivate the use of the GARCH models described below.

Four GARCH-type models are considered. The first is a standard GARCH(1,1):

$$R_t = \mu + \epsilon_t, \qquad \epsilon_t | \Omega_{t-1} \sim N(0, h_t), \tag{4.23}$$

$$h_t = \omega + \alpha \epsilon_{t-1}^2 + \beta h_{t-1}, \tag{4.24}$$

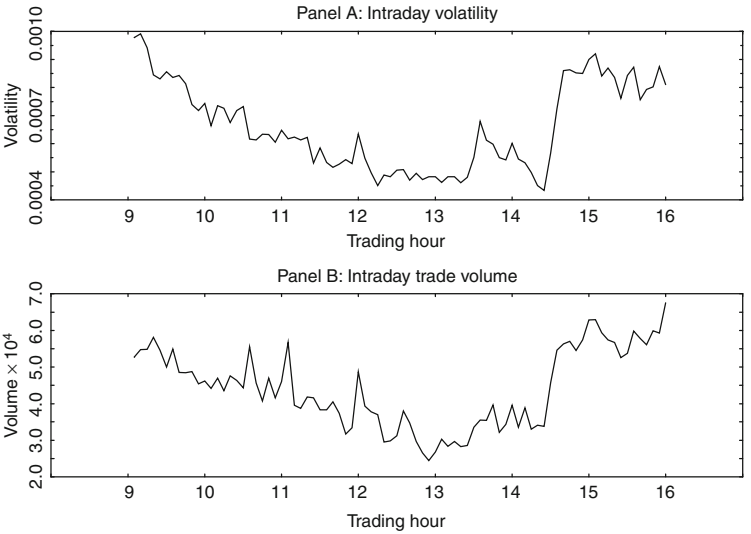


Figure 4.1 Intraday volatility of FTSE100 index futures returns and the trading volume of FTSE100 futures contracts

Note: Intraday volatility is calculated by taking the means of absolute returns during various five-minute intervals over the trading day. Intraday trading volume is calculated by taking the sum of trading volume during various five-minute intervals over the trading day.

where R_t is the nominal return on FTSE100 index futures. Given the similarity in the volatility and trading volume periodicity given in Figure 4.1, the above GARCH model is augmented by allowing the 7th lag of trading volume (i.e., trading volume observed during the same hour of the previous trading day) to enter the volatility equation. This gives the GARCH-V model:

$$h_t = \omega + \alpha\epsilon_{t-1}^2 + \beta h_{t-1} + \delta V_{t-7}, \tag{4.25}$$

where V denotes trading volume.

Two periodic GARCH (PGARCH) models were also considered to capture the intraday patterns. These models allow for six hourly dummy variables, ω_i, $i = 1, \ldots, 6$, in addition to the intercept ω in the GARCH and GARCH-V models, and are denoted by PGARCH and PGARCH-V.

These models are estimated by maximum likelihood using the Marquardt algorithm. The resulting parameter estimates and their associated heteroskedastic-consistent standard errors are given in

Table 4.2.[11] In addition, we record the fits of the models as given by the Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC). The PGARCH-V provides the best fit to the data.

Given the periodic pattern in volatility evident in Figure 4.1, estimates of conditional volatility from a 'good' model should reproduce these

Table 4.2 Volatility model estimates

Test	**Model**			
	GARCH	**GARCH-V**	**PGARCH**	**PGARCH-V**
$\hat{\mu}$	−3.6300	4.0700	5.7000	7.5500
	(9.9700)	(7.5700)	(7.7700)	(6.8700)
$\hat{\omega}$	3.1700	2.4600***	8.4300***	7.4100***
	(3.1500)	(0.6440)	(1.9500)	(1.8800)
$\hat{\omega}_1$			0.9600***	2.8500
			(1.9500)	(3.7200)
$\hat{\omega}_2$			−11.3400***	−11.4000***
			(2.9400)	(2.8000)
$\hat{\omega}_3$			−0.9130***	−7.6400***
			(0.2270)	(1.9900)
$\hat{\omega}_4$			−0.9330***	−7.7100***
			(0.2120)	(1.9600)
$\hat{\omega}_5$			−0.6380***	−5.7800***
			(0.2160)	(1.9500)
$\hat{\omega}_6$			−0.4170*	−4.2300*
			(0.2490)	(2.3300)
$\hat{\alpha}$	0.0204***	0.2061***	0.1197***	0.1126***
	(0.0065)	(0.0388)	(0.0261)	(0.0281)
$\hat{\beta}$	0.9775***	0.0383	0.6879***	0.4557***
	(0.0077)	(0.0568)	(0.0594)	(0.1197)
$\hat{\delta}$		0.0371***		0.0140***
		(0.0042)		(0.0037)
AIC	−8.4837	−8.4913	−8.5332	−8.5621
SIC	−8.4708	−8.4751	−8.5008	−8.5263

Note: The columns record the parameter estimates and heteroskedasticity-consistent standard errors (HCSEs) for the GARCH-type models discussed in the text. Significance at the 1%, 5% and 10% level are denoted by three, two and one asterisks, respectively.

patterns. This can be formally tested by an OLS (ordinary least squares) regression of the model-based estimates of return volatility on hourly dummy variables:

$$\hat{\sigma}_t = \gamma_0 + \sum_{i=1}^{6} \gamma_i D_{i,t} + \epsilon_t. \tag{4.26}$$

In this regression $\hat{\sigma}_t$ is the measure of the FTSE100 index intraday return volatility, and the $D_{i,t}$ are hourly dummy variables. In addition to the model-based measures of volatility, we also consider absolute returns and squared returns as model-free estimates of volatility (where $\hat{\sigma}_t = \sqrt{h_t}$). Table 4.3 records the results of these regressions and of tests that $\gamma_i = 0$, $i = 1, \ldots, 6$. The null of no periodicity in conditional volatility is clearly rejected for the model-free and the GARCH-V, PGARCH and PGARCH-V model estimates of volatility. The standard GARCH model is unable to capture the periodic pattern, and produces an estimate of conditional volatility that does not appear to have a periodic component. Consequently, we would expect interval forecasts from the GARCH model to be poorly specified, in the sense that the misses will tend to be concentrated in the hours at the beginning and end of each day.

Static and dynamic interval forecasts are generated over the entire sample period using full-sample parameter estimates. In the case of static forecasts this involves calculating percentile points of the empirical distribution function of futures returns. The dynamic interval forecasts are generated using estimated GARCH models according to the suggestion in Granger *et al*. (1989, p. 89), as described in Section 4.6.

Interval forecasts designed to cover 95% of future outcomes are presented in Figure 4.2. The dynamic interval forecasts appear to be preferable to static forecasts as they widen during periods of high volatility. However, as the results in Table 4.3 indicate the GARCH intervals fail to widen during the beginning and end of the floor trading period. By contrast, the PGARCH intervals have an intraday pattern of behaviour consistent with the observed intraday volatility.

The results of applying the interval evaluation tests to the static and dynamic interval forecasts with a nominal coverage of 95% are recorded in Table 4.4. The tests are the two Markov chain tests (denoted MC(1) and MC(7)), and various regression-based tests. In particular, four regression-based LR tests are carried out, each based on a logit regression. The first two exclude periodic dummies but allow various lagged values of the indicator variable as explanatory variables. The LOGIT(1) test includes the first lag only, and LOGIT(7) includes lags 1 through to 7. The third

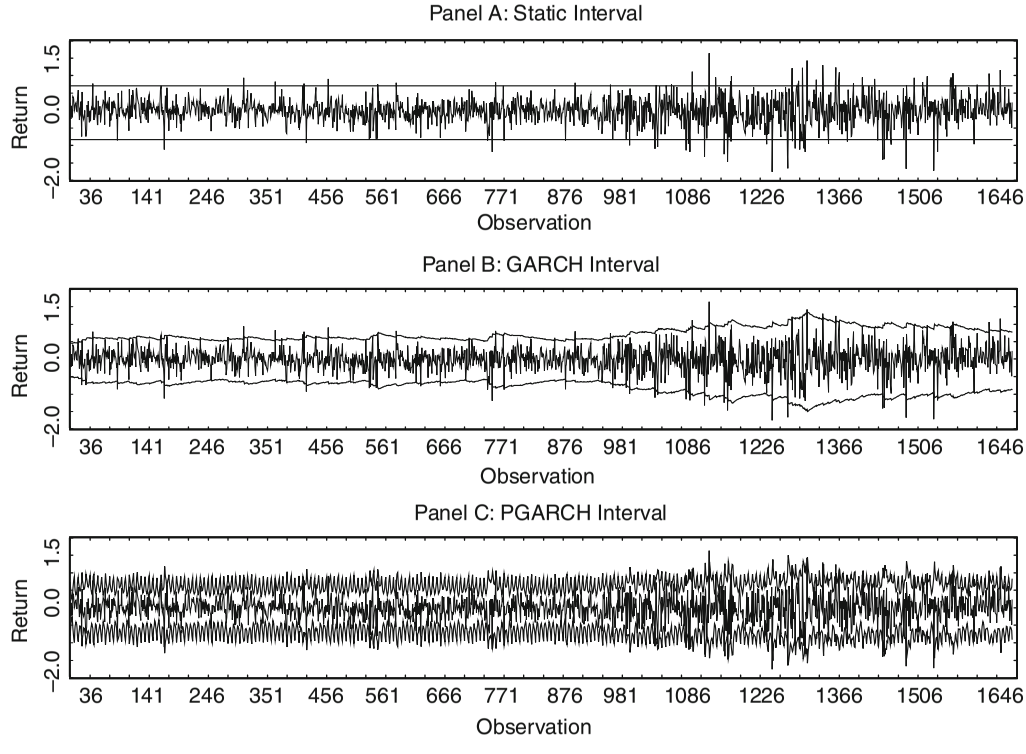


Figure 4.2 FTSE100 index futures returns (in percentage terms) together with the static interval forecasts (Panel A) and dynamic interval forecasts based on an estimated GARCH model (Panel B) and an estimated PGARCH model (Panel C)

Note: In each case the interval is based on 95% coverage.

Table 4.3 Testing for periodic heteroskedasticity

	Volatility measure					
	Abs. rtn	**Sq. rtn**	**GARCH**	**GARCH-V**	**PGARCH**	**PGARCH-V**
$\hat{\gamma}_0$	34.2230***	0.1932***	0.1325***	0.1616***	0.1727***	0.1785***
	(1.5987)	(0.0180)	(0.0047)	(0.0054)	(0.0040)	(0.0036)
$\hat{\gamma}_1$	2.25306	0.0544***	0.0009	0.0043	0.0628***	0.0626***
	(2.2609)	(0.0254)	(0.0066)	(0.0076)	(0.0050)	(0.0050)
$\hat{\gamma}_2$	−6.3464***	−0.0578***	0.0033	0.0053	−0.0311***	−0.0484***
	(2.2609)	(0.0254)	(0.0066)	(0.0076)	(0.0050)	(0.0050)
$\hat{\gamma}_3$	−11.6161***	−0.0958***	0.0033	−0.0276***	−0.0661***	−0.0778***
	(2.2609)	(0.0254)	(0.0066)	(0.0076)	(0.0050)	(0.0050)
$\hat{\gamma}_4$	−14.2861***	−0.1121***	0.0025	−0.0523***	−0.0967***	−0.1021***
	(2.2609)	(0.0254)	(0.0066)	(0.0076)	(0.0050)	(0.0050)
$\hat{\gamma}_5$	−13.8413***	−0.1148***	0.0015	−0.0551***	−0.0902***	−0.0952***
	(2.2609)	(0.0254)	(0.0066)	(0.0076)	(0.0050)	(0.0050)
$\hat{\gamma}_6$	−9.7216***	0.0800***	0.0003	−0.0300***	−0.0640***	−0.0671***
	(2.2609)	(0.0254)	(0.0066)	(0.0076)	(0.0050)	(0.0050)
F-test	17.0000***	12.4120***	0.0830	23.0960***	256.7750***	277.0100***

Note: The table records the estimated coefficients and standard errors obtained from the OLS regression,

$$\hat{\sigma}_t = \gamma_0 + \sum_{i=1}^{6} \gamma_i D_{i,t} + \epsilon_t,$$

where $\hat{\sigma}_t$ is a measure of FTSE100 index intraday return volatility and $D_{i,t}$ is a dummy variable that takes a value of unity during the ith trading hour and zero otherwise. The first two columns report the results for $\hat{\sigma}_t$ measured as the absolute value of the return at time t, and the square of the return at time t, respectively, and the remainder are the model-based volatility estimates. The F-test of the significance of all the periodic dummies is reported in the final row of the table. In each regression the dependent variable has been multiplied by 10,000. Significance at the 1%, 5% and 10% level are denoted by three, two and one asterisks, respectively.

Table 4.4 Evaluating interval forecasts

Test	95% interval forecast				
	Static	**GARCH**	**GARCH-V**	**PGARCH**	**PGARCH-V**
Panel A: Independence tests					
MC(1)	0.03	0.36	0.31	0.97	0.31
MC(7)	0.08	0.17	0.04	0.59	0.24
RUNS	0.02	0.40	0.25	0.93	0.25
LOGIT(1)	0.03	0.36	0.50	0.97	0.31
LOGIT(7)	0.02	0.13	0.12	0.16	0.07
PLOGIT(7)	0.00	0.00	0.00	0.59	0.37
PLOGIT(7)-V	0.00	0.00	0.00	0.03	0.31
PLOGIT(7)-VX	0.00	0.00	0.01	0.05	0.30
Panel B: Conditional coverage tests					
MC(1)	0.10	0.65	0.47	0.96	0.47
MC(7)	0.22	0.39	0.10	0.83	0.39
LOGIT(1)	0.10	0.65	0.50	0.96	0.50
LOGIT(7)	0.03	0.19	0.17	0.23	0.11
PLOGIT(7)	0.00	0.00	0.00	0.66	0.45
PLOGIT(7)-V	0.00	0.00	0.01	0.04	0.37
PLOGIT(7)-VX	0.00	0.00	0.01	0.06	0.36

Note: The table contains the *p*-values for various tests of interval forecast adequacy for intervals generated by a number of alternative models.

regression-based test allows periodic dummies as explanatory variables in addition to lags 1 to 7 of the indicator variable, and is referred to as the PLOGIT(7) test. The same set of regressors augmented with the seventh lag of trading volume is referred to as the PLOGIT(7)-V test. The test PLOGIT(7)-VX includes in addition the first lags of absolute returns and interval length as regressors. Finally, the independence of the indicator series is tested using a runs test.[12] For the other tests both independence and correct conditional coverage is examined.

In terms of independence, the static intervals are clearly rejected at the 10% level on the basis of all tests. However, only the PLOGIT(7), PLOGIT(7)-V and PLOGIT(7)VX tests decisively reject the adequacy of the GARCH and GARCH-V interval forecasts. The rejections on tests of these forecasts is due to the inability of the GARCH and GARCH-V models to generate intraday conditional volatility patterns that match the data. When the PGARCH forecast intervals are considered it is only the PLOGIT(7)-V and PLOGIT(7)-VX tests that reject their adequacy. Finally, the PGARCH-V interval forecasts appear adequate when these

tests are performed. This result is compatible with the superior fit of this model of the volatility process. The marginal contribution from allowing absolute returns and interval length as regressors (PLOGIT(7)-VX compared to PLOGIT(7)-V) is in all cases small. Finally, as is apparent from the table, similar conclusions are obtained when we test for correct conditional coverage.

4.8 Summary

This chapter considers the evaluation of interval forecasts or prediction intervals. An interval forecast gives a range within which the outcome is expected to fall with a given probability. We begin by considering the calculation of interval forecasts for linear autoregressions with independent disturbances, starting with the BJ approach, and then setting out a bootstrap approach. A bootstrap procedure for models with ARCH disturbances is also reviewed. The BJ and BS approaches are compared in a Monte Carlo that calculates the actual coverage rates of intervals with a given nominal coverage.

As well as having actual coverage rates close to the nominal, a 'good' set of interval forecasts should be conditionally efficient in the sense that the probability of a hit (interval includes the actual) should not vary in a systematic way with any variables in the agent's information set at the time the forecast was made. A number of tests of conditional efficiency are described, and applied to the evaluation of interval forecasts for hourly returns on FTSE100 index futures.

5
Density Forecasts

5.1 Introduction

In recent years there has been considerable interest in density forecasts. This has been fuelled by the rapidly expanding field of financial risk management, as well as by the literature on inflation forecasting.[1] For example, if the goal is to achieve an inflation rate in a certain range or target band, a point forecast of inflation is of limited value. A histogram (or density forecast) that assigns probabilities to inflation falling in certain intervals will be more informative about the likelihood of the target being met. That said, the forecast histogram will only be of value to the extent that the forecast probabilities accurately capture the true probabilities. As in previous chapters, our focus will be on evaluation, where the forecasts are now densities, or histograms, or probability distributions.

We begin by reviewing recent methods of evaluation based only on the sequence of forecasts and the outcomes. Thus, no recourse is made to the method of construction of the forecasts. This is appropriate when the probability distributions are survey-based or the method of construction is unknown to the investigator. Kling and Bessler (1989) consider purely model-based forecasts but evaluate the density forecasts without regard to the model, and 're-calibrate' their forecasts based solely on the past performance of previously issued forecasts. Section 5.2 describes the probability integral transform approach, Section 5.3 extensions to multivariate densities, and Section 5.4 the calibration of density forecasts. A density forecast can be viewed as being comprised of a sequence of interval forecasts generated by allowing the nominal coverage rate to vary over all values in the unit interval. The evaluation of a sequence of interval forecasts with a specific nominal coverage rate therefore assesses one aspect of the underlying sequence of forecast densities. In Section 5.5 we

exploit the relationship between interval and density forecasts to show how the interval forecast evaluation techniques reviewed in Section 4.4 can be extended to provide 'contingency-table-type' goodness-of-fit tests of densities, following Wallis (2003). Sections 5.6 and 5.7 describe applications to the evaluation of survey-based forecast densities of annual US inflation and to the Bank of England inflation forecasts (see also Diebold *et al*. (1999), Wallis (2003) and Clements (2003, 2004).

We then turn our attention to the evaluation of model-based conditional forecast densities, and discuss contributions by Andrews (1997), Li and Tkacz (2002) and Corradi and Swanson (2003) in Section 5.8.

5.2 Probability distribution forecast evaluation

The key tool in the recent literature on density forecast evaluation is the probability integral transform. This can be traced back at least to Rosenblatt (1952), with recent contributions by Shephard (1994), Kim *et al*. (1998) and Diebold *et al*. (1998). Suppose we have a series of 1-step forecast densities for the value of a random variable $\{Y_t\}$, denoted by $p_{Y,t-1}(y)$, where $t = 1, \ldots, n$. The probability integral transforms (p.i.ts) of the realizations of the variable with respect to the forecast densities are given by:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(y_t) \tag{5.1}$$

for $t = 1, \ldots, n$, where $P_{Y,t-1}(y_t)$ is the forecast probability of Y_t not exceeding the realized value y_t. In terms of the random variables $\{Y_t\}$, rather than their realized values $\{y_t\}$, we obtain random variables denoted by $\{Z_t\}$:

$$Z_t = \int_{-\infty}^{Y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(Y_t).$$

When the forecast density equals the true density, $f_{Y,t-1}(y)$, it follows that $Z_t \sim U(0,1)$, where $U(0,1)$ is the uniform distribution over $(0,1)$. Even though the actual conditional densities may be changing over time, provided the forecast densities match the actual densities at each t, then $Z_t \sim U(0,1)$ for each t, and the Z_t are independently distributed of each other, such that the realized time series $\{z_t\}_{t=1}^n$ is an i.i.d. sample from a $U(0,1)$ distribution.

If we let $q_{Z,t}(z)$ denote the density of Z_t, we can derive the result that $q_{Z,t}(\cdot)$ is the uniform distribution using a standard 'change-of-variables' argument. For $Z_t = P_{Y,t-1}(Y_t)$, we obtain $q_{Z,t}(\cdot)$ as:

$$q_{Z,t}(z) = f_{Y,t-1}(P^{-1}_{Y,t-1}(z)) \left| \frac{\partial P^{-1}_{Y,t-1}(Z_t)}{\partial Z_t} \right|$$
$$= \frac{f_{Y,t-1}(P^{-1}_{Y,t-1}(z))}{p_{Y,t-1}(P^{-1}_{Y,t-1}(z))}. \tag{5.2}$$

When $f_{Y,t-1}(\cdot)$ and $p_{Y,t-1}(\cdot)$ are the same, $q_{Z,t}(z) = 1$ for $z \in [0, 1]$, that is, $Z_t \sim U(0, 1)$. This holds for each t, $t = 1, \ldots, n$, so that the time subscript on $q_{Z,t}$ is redundant.

This suggests we can evaluate whether the conditional forecast densities match the true conditional densities by testing whether $\{z_t\}_{t=1}^n$ is i.i.d. $U(0, 1)$. This is a joint hypothesis of independence and uniformity.[2] Independence can be assessed informally by examining correlograms of $\{z_t - \bar{z}\}$, where $\bar{z} = n^{-1} \sum_{i=1}^n z_i$, and of powers of this series, $\{(z_t - \bar{z})^i\}$, $i = 2, 3, \ldots$, as a check for dependence in higher moments, which would be incompatible with the independence claim. Formal tests of autocorrelation can also be performed. Uniformity can also be assessed in a number of ways: whether the empirical cdf of the $\{z_t\}$ is significantly different from the theoretical uniform cdf (a 45° line) using, for example, the Kolmogorov–Smirnov (KS) test of whether the maximum difference between the two cdfs exceeds some critical value, or the Cramer-von-Mises 'integrated-squared' distance measure. Because of the possible distortions arising from dependence in $\{z_t\}$ (the i.i.d. assumption not holding) when testing for uniformity, and for testing for autocorrelation when 'identically distributed' fails, particularly in small samples, formal tests are often supplemented with graphical analyses.

Other ways of testing probability distributions are given in Thompson (2002), who suggests a frequency domain test of the uncorrelatedness of the $\{z_t\}$ based on the cumulative periodogram approach of Durbin (1969), and the generalized spectral approach of Hong (2001).

Finally, Berkowitz (2001) has suggested taking the inverse normal CDF transformation of the $\{z_t\}_{t=1}^n$ series, to give, say, $\{z_t^*\}_{t=1}^n$, on the grounds that more powerful tools can be applied to testing the null that the $\{z_t^*\}_{t=1}^n$ are i.i.d. $N(0, 1)$ (for $h = 1$) compared to one of i.i.d. uniformity of the original $\{z_t\}_{t=1}^n$ series. He proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a

three-degree of freedom test of zero-mean, unit variance and independence. In each case the maintained assumption is that of normality, so that standard likelihood ratio tests are constructed using the Gaussian likelihoods. The assumption of normality of $\{z_t^*\}_{t=1}^n$ is also amenable to testing, for example, using the Shenton and Bowman (1977) two-degree of freedom asymptotic chi-squared test or the test recommended by Doornik and Hansen (1994).

5.3 Joint probability distributions

Suppose now we have a joint forecast density for $\{y_{1t}, y_{2t}\}$, $p_{t-1,Y_1,Y_2}(y_1, y_2)$. The tests based on the probability integral transform described in Section 5.2 can still be used. We begin by factoring the joint density into the product of the conditional of y_{2t} given y_{1t}, and the marginal for y_{1t} (or vice versa),

$$p_{t-1,Y_1,Y_2}(y_1, y_2) = p_{t-1,Y_2 \mid Y_1}(y_2 \mid y_1) p_{t-1,Y_1}(y_1)$$

and calculate the probability integral transforms for the conditionals and marginals separately. Let the p.i.t. sequence for the conditional of y_{2t} given y_{1t} be $\{z_{2|1,t}\}_{t=1}^n$, and for the marginal for y_{1t} be $\{z_{1,t}\}_{t=1}^n$. Under the null that the predicted joint density is correct, that is, $p_{t-1,Y_1,Y_2}(y_1, y_2) = f_{t-1,Y_1,Y_2}(y_1, y_2)$, the two sequences will each be i.i.d. samples from a $U(0, 1)$ (and the two sequences will themselves be independent).

Diebold *et al.* (1998, p. 881) propose stacking the two sequences of p.i.ts into a single $2n \times 1$ vector $[z_{2|1,1}, \ldots, z_{2|1,n} : z_{1,1}, \ldots, z_{1,n}]'$ (or $[z_{1|2,1}, \ldots, z_{1|2,n} : z_{2,1}, \ldots, z_{2,n}]'$), and then testing for i.i.d. uniformity of the stacked vector. Clements and Smith (2000) propose instead basing a test on the n dimensional vector with typical element $\{z_t^j = z_{2|1,t} \times z_{1,t}\}$, that is, on the products. This has the advantage of preserving the temporal ordering and by doing so will have power to detect mis-specifications of the correlations between the two variables. Clements and Smith (2000) derive the distribution function for the 'product' for up to three variables (and indicate how it can be derived for any number of variables) and this can be used to transform the $\{z_t^j\}$ series to an i.i.d. $U(0, 1)$ sample under the null that the forecast density is correctly specified. Clements and Smith (2002) consider in addition tests based on ratios of the conditional and marginal p.i.ts, with typical element $\{z_{2|1,t}/z_{1,t}\}$. The appendix records the derivation of the distribution functions for the product and ratio under the null. Clements and Smith (2002) provide Monte Carlo evidence that compares the 'stacked', 'product' and 'ratio' tests under

various forms of mis-specification of the forecast density for the true (joint) density.

5.4 Calibration

Provided a series of forecast densities are dynamically well-specified, it may be possible to 'correct' future forecast densities for mis-specifications that result in the $\{z_t\}$ sequence not being $U(0, 1)$. More precisely, Diebold *et al.* (1999, p. 663) assume that $p_{Y,t-1}(y)$ is a different member of the location-scale family of distributions than $f_{Y,t-1}()$, albeit that the conditional mean and variance are correctly specified. 'Calibration' or 'recalibration' (e.g., Dawid 1984; Kling and Bessler 1989) is the process of correcting future forecast densities for the mis-specifications apparent from a consideration of $p_{Y,t-1}(y)$ and y_t, $t = 1, \ldots, n$. It requires that the relationship between the (uncorrected) $p_{Y,t-1}(y)$ and $f_{Y,t-1}(y)$ remains the same in future, $t = n + 1, \ldots$, as it was in the past ($t = 1, \ldots, n$). The assumption that the $\{z_t\}_{t=1}^{n}$ are i.i.d. allows us to rearrange (5.2) to give:

$$f_{Y,t-1}(y) = p_{Y,t-1}(y)q_Z(z)$$

dropping the time subscript on $q_{Z,t}()$. The i.i.d. assumption allows us to estimate $q_Z()$ as the empirical pdf of $\{z_t\}_{t=1}^{n}$, which we denote as $\hat{q}_Z()$, and using this in place of $q_Z()$, the re-calibrated forecast densities are given by:

$$\hat{p}_{Y,t-1}(y) = p_{Y,t-1}(y)\hat{q}_Z(z)$$

for $t = n + 1, \ldots$. The $\hat{p}_{Y,t-1}(y)$ may provide a closer match to $f_{Y,t-1}(y)$ than the $p_{Y,t-1}(y)$.

Calibration extends in a straightforward fashion to the multivariate case. Suppose we have two variables, with actual joint pdf. $f_{t-1,Y_1,Y_2}(y_1, y_2)$, forecast pdf $p_{t-1,Y_1,Y_2}(y_1, y_2)$, then:

$$f_{t-1,Y_1,Y_2}(y_1, y_2) = p_{t-1,Y_1,Y_2}(y_1, y_2)q(z_1, z_2),$$

where we have assumed the $q_{t,Z_1,Z_2}()$ time subscript can be dropped. The re-calibrated forecasts are:

$$\hat{p}_{t-1,Y_1,Y_2}(y_1, y_2) = p_{t-1,Y_1,Y_2}(y_1, y_2)\hat{q}(z_1, z_2).$$

Diebold *et al*. (1999) provide an illustration of the evaluation and calibration of multivariate density forecasts of high-frequency exchange rate data.

5.5 Density and interval forecasts

In Section 4.4 we tested for correct unconditional coverage of a sequence of interval forecasts by performing a likelihood ratio test of the restricted likelihood for a nominal coverage rate of p against the likelihood for a coverage rate of $\hat{\pi} = n_1/n$. The likelihoods were given by:

$$L(p) = (1-p)^{n_0} p^{n_1} \tag{5.3}$$

and:

$$L(\hat{\pi}) = (1-\hat{\pi})^{n_0} \hat{\pi}^{n_1}, \tag{5.4}$$

respectively. As discussed by Wallis (2003), this can be viewed as a likelihood ratio goodness-of-fit test where there are two classes with unequal probabilities. The first class is that a hit occurs with probability p under the null, and the second class is 'no-hit' with probability of $(1-p)$. In terms of Pearson's chi-squared statistic:

$$\sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i},$$

$K = 2$, and O_1 and O_2 are the number of observed hits and misses, n_1 and n_0, and E_1 and E_2 are the expected number under the null, pn and $(1-p)n$ respectively.

We can obtain goodness-of-fit tests for densities if we extend K beyond 2, and in so doing use equiprobable classes. For example, for $K = 4$, the classes are defined by $(-\infty, c_{0.25}]$, $(c_{0.25}, c_{0.5}]$, $(c_{0.5}, c_{0.75}]$ and $(c_{0.75}, \infty)$, where c_α is such that $\alpha = F(c_\alpha)$ and F is the distribution function (which typically will be time-subscripted). Then $E_i = n/K$ for all i, so the Pearson statistic is given by:

$$\sum_{i=1}^{K} \frac{(n_i - n/K)^2}{n/k} = \frac{K}{n} \sum_{i=1}^{K} n_i^2 - n,$$

which has a limiting chi-squared distribution with $K-1$ degrees of freedom.

An asymptotically equivalent statistic can be written down that generalizes the likelihoods given by (5.3) and (5.4) for interval forecast evaluation to the evaluation of density forecasts. Suppose the classes are defined for the p.i.ts, $\{z_t\}$, such that for K equiprobable classes we have boundaries j/K where $j = 0, 1, \ldots, K$. Then $n_j = \sum_{i=1}^{n} 1((j-1)/K < z_i < j/K)$, $j = 1, \ldots, K$. Under the null, $\Pr((j-1)/K < z < j/K) = K^{-1}$ for all j. Under the alternative, $\Pr((j-1)/K < z < j/K) = n_j/n$ for $j = 1, \ldots, K$ (the MLEs).

Thus, the likelihoods under the null and alternative are given by:

$$L\left(\frac{1}{K}, \ldots, \frac{1}{K}\right) = \prod_{i=1}^{K} \left(\frac{1}{K}\right)^{n_i} = \left(\frac{1}{K}\right)^{n} \tag{5.5}$$

given that $\sum_{i=1}^{K} n_i = n$, and:

$$L\left(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\right) = \prod_{i=1}^{K} \left(\frac{n_i}{n}\right)^{n_i}. \tag{5.6}$$

The arguments of $L(\cdot)$ are the probabilities at which the functions are evaluated. The likelihood ratio statistic is:

$$\begin{aligned} LR &= -2 \ln \left(\frac{L(\frac{1}{K}, \ldots, \frac{1}{K})}{L(\frac{n_1}{n}, \ldots, \frac{n_k}{n})} \right) \\ &= -2 \left[-n \ln K - \sum_{i=1}^{K} n_i (\ln n_i - \ln n) \right] \\ &= 2 \sum_{i=1}^{K} n_i \ln \left(\frac{K n_i}{n} \right), \end{aligned}$$

which is chi-squared with $K - 1$ degrees of freedom.

Wallis (2003) also notes that tests of independence can be based on Pearson goodness-of-fit tests. Such tests arise quite naturally for interval forecasts where the 2×2 contingency-table is appropriate, as in Granger *et al.* (1989). But they would appear to be less suitable for testing independence in the case of density forecasts, where increasing K much beyond 2 is likely to result in cells with zero entries unless the number of forecasts is large.

5.6 Empirical illustration (I): the SPF probability distributions

The SPF[3] is a quarterly survey of macroeconomic forecasters of the US economy that began in 1968 as the ASA–NBER survey, administered by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER), and since June 1990 has been run by the Philadelphia Fed, as the Survey of Professional Forecasters (SPF). The majority of the survey questions ask respondents to report their point forecasts for a number of variables at various forecast horizons, from which median forecasts are calculated, but respondents are also asked to report discrete probability forecasts, or histograms, for output growth and inflation for the current and following year, which are then averaged to produce *the* forecast distributions.[4]

Diebold *et al*. (1999) discuss the survey and the complications that arise in using the inflation forecasts. In order to obtain a non-overlapping series of forecasts – in the sense that the realization of inflation in period t is known before making the forecast of the next period – they take the density forecasts made in the first quarter of each year of the annual change in that year on the preceding year. This avoids the counterpart of the problem in the point forecast evaluation literature that optimal h-step forecasts that overlap will be autocorrelated: see Section 2.1.1. Further complications are that both the base years of the price indices and the indices themselves have changed over time. The change in base years is likely to have had a minor effect on the inflation rate, and we construct a series of realizations of annual inflation that matches the indices for which probability assessments were requested. Thus, for 1969 to 1991 we use the implicit GNP deflator, for 1992 to 1995 the implicit GDP deflator, and for 1996 to 2002 the chain-weighted deflator, correcting for the changes in the definition of the index but not for base-year changes. Moreover, we use the latest available estimates of the realized values.[5] Finally, as documented by the Philadelphia Fed, the form in which the respondents report their probability assessments has changed over time, with changes in the number of bins and/or their locations and lengths as the perceived likely ranges of the target variables has changed. We calculate critical values and probabilities from the histograms by piecewise linear approximation.[6]

Figure 5.1 portrays the inflation density forecasts as Box–Whisker plots along with the realizations. The observations for 1969 to 1996

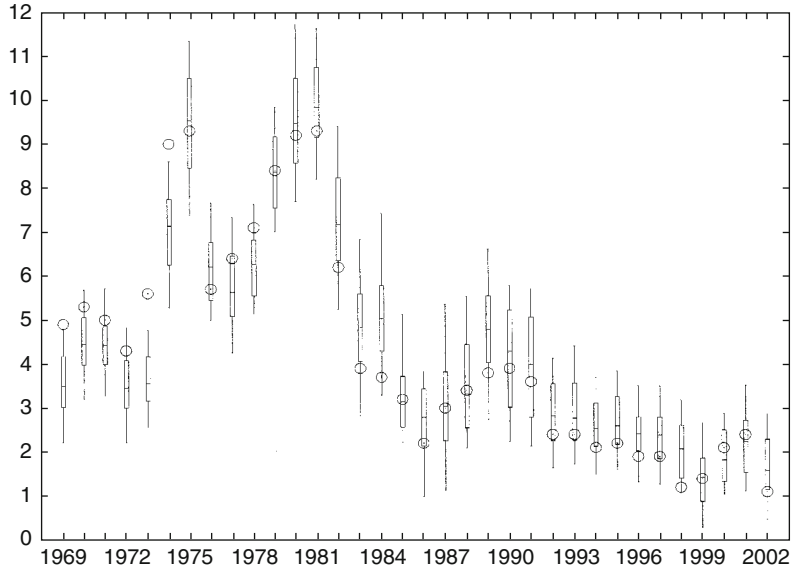


Figure 5.1 Inflation forecast probability distributions shown as Box–Whisker plots and realizations

Notes: The boxes represent the inter-quartile range, the outer 'whiskers' the 10 and 90th percentiles, and the inner line the median. The realisations are circles with dots at the centres.

are discussed by Diebold *et al.* (1999). The forecasts and realizations for 1997 and 1998 indicate a continuation of the tendency in the 1990s to both overestimate the uncertainty and level of inflation. The forecast distributions appear too dispersed and the central tendencies consistently indicate higher inflation rates than actually materialize.

Table 5.1 presents the results of the formal tests. The Kolmogorov–Smirnov test statistic of the uniformity of $\{z_t\}_{t=1}^n$ offers no evidence against the distributional assumption. The Pearson goodness-of-fit test (outlined in Section 5.5) applied to $\{z_t\}$ with $K = 4$ did not offer any evidence against uniformity at conventional significance levels. The tests of independence and normality are based on $\{z_t^*\}_{t=1}^n$. Both the 1 and 3-degree-of-freedom tests reject at the 1% level. Tests of the normality of $\{z_t^*\}$ return p-values of 0.10 and 0.01, although these tests of distribution do assume a random sample. The formal tests reject the SPF densities confirming the impression gained from the Box–Whisker plots.

Table 5.1 Tests of SPF density forecasts of inflation (1969–2002) based on p.i.ts

Null hypothesis	Test	Outcome
Distribution		
	KS test of uniformity	0.16
Independence		
	Bowman–Shenton	0.10
	Doornik–Hansen	0.01
	Berkowitz I	0.00
	Berkowitz II	0.00

Notes: The test outcomes are recorded as *p*-values, except for the KS test, which is the test statistic value, for which the 5% critical value is 0.23. The Bowman–Shenton test is a two-degree of freedom test with an asymptotic chi-squared distribution, whilst the Doornik–Hansen test may have better small-sample properties. Berkowitz I is a 1-degree-of-freedom test of no first-order autocorrelation of the transformed p.i.ts assuming $N(0, 1)$. Berkowitz II is a 3-degree-of-freedom test of zero-mean, unit-variance and no first-order autocorrelation of the transformed p.i.ts assuming normality.

5.7 Empirical illustration (II): the MPC inflation forecasts

Since August 1997, in the mid month of every quarter the Bank of England Inflation Report has contained density forecasts of RPIX inflation for the current quarter and for every quarter up to two years ahead, produced by the Monetary Policy Committee (MPC).[7] The forecasts are given analytically by the two-piece normal (2PN) distribution, and graphically by the 'rivers of blood' fan chart.[8] This is a useful way of capturing perceived asymmetries between upside and downside risks, whilst allowing probability calculations to be undertaken using the standard normal distribution. The 2PN distribution can be parameterized as $\{\mu, \sigma_1, \sigma_2\}$ where μ is the (common) mode and σ_1 and σ_2 are the standard deviations of the two normal distributions on which the 2PN is based.

Two sets of projections are included in each report, one based on the assumption that interest rates remain constant throughout the forecast period and the other based on the assumption that interest rates follow market expectations. Only for the former are projections available back to August 1997, and this is the series which we use. The RPIX inflation rate is the annual percentage growth in quarterly RPIX (RPI excluding mortgage interest payments, ONS code CHMK). The current quarter forecasts can be viewed as 1-step ahead forecasts, and the year-ahead forecasts correspond to a five-step ahead horizon.

Table 5.2 MPC one-year ahead inflation forecasts

Inflation report	Mode	σ_1	σ_2	Skew (Mean – Mode)	Outcome	z
Aug. 97	1.99	0.651	0.914	0.210	2.55	0.683
Nov. 97	2.19	0.385	1.05	0.530	2.53	0.454
Feb. 98	2.44	0.447	0.557	0.088	2.53	0.513
May 98	2.37	0.790	0.515	−0.220	2.30	0.563
Aug. 98	2.86	0.531	0.706	0.140	2.17	0.084
Nov. 98	2.59	0.554	0.717	0.130	2.16	0.190
Feb. 99	2.52	0.586	0.661	0.060	2.09	0.219
May. 99	2.23	0.533	0.671	0.110	2.07	0.335
Aug. 99	1.88	0.488	0.676	0.150	2.13	0.584
Nov. 99	1.84	0.584	0.521	−0.050	2.11	0.717
Feb. 00	2.32	0.508	0.633	0.100	1.87	0.168
May 00	2.47	0.521	0.584	0.050	2.26	0.325
Aug. 00	2.48	0.540	0.540	0.000	2.38	0.426
Nov. 00	2.19	0.531	0.594	0.050	1.95	0.309
Feb. 01	2.09	0.584	0.521	−0.050	2.37	0.721
May 01	1.94	0.584	0.521	−0.050	1.86	0.473
Aug. 01	1.96	0.550	0.550	0.000	1.98	0.509
Nov. 01	2.06	0.464	0.714	0.200	2.61	0.732
Feb. 02	2.13	0.454	0.705	0.200	2.89	0.829

Notes: These are one-year ahead forecasts, so that the forecast published in the Aug. 1997 Inflation Report, for example, relates to the third quarter of 1998. The mode μ, σ_1 and σ_2 correspond to the parameterisation of the 2PN discussed in the text. z is the probability integral transform.

The numerical values of the 2PN parameters for the $\{\mu, \sigma_1, \sigma_2\}$ parameterisation are contained in Tables 5.2 and 5.3 for the current and year-ahead forecasts, respectively. We also report the skewness of the 2PNs (calculated as mean minus mode) as well as the outcomes and the p.i.ts $\{z_t\}$. Notice that when $\sigma_1 = \sigma_2 = \sigma$ (say), the distribution is symmetric and the 2PN collapses to the $N(\mu, \sigma^2)$. When $\sigma_2 > \sigma_1$, the mean exceeds the mode and the 'upside' risks of inflation exceeding the mode outweigh the 'downside' risks.

5.7.1 Point forecast performance

Table 5.4 shows the point forecast performance of the mean of the MPC 2PN forecast densities for the current, next quarter and the year ahead. None of the forecasts are significantly biased. Wallis (2003) finds the year ahead forecasts significantly overestimate the rate of inflation at the 5% level using a one-sided t-test on a shorter data set, but the positive

Table 5.3 MPC current quarter inflation forecasts

Inflation report	Mode	σ_1	σ_2	Skew (Mean − Mode)	Outcome	z
Aug. 97	2.65	0.131	0.181	0.040	2.81	0.788
Nov. 97	2.60	0.077	0.210	0.106	2.80	0.744
Feb. 98	2.60	0.180	0.223	0.035	2.59	0.427
May 98	2.83	0.318	0.205	−0.090	2.94	0.772
Aug. 98	2.51	0.215	0.277	0.050	2.55	0.495
Nov. 98	2.54	0.166	0.216	0.040	2.53	0.414
Feb. 99	2.49	0.175	0.200	0.020	2.53	0.541
May 99	2.48	0.161	0.199	0.030	2.30	0.118
Aug. 99	2.31	0.148	0.198	0.040	2.17	0.151
Nov. 99	2.20	0.175	0.156	−0.015	2.16	0.429
Feb. 00	1.93	0.152	0.190	0.030	2.09	0.791
May 00	1.88	0.156	0.175	0.015	2.07	0.851
Aug. 00	2.38	0.162	0.162	0.000	2.13	0.059
Nov. 00	2.36	0.159	0.178	0.015	2.11	0.061
Feb. 01	1.94	0.175	0.156	−0.015	1.87	0.373
May 01	1.90	0.175	0.156	−0.015	2.26	0.990
Aug. 01	2.31	0.165	0.165	0.000	2.38	0.665
Nov. 01	2.00	0.232	0.357	0.100	1.95	0.328
Feb. 02	2.14	0.227	0.352	0.100	2.37	0.685
May 02	2.02	0.259	0.259	0.000	1.86	0.272
Aug. 02	1.84	0.253	0.253	0.000	1.98	0.704
Nov. 02	2.64	0.242	0.242	0.000	2.61	0.458
Feb. 03	2.77	0.280	0.280	0.000	2.89	0.668

Note: The entries in this table relate to current quarter (one-step ahead forecasts), so that the forecast published in the August 1997 Inflation Report, for example, is of 1997: 3.

Table 5.4 Point forecast evaluation summary statistics

Test	1-step (current qtr)	2-step (next qtr)	5-step (year ahead)
MPC forecasts			
Bias	0.007	0.024	−0.073
Standard error of bias	0.034	0.058	0.086
MSFE	0.025	0.070	0.137
No change benchmark forecasts			
Bias	0.014	0.006	−0.105
Standard error of bias	0.060	0.074	0.086
MSFE	0.078	0.115	0.144

Notes: The MPC point forecasts are taken to be the means of the 2PN densities. The current quarter no-change forecasts are based on the previous quarters' inflation rate, so that the forecast for 1997: 3, for example, is the actual rate of inflation in 1997: 2. The 2-step or next quarter no-change forecast of 1997: 4 is the recorded rate in 1997: 2, etc.

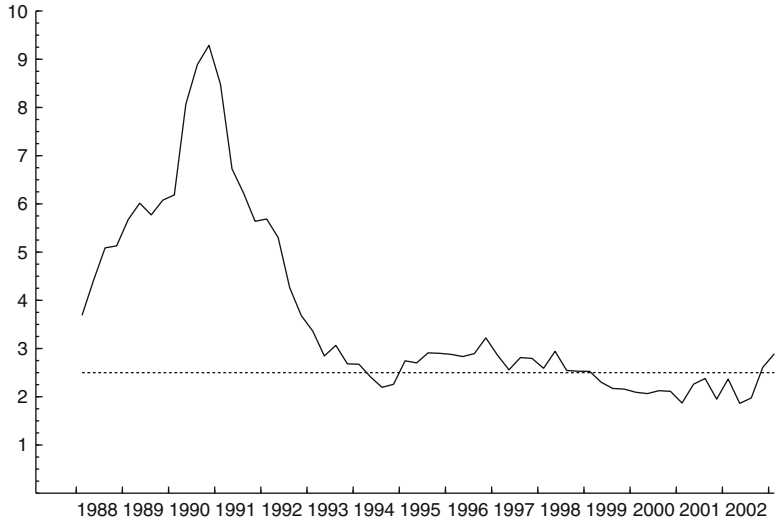


Figure 5.2 Annual rate of quarterly price inflation – RPI excluding mortgage interest payments

forecast errors of 0.35 and 0.55 for November 2002 and February 2003 respectively reduce the overall average bias to around −0.07 percentage points. The third row gives the mean squared forecast error (MSFE) associated with the point forecasts. These are hard to interpret except relative to a rival set of forecasts or a benchmark (as discussed in Section 2.3). Any one of a number of time-series models, or more structural models, could be used to generate forecasts for this purpose, but would be unlikely to pose a particularly stern test unless they adequately captured the dramatic slowdown in inflation to the historically low rates observed over the last ten years: see Figure 5.2. Given the arguments in Clements and Hendry (1999), we record the results for a 'no-change' forecast.[9]

The MPC current quarter performance is well judged against the benchmark. The MPC forecasts are significantly more accurate than the no-change forecasts using the Harvey *et al.* (1997) small-sample modifications to the Diebold and Mariano (1995) statistic. A p-value of 0.013 was obtained for a one-sided test of the null of equal forecast accuracy versus the alternative that the MPC forecasts are more accurate. At a year ahead the MPC's MSFE is similar to that of the unconditional benchmark. The relatively poor performance of the no-change forecasts of the current quarter compared to those of the MPC is consistent with the

finding of Montgomery *et al.* (1998). They show that using information on the first month in a quarter to forecast that quarter may significantly improve forecast accuracy. The MPC forecasts make use of current quarter information although the no-change forecasts do not.

5.7.2 Evaluation of forecast densities

Figure 5.3 presents time series and histograms (of the quartiles) for the p.i.ts for the MPC forecasts. The results of applying the testing procedures outlined in Section 5.2 are recorded in Table 5.5.

There is no evidence against the MPC current or next quarter forecasts. The year-ahead forecasts are rejected by the Berkowitz test of zero mean and unit variance of the transformed p.i.ts. This finding for the year-ahead forecasts is consistent with Wallis (2003), who finds that too much probability mass is put on relatively high inflation rates: from Figure 5.3 it is apparent that only the last z-value is in excess of 0.75, whereas we would expect 25% of the $\{z_t\}$ values to exceed 0.75. The forecasts are unduly pessimistic one year-ahead. Recall from the evaluation of the

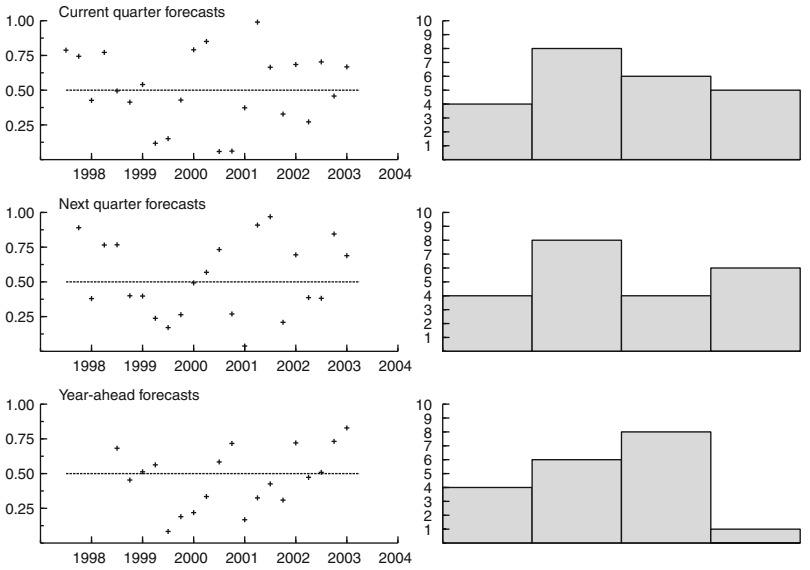


Figure 5.3 Probability integral transforms of the MPC two-piece normal density forecasts of the quarterly annual inflation rate

Note: The figures in the left column are time-series plots, those in the right column are the corresponding histograms.

Table 5.5 Probability integral transform-based testing of inflation density forecasts

Test	1-step (current qtr)	2-step (next qtr.)	5-step (year ahead)
Bowman–Shenton normality	0.850	0.913	0.759
Doornik–Hansen normality	0.252	0.893	0.859
Berkowitz I	0.694	0.568	0.025
Berkowitz II	0.774	0.745	0.031

Notes: The entries in the table are *p*-values. The first Berkowitz test is a 2-degree-of-freedom test of zero mean and unit variance, with a maintained hypothesis of normality. The second is a 3-degree-of-freedom test of zero-mean, unit variance and zero first-order autocorrelation, with a maintained hypothesis of normality of the inverse-normal cdf transformation of the p.i.ts.

point forecasts that the year ahead forecasts appear to be unbiased, so that the rejection on the Berkowitz test is not due to the forecast densities being incorrectly centred.

5.8 Model-based density evaluation

So far in this chapter, we have considered the evaluation of sequences of density forecasts when the forecasts are not explicitly derived from a given model. The forecast densities are *given*: in the case of the SPF probability distributions, as histograms; for the Bank of England inflation forecasts, specified as 2PN distributions with particular parameter values. A natural question is how to evaluate whether the forecast densities from a particular conditional parametric model are correctly specified, where the model is defined up to an unknown vector, $\boldsymbol{\theta}$. That is, the parametric family of conditional densities is given by, say,

$$\{f(y \mid x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

so the question is whether there is an admissible $\boldsymbol{\theta}$ such that $f(y \mid x, \boldsymbol{\theta})$ equals the true conditional distribution of Y given X. The unknown parameter vector $\boldsymbol{\theta}$ is replaced by an estimator $\hat{\boldsymbol{\theta}}$ which is assumed to be a '$\sqrt{T}$'-consistent estimator of the true value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_0$, when the null is true. The evaluation procedure will need to take into account the effect of parameter estimation error, as discussed in Section 2.4.2 in the context of

evaluating point forecasts. We assume strict stationarity, so that the same conditional density holds at all $t = 1, \ldots, T$. Recall that in Section 5.7 the 'unconditional' 2PN distributions that constituted the Bank of England inflation forecasts had parameter values that depended on t, and that in the case of the SPF probability distributions (Section 5.6) not even the form of the distribution was assumed constant over time.

Andrews (1997) introduces a specification test for independent observations, and Corradi and Swanson (2003) discuss the implementation of this test in a time-series context (where the observations are generally dependent). To make matters concrete, suppose the null hypothesis is that the family of conditional distributions are generated by an AR(1) with Gaussian disturbances. That is, the parametric model is given by:

$$y_t = \alpha_1 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}, \ N(0, \sigma_\varepsilon^2)$$

so that the conditional distributions are given by:

$$y_t \mid y_{t-1} \sim N(\alpha_1 y_{t-1}, \sigma_\varepsilon^2)$$

under the null, or:

$$\begin{aligned} F(y \mid y_{t-1}, [\alpha_1, \sigma_\varepsilon]) &\equiv \Pr(y_t < y \mid y_{t-1}, [\alpha_1, \sigma_\varepsilon]) \\ &= \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \int_{-\infty}^{y} \exp\left(\frac{(y_t - \alpha_1 y_{t-1})^2}{2\sigma_\varepsilon^2} \right) dy_t. \end{aligned} \tag{5.7}$$

Under the null that the data are generated by a Gaussian AR(1), α_1 is consistently estimated by the maximum likelihood estimator (MLE) (conditional on y_0):

$$\hat{\alpha}_1 = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}$$

and σ_ε^2 by:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{\alpha}_1 y_{t-1})^2.$$

The Conditional Kolmogorov (CK) statistic of Andrews (1997) compares the empirical joint distribution function (EDF) of $\{y_t, y_{t-1}; t = 1, \ldots, T\}$

with the product of the conditional distribution given by the parametric model and the EDF of $\{y_{t-1}, t = 1, \ldots, T\}$. That is:

$$\mathrm{CK}_T = \sqrt{T} \sup_{u \times v \in U \times V} \left| \hat{H}_T(u, v) - \hat{F}_T(u, v) \right|.$$

$\hat{H}_T(u, v)$ denotes the joint EDF of $\{y_t, y_{t-1}\}$:

$$\hat{H}_T(u, v) = \frac{1}{T} \sum_{t=1}^{T} 1(y_t \leq u) 1(y_{t-1} \leq v),$$

and $\hat{F}_T(u, v)$ is the 'semi-*parametric*/semi-*empirical*' distribution function of $\{y_t, y_{t-1}\}$ defined by:

$$\hat{F}_T(u, v) = \hat{F}(u \mid y_{t-1}, [\hat{\alpha}_1, \hat{\sigma}_\varepsilon]) \times \hat{G}_T(v).$$

Here, $\hat{F}(u \mid y_{t-1}, [\hat{\alpha}_1, \hat{\sigma}_\varepsilon])$ is the *parametric* conditional distribution from (5.7) (with $[\alpha_1, \sigma_\varepsilon]$ replaced by the MLEs $[\hat{\alpha}_1, \hat{\sigma}_\varepsilon]$):

$$\hat{F}(u \mid y_{t-1}, [\hat{\alpha}_1, \hat{\sigma}_\varepsilon]) \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}\hat{\sigma}_\varepsilon} \exp\left(\frac{(y_t - \hat{\alpha}_1 y_{t-1})^2}{2\hat{\sigma}_\varepsilon^2} \right) dy_t,$$

and $\hat{G}_T(v)$ is the *empirical* DF of $\{y_{t-1}, t = 1, \ldots, T\}$:

$$\hat{G}_T(v) = \frac{1}{T} \sum_{t=1}^{T} 1(y_{t-1} \leq v).$$

Thus:

$$\hat{F}_T(u, v) = \frac{1}{T} \sum_{t=1}^{T} \hat{F}(u \mid y_{t-1}, [\hat{\alpha}_1, \hat{\sigma}_\varepsilon]) 1(y_{t-1} \leq v).$$

The test statistic can then be written as:

$$\begin{aligned} \mathrm{CK}_T &= \sqrt{T} \sup_{u \times v \in U \times V} \left| \hat{H}_T(u, v) - \hat{F}_T(u, v) \right| \\ &= \sup_{u \times v \in U \times V} \left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [1(y_t \leq u) - \hat{F}(u \mid y_{t-1}, [\hat{\alpha}_1, \hat{\sigma}_\varepsilon])] 1(y_{t-1} \leq v) \right|. \end{aligned}$$

It is based on the maximum difference between $\hat{H}_T(u, v)$ and $\hat{F}_T(u, v)$.[10] Andrews (1997) shows that the asymptotic null distribution of the CK statistic depends upon nuisance parameters, so that general-purpose critical values cannot be tabulated, and instead proposes a parametric bootstrap. Corradi and Swanson (2003) discuss bootstrap techniques relevant for time-series models.

An alternative approach by Li and Tkacz (2002) compares a kernel estimate of the true conditional density function to the model's parametric conditional density function. They obtain a test statistic which has a limiting standard normal distribution under the null of correct specification, although Monte Carlo evidence suggests that bootstrapped critical values are likely to prove more reliable for inference.

Finally, if we wished to test the null that the AR(1) model conditional forecast densities were correctly specified for a given $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, say (i.e., $[\alpha_1, \sigma_\varepsilon] = [0.9, 1]$) then we could simply calculate:

$$z_t = F(y_t \mid y_{t-1}, [0.9, 1]) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_t} \exp\left(\frac{1}{2}(w - 0.9y_{t-1})^2\right) dw$$

for $t = 1, \ldots, n$ and apply the tests outlined in Section 5.2 to the $\{z_t\}_{t=1}^n$.

5.8.1 Model mis-specification

The approach to density specification testing in Corradi and Swanson (2003) allows for dynamic mis-specification under the null hypothesis, in the sense that the test is for correct specification *given a particular information set*, rather than being a test of correct specification with all the relevant history included. As an example, consider testing the AR(1) model density forecasts described above, when the data generating process is actually a second-order Gaussian AR model:

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + v_t, \quad v_t \sim N(0, \sigma_v^2). \tag{5.8}$$

The true conditional densities are $y_t \mid y_{t-1}, y_{t-2} \sim N(\gamma_1 y_{t-1} + \gamma_2 y_{t-2}, \sigma_v^2)$. Nevertheless, if we take the information set to be y_{t-1} (rather than y_{t-1} and y_{t-2}) the AR(1) model densities, $y_t \mid y_{t-1}$, are still 'well specified', given the properties of the normal distribution, in that $y_t \mid y_{t-1}$ for (5.8) has a normal distribution, but note that α_1 and σ_ε in $y_t \mid y_{t-1} \sim N(\alpha_1 y_{t-1}, \sigma_\varepsilon^2)$ will not equal γ_1 and σ_v (e.g., α_1 will be the first-order autocorrelation coefficient for an AR(2)). Ignoring parameter estimation uncertainty, the $\{z_t\}$ for the $y_t \mid y_{t-1}, y_{t-2}$ densities will be i.i.d. $U(0, 1)$, while the $\{z_t\}$ for the $y_t \mid y_{t-1}$ densities (with mean and variance derived

from the AR(2)) will be $U(0,1)$ but not i.i.d. because of the dynamic mis-specification. Corradi and Swanson (2003) outline the derivation of appropriate bootstrap critical values for CK-type tests that allow for dynamic mis-specification under the null.

5.9 Summary

This chapter reviews some of the recent literature on the evaluation of density (or histogram) forecasts. Whilst an interval forecast posits a range and a probability of the actual falling within that range, a density forecast gives a complete description of the probabilities attached to all possible values or ranges of values of the outcome variable. The key tool for the evaluation of survey-based histograms is the probability integral transform. This is described and used in an evaluation of the SPF US inflation probability distributions and the Bank of England inflation density forecasts

We also describe the application of this approach to the evaluation of joint probability distribution functions. Calibration is discussed – the adjustment of future probability assessments prompted by systematic past errors. A relationship between interval and density forecast evaluation is established by showing that certain interval forecast evaluation techniques can be extended to provide 'contingency-table-type' goodness-of-fit tests of densities.

Finally, we consider a different problem in density evaluation, which is whether an observed sample comes from a particular parametric density defined up to an unknown parameter vector. This approach to the evaluation of model-based conditional forecast densities contrasts with the evaluation of the survey-based SPF histograms and the Bank densities. For both the SPF and Bank forecasts the question is whether a sequence of histograms (SPF), or parametric densities with given location and scale parameters (Bank), is consistent with an associated sequence of outcomes.

5.10 Appendix: multivariate forecast density probability integral transform tests

Let Z_1 and Z_2 be independent $U(0,1)$ random variables. We begin by deriving the distribution function for their product. Because of independence, the joint distribution function $F_{Z_1Z_2}$ is the product of the distribution functions of Z_1 and Z_2, $F_{Z_1Z_2}(z_1, z_2) = z_1z_2$, and $f_{Z_1Z_2} = f_{Z_1}f_{Z_2} = 1$.

Using a change of variables:

$$\begin{aligned} Z_1^* &= Z_1 Z_2 \\ Z_2^* &= Z_2 \end{aligned}$$

for which the determinant of the Jacobian for the inverse transformation is:

$$J = \det \frac{\partial(Z_1, Z_2)}{\partial(Z_1^*, Z_2^*)} = \begin{vmatrix} \frac{1}{Z_2^*} & -\frac{Z_1^*}{Z_2^{*2}} \\ 0 & 1 \end{vmatrix} = \frac{1}{Z_2^*}$$

the joint density function of (Z_1^*, Z_2^*) is then:

$$f_{Z_1^* Z_2^*} = f\left(\frac{Z_1^*}{Z_2^*}, Z_2^*\right) \times \frac{1}{Z_2^*} = \frac{1}{Z_2^*}, \tag{5.9}$$

where $0 < Z_1^* < Z_2^* < 1$.

Since Z_1^* is the random variable of interest, integrating Z_2^* out of $f_{Z_1^* Z_2^*}$ over the permissible range gives:

$$f_{Z_1^*} = \int_{Z_1^*}^{1} Z_2^{*-1} dZ_2^* = [\ln Z_2^*]_{Z_1^*}^{1} = -\ln Z_1^*.$$

The distribution function is:

$$F_{Z_1^*} = Z_1^* - Z_1^* \ln Z_1^*, \quad 0 < Z_1^* < 1.$$

The probability integral transform of Z_1^* with respect to $f_{Z_1^*}$, that is:

$$z_t^* = \int_0^{z_t^*} f_{Z_1^*}(u) du \quad t = 1, \ldots, n \tag{5.10}$$

yields an i.i.d. $U(0, 1)$ sequence under the null. Thus, we need simply to evaluate $F_{Z_1^*}$ at $Z_{1t} \times Z_{2t}$, $t = 1, \ldots, n$, and test whether $\{z_t^*\}_{t=1}^n$ constitutes an i.i.d. sample from a $U(0, 1)$, using the methods described in Section 5.2.

Now consider the ratio. Again using a change of variables:

$$\begin{aligned} Z_1^* &= Z_1 / Z_2 \\ Z_2^* &= Z_2 \end{aligned}$$

for which the determinant of the Jacobian for the inverse transformation is:

$$J = \det \frac{\partial(Z_1, Z_2)}{\partial(Z_1^*, Z_2^*)} = \begin{vmatrix} Z_2^* & Z_1^* \\ 0 & 1 \end{vmatrix} = Z_2^*$$

the joint density function of (Z_1^*, Z_2^*) is then:

$$f_{Z_1^* Z_2^*} = f\left(\frac{Z_1^*}{Z_2^*}, Z_2^*\right) \times Z_2^* = Z_2^*, \tag{5.11}$$

where $0 < Z_2^* < 1$, $0 < Z_1^* < \infty$.

Since Z_1^* is the random variable of interest, integrating Z_2^* out of $f_{Z_1^* Z_2^*}$ over the permissible range gives:

$$f_{Z_1^*} = \int_0^1 Z_2^* dZ_2^* = \frac{1}{2},$$

when $Z_1^* < 1$ or:

$$f_{Z_1^*} = \int_0^{1/Z_1^*} Z_2^* dZ_2^* = \frac{1}{2(Z_1^*)^2},$$

when $Z_1^* > 1$. The distribution function is:

$$F_{Z_1^*} = \begin{cases} Z_1^*/2, & 0 < Z_1^* < 1, \\ 1 - (1/2Z_1^*), & 1 < Z_1^* < \infty. \end{cases}$$

6
Decision-based Evaluation

6.1 Introduction

Forecasts are generally made for a purpose. If we suppose an environment whereby agents make decisions (equivalently, select actions) based on a particular forecast, then we can evaluate that forecast in terms of its expected economic value (equivalently, expected loss), where the expectation is calculated using the actual probabilities of the states of nature. Typically, we might expect users to have different economic value (or loss) functions, so that the actions and expected losses induced by two rival sets of forecasts need not be such that each user's expected economic value is maximized by the same set of forecasts. In Section 6.2 we show following Diebold *et al.* (1998)[1] that only when a density forecast coincides with the true conditional density will it be optimal (in the sense of maximizing economic value) for all users regardless of their loss functions. This is a compelling reason to assess how well the forecast distribution matches the actual distribution, as in Section 5.2 – a forecast density that provides a close match to the true density can be used by all with equanimity, no matter what their individual loss functions. For decision-based evaluation in general we require the whole forecast density. Only in exceptional circumstances will the minimum mean squared forecast error (MSFE) point forecast be sufficient to generate optimal decisions and maximize economic value – see Section 6.3.

The importance of the principle of decision-based assessment of forecasts has been widely accepted for a long time. That said, there is little decision-based forecast evaluation in macroeconomics, although it is common in meteorology (see, e.g., Katz and Murphy (1997)), and occurs in empirical finance (see, e.g., Leitch and Tanner (1991, 1995)).

As will become evident from the description of the simple two-state, two-action decision problem in Section 6.4, the implementation of this approach to forecast evaluation requires information that may not be readily available. For example, a full specification of the decision problem is needed, including a mapping from forecasts to decisions (or actions), and quantification of the economic costs and benefits emanating from those actions in different states of nature. Moreover, when we consider the behaviour of the government, as in the case of the Central Bank making decisions on interest rates, actions may affect the probabilities of outcomes. In meteorological applications, the likelihoods of different outcomes will usually be independent of agents' actions.[2] In Section 6.5 we consider how the simplest decision problem in Section 6.4 might be generalized to make it relevant for an analysis of the Bank of England's inflation targeting and interest-rate setting problem. Section 6.6 examines the relationship between economic value and simple commonly used measures of forecast performance, such as the LPS and QPS scores. Section 6.7 calculates these measures for a simplified version of the decision problem faced by the Bank of England.

In the decision approach, the loss function depends on actions and the state of nature. Because the (optimal) action depends on the forecast, this induces a loss function defined on the forecast and the state of nature. By highlighting the linkage between actions and forecasts, and the resulting dependence of the loss function on the context, the decision approach leads us to question the general reliance on squared-error loss, when obtaining the properties of optimal forecasts, for example. We saw in Section 3.3 that optimal forecasts will be biased when the loss function is asymmetric, and that in the specific case of linex loss, an analytic expression for the bias can be obtained in terms of the degree of asymmetry and the conditional forecast error variance. As noted by Patton and Timmermann (2003), for example, this suggests that one should be wary of testing for 'forecast rationality' by testing for the properties that optimal forecasts have under squared-error loss. That optimal forecasts may be biased is perhaps unsurprising, but as Patton and Timmermann (2003) show, a number of other properties, such as that h-step forecasts will be correlated of order at most $h - 1$ (see Section 2.1.1), may not characterize optimal forecasts allowing for general loss functions and data generating processes. Given that the form of the loss function may not be known to the econometrician, in Section 6.8 we consider testing forecast rationality allowing for general loss functions.

6.2 Decision-based evaluation – some basic results

Diebold *et al*. (1998) consider a user with a loss function $L_1(a, y)$, where a refers to an action, and y is the random variable $y \sim f(y)$; $f(y)$ is unknown to the user. The user chooses an action to minimize their expected loss over all possible states (here, values of y) based on a forecast of the probabilities of the states, $p_1(y)$. Here y is a continuous random variable so that there is a continuum of states. (In Sections 6.4 and 6.5 we illustrate with a few discrete states). Thus a_1^* is chosen such that:

$$a_1^* = \underset{a \in A}{\operatorname{argmin}} \int L_1(a, y) p_1(y) dy. \tag{6.1}$$

That is, given $p_1(y)$, a_1^* is the action in the admissible set A which minimizes the user's expected loss. Substituting a_1^* in the loss function gives loss $L_1(a_1^*, y)$, a random variable as it depends on y. The loss that will eventuate on average is determined by the actual density for y, $f(y)$, that is:

$$E[L_1(a_1^*, y)] = \int L_1(a_1^*, y) f(y) dy. \tag{6.2}$$

The dependence of the optimal action a^* on $L()$ is apparent. For L_1 (and $p_1(y)$) we obtain a_1^*, and expected loss given by (6.2). For the same forecasts $p_1(y)$ changing the loss function will typically change the optimal action, and for the same loss function $L_1()$, changing $p(y)$ will typically give rise to a different a^* and value of (6.2). For a given loss function and forecasts $p_1(y)$ and $p_2(y)$, the better forecast will be that with a lower value of (6.2).

Formally, $p_1(y)$ is preferred to $p_2(y)$ if:

$$E[L_1(a_1^*(p_1(y)), y)] < E[L_1(a_2^*(p_2(y)), y)],$$

where the notation $a_i^*(p_i(y))$ makes it plain that a_i^* is the optimal action for forecast $p_i(y)$.

However, for arbitrary density forecasts $p_1(y)$ and $p_2(y)$ we can always find a loss function $L_2()$ such that:

$$E[L_2(a_1^*(p_1(y)), y)] > E[L_2(a_2^*(p_2(y)), y)],$$

establishing that we are in general unable to establish a ranking of $p_1(y)$ and $p_2(y)$ with which all users will agree, regardless of their loss functions.

Diebold *et al.* (1998, p. 866) illustrate for specific $f(y)$, $p_i()$ and $L_i()$, $i = 1, 2$). This suggests that it might in general be difficult to rank density forecasts without recourse to a specific loss function.

The key result stated in the introduction is that if $p_1(y) = f(y)$ (the first forecast density coincides with the true density) then:

$$E[L(a^*(f(y)), y)] < E[L(a_2^*(p_2(y)), y)].$$

The optimal action with respect to $f(y)$ will have smaller expected loss than the optimal action for all other forecast densities, here given by $p_2(y)$ (not equal to $f(y)$) for *all* loss function. That is, $p_1(y) = f(y)$ will be ranked first by all users irrespective of their loss functions.

6.3 Quadratic loss functions

When the cost function is quadratic, the optimal decision depends only on the conditional expectation of the variable being forecast rather than the whole forecast density. Further, economic value, the evaluation criterion in the decision-based context, is proportional to the MSFE criterion. In these circumstances, the minimum MSFE point forecast is sufficient to generate optimal decisions and maximize economic value. This result holds when there are constraints but these are linear, giving rise to the Linear-Quadratic (LQ) decision problem, but here we illustrate without any constraints. In general, though, decision-based forecast evaluation criteria will not coincide with traditional statistical measures of forecast accuracy, such as MSFE.

Consider a quadratic loss function:

$$L(a_t, y_{t+1}) = c_1 a_t^2 + 2c_2 a_t y_{t+1} + c_3 y_{t+1}^2, \tag{6.3}$$

where relative to Section 6.2 we have time-dated the state variable y and the decision variable a to make it explicit that at time t a forecast will be required of y_{t+1}. The parameters c_1, c_2 and c_3 of the loss function must satisfy $c_1 > 0$ and $c_1 c_3 - c_2^2 > 0$ to ensure $L()$ is globally convex in both its arguments. We need to solve for a_t^* such that:

$$a_t^* = \underset{a_t \in A}{\operatorname{argmin}} \int L(a_t, y_{t+1}) p_t(y) dy \tag{6.4}$$

$$= \underset{a_t \in A}{\operatorname{argmin}} \int (c_1 a_t^2 + 2c_2 a_t y_{t+1} + c_3 y_{t+1}^2) p_t(y) dy, \tag{6.5}$$

where $p_t(y)$ is the forecast density of y_{t+1} based on period t information. We can take the derivative of $L(a_t, y_{t+1})$ within the integral (i.e., expectations operator) to obtain the first-order condition:

$$\int \frac{\partial L(a_t, y_{t+1})}{\partial a_t} p_t(y)\, dy = 0,$$
$$\int (2c_1 a_t + 2c_2 y_{t+1}) p_t(y)\, dy = 0,$$

which gives $a_t^* = -(c_2/c_1) \int y_{t+1}\, p_t(y) dy = -(c_2/c_1)\hat{E}_t(y_{t+1})$, where $\hat{E}_t(y_{t+1})$ denotes the 1-step ahead conditional expectation of y_{t+1} based on the *forecast* distribution.

Substituting a_t^* in (6.3) gives:

$$\begin{aligned} L(a_t^*, y_{t+1}) &= \frac{c_2^2}{c_1}\hat{E}_t(y_{t+1})^2 - 2\frac{c_2^2}{c_1}\hat{E}_t(y_{t+1})y_{t+1} + c_3 y_{t+1}^2 \\ &= \left(c_3 - \frac{c_2^2}{c_1}\right) y_{t+1}^2 + \frac{c_2^2}{c_1}(y_{t+1} - \hat{E}_t(y_{t+1}))^2. \end{aligned}$$

The expected loss is determined by the actual density for y_{t+1}, $f_{t+1}(y)$, that is (dropping the subscript on f for convenience):

$$\begin{aligned} E[L(a_t^*, y_{t+1})] &= \int L(a_t^*, y_{t+1}) f(y) dy \\ &= \int \left[\left(c_3 - \frac{c_2^2}{c_1}\right) y_{t+1}^2 + \frac{c_2^2}{c_1}(y_{t+1} - \hat{E}_t(y_{t+1}))^2 \right] f(y) dy \\ &= \left(c_3 - \frac{c_2^2}{c_1}\right) \int y_{t+1}^2 f(y) dy + \frac{c_2^2}{c_1} \int (y_{t+1} - \hat{E}_t(y_{t+1}))^2 f(y) dy. \end{aligned}$$

The first term is $(c_3 - (c_2^2/c_1)) > 0$ times the second (uncentred) moment of y_{t+1}, and so does not depend on the $p_t(y)$. The second term $\int (y_{t+1} - \hat{E}_t(y_{t+1}))^2 f(y) dy$ is the expected squared-error loss, so that expected economic loss is proportional to the MSFE. Thus, an alternative forecast density $\tilde{p}_t(y)$ would be preferred (worse) on economic value if the corresponding point forecast ($\int y_{t+1} \tilde{p}_t(y) dy$) was preferred (worse) in terms of MSFE. That is, we can decide between $p_t(y)$ and $\tilde{p}_t(y)$ by considering the MSFEs of the associated point forecasts, without needing to know the parameters of the quadratic loss function (c_1, c_2 and c_3).

Table 6.1 Payoff matrix for a two-state, two-action decision problem

		States (s_{t+1})	
		Bad ($s_{t+1} = 1$)	Good ($s_{t+1} = 0$)
Decisions (y_t)	Yes ($y_t = 1$)	$U_{t+1,by}$	$U_{t+1,gy}$
	No ($y_t = 0$)	$U_{t+1,bn}$	$U_{t+1,gn}$

6.4 Two-state, two-action decision problems

The ideas in the previous section can be illustrated with a simple example where there are two states and two actions. Our treatment follows Granger and Pesaran (2000a, b) and Pesaran and Skouras (2002), although the analysis is standard. The two possible states in period $t+1$ are 'Bad' ($s_{t+1} = 1$) and 'Good' ($s_{t+1} = 0$), and there are two possible actions open to the decision-maker in period t. To take action, 'Yes', indicated by $y_t = 1$, or to decline to take action, 'No', $y_t = 0$. Thus, actions are taken in advance. The payoff matrix associated with this decision problem is given in Table 6.1. In the Payoff Matrix, $U_{t+1,by}$ represents the decision maker's utility if the *b*ad state occurs after the *y*es decision is taken, and so on.

Given this Payoff Matrix, which maps combinations of outcomes (or realized states) and (prior) actions to economic values (U), what is the value at period t to the decision maker of a particular forecast probability of the Bad event occurring in period $t+1$? Let π_{t+1} be the actual probability that $s_{t+1} = 1$, $\pi_{t+1} = \Pr(s_{t+1} = 1)$, and $\hat{\pi}_{t+1}$ the forecast probability.[3] We assume that the probabilities are independent of actions. Then, the expected utility of taking action ($y_t = 1$) based on the forecast probabilities is given by:

$$U_{t+1,by}\hat{\pi}_{t+1} + U_{t+1,gy}(1 - \hat{\pi}_{t+1}) \tag{6.6}$$

and of not acting:

$$U_{t+1,bn}\hat{\pi}_{t+1} + U_{t+1,gn}(1 - \hat{\pi}_{t+1}). \tag{6.7}$$

For the forecast probabilities of the two states given by $\hat{\pi}_{t+1}$ and $(1-\hat{\pi}_{t+1})$, action will be taken if (6.6) exceeds (6.7), that is, if:

$$\hat{\pi}_{t+1} > q_{t+1}$$

where[4]

$$q_{t+1} = \frac{\delta_{t+1,g}}{\delta_{t+1,g} + \delta_{t+1,b}},$$

and

$$\delta_{t+1,g} = U_{t+1,gn} - U_{t+1,gy}, \qquad \delta_{t+1,b} = U_{t+1,by} - U_{t+1,bn}.$$

Thus, the link between the decision (y_t) and the forecast of the bad state ($\hat{\pi}_{t+1}$) is given by:

$$y_t^* = 1(\hat{\pi}_{t+1} > q_{t+1}),$$

so action will be taken ($y_t = 1$) if $\hat{\pi}_{t+1} > q_{t+1}$.

The economic benefit that accrues at period $t+1$ will depend on which state materializes and the action taken at period t:

$$\begin{aligned} \upsilon_{t+1}(y_t; s_{t+1}) = {} & U_{t+1,by} s_{t+1} y_t + U_{t+1,gy}(1 - s_{t+1}) y_t \\ & + U_{t+1,bn} s_{t+1}(1 - y_t) + U_{t+1,gn}(1 - s_{t+1})(1 - y_t). \end{aligned}$$

Using the optimal decision rule:

$$\begin{aligned} \upsilon_{t+1}(y_t^*; s_{t+1}) = {} & U_{t+1,by} s_{t+1} y_t^* + U_{t+1,gy}(1 - s_{t+1}) y_t^* \\ & + U_{t+1,bn} s_{t+1}(1 - y_t^*) + U_{t+1,gn}(1 - s_{t+1})(1 - y_t^*). \end{aligned} \tag{6.8}$$

If we substitute $y_t^* = 1(\hat{\pi}_{t+1} > q_{t+1})$, we obtain υ_{t+1} as a function of the forecast probability, $\upsilon_{t+1}(\hat{\pi}_{t+1}; s_{t+1})$:

$$\upsilon_{t+1}(\hat{\pi}_{t+1}; s_{t+1}) = a_{t+1} + b_{t+1}(s_{t+1} - q_{t+1}) 1(\hat{\pi}_{t+1} > q_{t+1}), \tag{6.9}$$

where $a_{t+1} = s_{t+1} U_{t+1,bn} + (1 - s_{t+1}) U_{t+1,gn}$ and $b_{t+1} = U_{t+1,by} - U_{t+1,bn} + U_{t+1,gn} - U_{t+1,gy}$. Notice that the part of the economic value given by the term a_{t+1} does not depend on the probability forecast estimate, $\hat{\pi}_{t+1}$, and can be ignored when comparing two or more rival forecast probabilities (say, $\hat{\pi}_{t+1}$ and $\tilde{\pi}_{t+1}$).

The expected economic value of using the probability forecast $\hat{\pi}_{t+1}$ is given by

$$\begin{aligned} & E[\upsilon_{t+1}(\hat{\pi}_{t+1}; s_{t+1}) \mid \Omega_t] = E(a_{t+1} \mid \Omega_t) \\ & \quad + b_{t+1}(\pi_{t+1} - q_{t+1}) 1(\hat{\pi}_{t+1} > q_{t+1}), \end{aligned} \tag{6.10}$$

where expectations are taken with respect to the true conditional probability distribution of s_{t+1}, denoted by $E(\cdot \mid \Omega_t)$, and $\pi_{t+1} = E(s_{t+1} \mid \Omega_t) = \Pr(s_{t+1} = 1 \mid \Omega_t)$, $1 - \pi_{t+1} = \Pr(s_{t+1} = 0 \mid \Omega_t)$. In our set-up, the optimal decision is that $y_t = 1$ whenever the true probability of the Bad state, π_{t+1}, exceeds q_{t+1}, and $y_t = 0$ whenever $\pi_{t+1} < q_{t+1}$. So any forecast $\hat{\pi}_{t+1}$ that satisfies $\hat{\pi}_{t+1} > q_{t+1}$ when $\pi_{t+1} > q_{t+1}$, and $\hat{\pi}_{t+1} < q_{t+1}$ when $\pi_{t+1} < q_{t+1}$, belongs to the optimal set. But only when $\hat{\pi}_{t+1} = \pi_{t+1}$, is it the case that correct decisions will be made on the basis of $\hat{\pi}_{t+1}$ regardless of the relative values of $\delta_{t+1,g}$ and $\delta_{t+1,b}$, illustrating the point made in Section 6.2 (because different values of $\delta_{t+1,g}$ and $\delta_{t+1,b}$ can be viewed as defining different loss functions).

6.5 Decision problem for inflation-targeting and interest rate setting

Suppose a central bank seeks to maintain the rate of inflation in a certain band. There are three events: inflation falls below the band ($s_{t+1} = 1$), within the band ($s_{t+1} = 2$) and above the band ($s_{t+1} = 3$). There are also three decisions/actions to be made in the previous period: cut interest rates ($y_t = 1$), leave interest rates unchanged ($y_t = 2$) and raise rates ($y_t = 3$). We assume that the actions are taken in period t in order to influence the likelihoods of the three states, so that we are immediately at odds with the assumption in Section 6.4 that s_{t+1} does not depend on y_t. The payoff matrix associated with this decision problem is given in Table 6.2, where $U_{t+1,ij}$ is the payoff when $y_t = i$ and $s_{t+1} = j$ eventuates.

As in the two-state, two-action case, combinations of outcomes and actions are thus mapped into economic values (U). When the probabilities of the states depend on the actions, we define the true joint

Table 6.2 Payoff matrix for a three-state, three-action decision problem

		States		
		Below ($s_{t+1} = 1$)	Within ($s_{t+1} = 2$)	Above ($s_{t+1} = 3$)
Decisions	Cut ($y_t = 1$)	$U_{t+1,11}$	$U_{t+1,12}$	$U_{t+1,13}$
	Unchanged ($y_t = 2$)	$U_{t+1,21}$	$U_{t+1,22}$	$U_{t+1,23}$
	Raise ($y_t = 3$)	$U_{t+1,31}$	$U_{t+1,32}$	$U_{t+1,33}$

probabilities as:

$$\pi_{t+1,ij} = \Pr(y_t = i \text{ and } s_{t+1} = j).$$

To apply the decision approach, we require nine forecast probabilites:

$$\hat{\pi}_{t+1,ij} = \Pr(y_t = i \text{ and } s_{t+1} = j).$$

The expected utility of $y_t = i$, given these forecast probabilities, will be:

$$EU_{t+1,i} = \sum_{j=1}^{3} U_{t+1,ij}\hat{\pi}_{t+1,ij}$$

for $i = 1, 2, 3$. The optimal action (conditional on $\{\hat{\pi}_{t+1,ij}\}$) will be $y_t = i$ for that i which is the maximum of $\{EU_{t+1,i}\}$, $i = 1, 2, 3$. We can denote this value of y_t as y_t^*, as in Section 6.4. The economic value that accrues in $t + 1$ from y_t^* is given by:

$$v_{t+1}(y_t^*; s_{t+1}) = \sum_{j=1}^{3} 1(s_{t+1} = j)U_{t+1,y_t^*,j},$$

where the dependence of y_t^* on $\{\hat{\pi}_{t+1,ij}\}$ is implicit. The expected economic value of using these forecasts (evaluated using the true probabilities) is given by:

$$E[v_{t+1}(y_t^*; s_{t+1})] = E\left(\sum_{j=1}^{3} 1(s_{t+1} = j)U_{t+1,y_t^*,j}\right) \tag{6.11}$$

$$= \sum_{j=1}^{3} (E[1(s_{t+1} = j)])U_{t+1,y_t^*,j}, \tag{6.12}$$

where the $E[1(s_{t+1} = j)]$ are the actual probabilities of $s_{t+1} = j$ ($j = 1, 2, 3$) *given* y_t^*, because the probabilities depend on the actions taken.

The informational requirements for a decision-based evaluation of the Bank of England inflation density forecasts (described in Section 5.7) along the lines outlined in this section make such a task infeasible. For example, we can derive forecast probabilities of the states, but these are perhaps best viewed as being conditional on the interest rate changes made at the times the forecasts are released.[5] Moreover, it is difficult to

assign values to the nine U's in Table 6.2. Some progress is possible if the problem is shoe-horned into the two–state, two-action framework of Section 6.4.

We now outine how that might be done. In principle, the MPC sets the base rate to bring about a satisfactory profile of inflation up to the medium term, two year ahead horizon, taking into account likely developments in the economic environment. Given the delays in the effects of changes in monetary policy impacting on the economy, and especially the rate of inflation *excluding* mortgage interest payments, one might treat the short-horizon (say, current and next quarter) actual probabilities and forecast probabilities as being independent of changes in the Repo rate. Second, we assume that there are just two states. The Good state is defined by inflation being below the target rate of $2\frac{1}{2}\%$ ($s_{t+1} = 0$). Then the standard two-state, two-action decision-based evaluation framework becomes applicable. If we assume that the MPC's credibility may be less harmed when inflation exceeds the target in $t + 1$ if interest rates had been raised in period t, that is, it is better to 'be seen to be doing something', then $\delta_{t+1,b} = U_{t+1,by} - U_{t+1,bn} > 0$ as assumed in the Section 6.4. Moreover, $\delta_{t+1,g} = U_{t+1,gn} - U_{t+1,gy} > 0$ as there are output costs to raising interest rates unnecessarily.

6.6 Statistical measures related to economic value

Ignoring the dependence of probabilities on actions, from (6.10) the part of the expected economic value that depends on the probability forecast $\hat{\pi}_{t+1}$ is given by:

$$E[\upsilon_{t+1}(\hat{\pi}_{t+1}, s_{t+1}) \mid \Omega_t] = b_{t+1}(\pi_{t+1} - q_{t+1})1(\hat{\pi}_{t+1} > q_{t+1}). \tag{6.13}$$

Suppose we have a set of probability forecasts and states for $t = 1, \ldots, T$, then the expectation in (6.13) can be evaluated by averaging over the observations to give the average realized economic value (that depends on the forecasts) as:

$$\upsilon = \frac{1}{T}\sum_{t=1}^{T} b_t(s_t - q_t)1(\hat{\pi}_t > q_t). \tag{6.14}$$

Probability forecasts are often evaluated using the quadratic probability score (QPS) of Brier (1950) and the log probability score (LPS). These are

defined as:

$$\text{QPS} = \frac{2}{T}\sum_{t=1}^{T}(\hat{\pi}_t - s_t)^2$$

$$\text{LPS} = -\frac{1}{T}\sum_{t=1}^{T}[s_t \ln \hat{\pi}_t + (1 - s_t)\ln(1 - \hat{\pi}_t)]. \tag{6.15}$$

The QPS is bounded between 0 and 2, with lower numbers denoting more accurate. It is of course just twice the standard MSFE measure. The MSFE (or root MSFE) is a popular measure for comparing prediction errors from point-forecasting exercises that range over the real line, as discussed in Chapter 2. As a squared measure, large mistakes attract disproportionately greater penalties than small errors. The LPS is non-negative, and penalizes large mistakes more heavily than QPS. The LPS is the negative of the average log likelihood for the logit binary choice model.

The Kuipers score (Ks) is defined as:

$$\text{Ks} = H - F,$$

where H is the 'hit rate', the proportion of the total number of Bad states that were correctly forecast, and F is the 'false alarm' rate, defined as the proportion of the total number of Good states that were incorrectly forecast as being Bad states. The advantage of the Ks statistic over measures such as QPS and LPS is that always forecasting the Bad state to occur (or always forecasting the Good state) will score zero, whereas such strategies may fare well on QPS and LPS. Notice that the Ks evaluates forecasts of events (whether the Bad or Good state is forecast to occur) rather than forecasts of the probabilities of events. Given the latter, we can obtain the former using $1(\hat{\pi}_{t+1} > q_{t+1})$ in the spirit of decision-based evaluation, so that the Bad state is forecast to occur when $1(\hat{\pi}_{t+1} > q_{t+1}) = 1$. Then, H and F can be expressed as:[6]

$$H = \frac{\sum_{t=1}^{T} s_t 1(\hat{\pi}_t > q_t)}{\sum_{t=1}^{T} s_t}, \qquad F = \frac{\sum_{t=1}^{T}(1 - s_t)1(\hat{\pi}_t > q_t)}{\sum_{t=1}^{T}(1 - s_t)}.$$

Granger and Pesaran (2000b) show that in special circumstances the economic value criterion is proportional to Ks:

$$\upsilon = b\bar{s}(1 - \bar{s})\text{Ks},$$

where $\bar{s} = T^{-1} \sum_{t=1}^{T} s_t$, the estimate of the (unconditional) probability of the Bad state. To obtain this expression, the decision problem has to be simplified by assuming that $b_t = b$, all t, $q_t = q = \bar{s}$, all t.

However, we can also calculate economic value under a less restrictive set of assumptions concerning the decision problem. In that case, the proportionality between economic value and Ks does not hold, and the simple decision-based framework may yield useful additional information on the relative quality of the MPC forecasts compared to the benchmark forecasts. We again assume that $b_{t+1} = b$ for all t (as above), so that the relative comparisons of economic value do not depend on the b's. We also need to assume that $q_{t+1} = q$, for all t (again as above). Then the relative economic value of the MPC to benchmark forecasts is given by:

$$\frac{\upsilon_{\text{MPC}}}{\upsilon_{\text{bench.}}} = \frac{(1/T) \sum_{t=1}^{T} (s_t - q) 1(\hat{\pi}_t > q)}{(1/T) \sum_{t=1}^{T} (s_t - q) 1(\tilde{\pi}_t > q)}. \tag{6.16}$$

(The $\tilde{\pi}_t$ denote the benchmark forecast probabilities of the bad state). Recall that q depends on δ_g and δ_b: the output costs of raising interest rates in the good state (in utility terms), and the utility gains of raising rates in the bad state. In ignorance of the relative importance to be attached to these two, we consider a grid of values on the unit interval $[0, 1]$ for q. Finally, we stress that the interpretation of these calculations as representing 'economic value' rests on the simple decision-problem setting with action-independent state probabilities. At best this may be a reasonable interpretation for the current and next quarter MPC forecasts.

6.7 The Bank of England MPC inflation forecasts

The MPC inflation forecasts are described in Section 5.7. Here we record some of the measures of forecast performance described in this chapter for a simplified decision-based evaluation of the forecasts. Table 6.3 gives the LPS and QPS scores of the MPC probability forecasts of the Bad event (inflation in excess of $2\frac{1}{2}\%$), along with the scores for unconditional probability forecasts, for the current, next quarter, and one-year ahead horizons. The unconditional forecasts of the bad state, calculated as the sample frequency of $s_t = 1$ ($\bar{s} = T^{-1} \sum_{t=1}^{T} s_t$) can be viewed as a benchmark for the MPC forecasts. The MPC forecast probabilities are calculated by reading off the probabilities that inflation will exceed $2\frac{1}{2}\%$ given the 2PN with the specific values of the mode and two standard deviations.

Table 6.3 Comparisons based on statistical measures, LPS and QPS

	LPS		QPS	
	$\hat{\pi}$	$\bar{s}$	$\hat{\pi}$	$\bar{s}$
Current quarter	0.208	0.669	0.106	0.474
Next quarter	0.399	0.656	0.259	0.463
One year ahead	0.623	0.576	0.441	0.388

Note: $\hat{\pi}$ are the MPC probability forecasts, and $\bar{q}$ are the unconditional forecasts, defined by $\bar{s} = T^{-1}\sum_{t=1}^{T} s_t$, of the bad state (inflation exceeds 2.5%).

Table 6.4 Comparisons based on economic value

q	Current quarter		Next quarter		Year ahead	
	$\hat{\pi}$	$\bar{s}$	$\hat{\pi}$	$\bar{s}$	$\hat{\pi}$	$\bar{s}$
0.1	0.326	0.291	0.286	0.264	0.168	0.163
0.2	0.287	0.191	0.236	0.164	0.095	0.063
0.3	0.261	0.091	0.150	0.064	0.042	0.000
0.4	0.217	0.000	0.100	0.000	−0.063	0.000
0.5	0.174	0.000	0.091	0.000	−0.053	0.000
0.6	0.122	0.000	0.091	0.000	−0.063	0.000
0.7	0.035	0.000	0.010	0.000	0.000	0.000
0.8	0.000	0.000	0.000	0.000	0.000	0.000
0.9	0.000	0.000	0.000	0.000	0.000	0.000
1.0	0.000	0.000	0.000	0.000	0.000	0.000

Note: Columns 2–7 evaluate equation (6.14) with $b_t = b = 1$, $\forall t$ and $q_t = q$, $\forall q$, at different values of q. $\hat{\pi}$ are the MPC event probability forecasts, and $\bar{s}$ are the unconditional forecasts, defined by $\bar{s} = T^{-1}\sum_{t=1}^{T} s_t$. The bad state is that inflation exceeds 2.5%.

At the short-horizons the MPC forecasts are clearly preferred, though $\bar{s}$ has smaller LPS and QPS scores than $\hat{\pi}$ at the one-year ahead horizon.

Table 6.4 records the decision-based evaluation calculations at values of q between 0.1 and 1 (in steps of 0.1). A range of values of q is used because of ignorance of the payoffs: see Section 6.6. When $q = \bar{s} = 0.35$ (for the current quarter forecasts), economic value is proportional to the Ks (see Section 6.6). For both the current and one-quarter ahead horizons the MPC forecasts have higher economic value than $\bar{s}$ for $q < 0.8$. For the year-ahead forecasts the interpretation of the figures in the table as

representing 'economic value' is more strained, but the superiority of the unconditional forecasts for $q > 2$ is clear. Intuitively, the MPC probability forecasts generally overstate the likelihood of the Bad state relative to $\bar{s}$, which is the proportion of Bad states that actually materializes. Thus $1(\hat{\pi}_t > q_t) = 1$ in (6.14) for a number of observations for which $s_t - q_t < 0$ (because $s_t = 0$), for values of q in the range 0.3–0.6. For $q > 0.6$, the part of economic value that varies with the forecast probability is zero because $1(\hat{\pi}_t > q_t) = 0$ for all t.

6.8 Properties of optimal forecasts for general loss functions

Under squared-error loss, the following four properties of optimal forecasts can be easily established:

Property (1). Unbiased forecasts.

Property (2). Forecast-error variance is monotonically non-decreasing in the forecast horizon.

Property (3). h-step forecast errors follow an MA process which is at most of order $h - 1$.

These properties follow more or less immediately for the AR(1) process discussed in Section 2.1.1, and can be obtained with little more effort for a stationary ARMA process via the Wold representation (infinite moving average). For the stationary AR(1) $y_t = \phi y_{t-1} + \nu_t$,[7] property (1) follows from:

$$e_{t+h|t} = y_{t+h} - y_{t+h|t} = \sum_{i=0}^{h-1} \phi^i \nu_{t+h-i}, \tag{6.17}$$

where $y_{t+h|t} \equiv E(y_{t+h} \mid \Omega_t)$ is the MMSEP (the optimal forecast for squared-error loss),

$$y_{t+h|t} = E\left[\left(\phi^h y_t + \sum_{i=0}^{h-1} \phi^i \nu_{t+h-i}\right) \middle| \Omega_t\right] = \phi^h y_t.$$

Then $E(e_{t+h|t}) = 0$.

Property (2) is shown by considering $\mathrm{Var}(e_{t+h+s|t}) = E(e^2_{t+h+s|t})$ compared to $\mathrm{Var}(e_{t+h|t}) = E(e^2_{t+h|t})$, and establishing that $\mathrm{Var}(e_{t+h+s|t}) \geq \mathrm{Var}(e_{t+h|t})$ for

$s > 0$.[8] Thus:

$$E(e^2_{t+h+s|t}) = E\left(\sum_{i=0}^{h+s-1} \phi^i v_{t+h+s-i}\right)^2 = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i} + \sigma^2 \sum_{i=h}^{h+s-1} \phi^{2i}$$
$$\geq E(e^2_{t+h|t}) = \sigma^2 \sum_{i=0}^{h-1} \phi^{2i}$$

Property (3). Consider the h-step forecast errors for $s = 1, 2, \ldots$:

$$e_{t+s+h|t+s} = y_{t+s+h} - y_{t+s+h|t+s} = \sum_{i=0}^{h-1} \phi^i v_{t+s+h-i} \tag{6.18}$$

and the covariances:

$$\text{Cov}(e_{t+s+h|t+s}, e_{t+r+h|t+r}) = E(e_{t+s+h|t+s}, e_{t+r+h|t+r})$$

for $r = 1, 2, \ldots$ and $s \geq r$.

$$E(e_{t+s+h|t+s}, e_{t+r+h|t+r}) = E\left[\left(\sum_{i=0}^{h-1} \phi^i v_{t+s+h-i}\right)\left(\sum_{i=0}^{h-1} \phi^i v_{t+r+h-i}\right)\right]. \tag{6.19}$$

The forecast errors will be correlated whenever some same-dated v's are common to both the summations. This will not be the case when $s - r > h - 1$. For $h = 1$, this condition is satisfied when $s > r$, so that sequences of optimal 1-step forecasts are serially uncorrelated ($s = r$ gives the variance). For $s - r \leq h - 1$, examination of (6.19) indicates that the maximum order of correlation is $h - 1$.

Granger (1999) and Patton and Timmermann (2003) establish that none of these properties need hold for optimal forecasts when the loss function is something other than squared-error loss. Patton and Timmermann (2003) show that under certain assumptions about the nature of the data generating process, property (3) ('lack of serial correlation in excess of $h-1$') can be established for asymmetric loss functions, thus allowing one aspect of forecast rationality to be tested without needing to assume squared-error loss. The restrictions on the process are that higher-order conditional moments (e.g., the conditional variance) are constant. They allow for non-linear regime-switching processes, that is, conditional mean dynamics, but rule out non-constancy in all other conditional moments. As Patton and Timmermann (2003) argue, such

an assumption is clearly unattractive for financial time series, but may be less contentious for macroeconomic time series. The example of the linex loss function with conditional variance dynamics in Section 3.3 is clearly at odds with these conditions. If we replace the data generating process we used there:

$$y_{t+h} \mid Y_t \sim N(y_{t+h|t}, \sigma^2_{t+h|t}),$$

(where $y_{t+h|t}$ is the conditional expectation and $\sigma^2_{t+h|t}$ is the conditional variance) with:

$$y_{t+h} \mid Y_t \sim N(y_{t+h|t}, \sigma^2_h)$$

then we can obtain the optimal predictor assuming a constant conditional variance by simply re-working the steps in Section 3.3. That is, the optimal predictor solves:

$$\underset{\alpha_{t+h}}{\operatorname{argmin}} E_t[b(\exp(a(y_{t+h} - y_{t+h|t} - \alpha_{t+h})) - a(y_{t+h} - y_{t+h|t} - \alpha_{t+h}) - 1)],$$

where $\tilde{y}_{t+h} = y_{t+h|t} - \alpha_{t+h}$. Noting that:

$$E_t[\exp(ay_{t+h})] = \exp\left(ay_{t+h|t} + \frac{a^2\sigma^2_h}{2}\right),$$

gives:

$$\underset{\alpha_{T+h}}{\operatorname{argmin}}\, b\left[\exp\left(\frac{a^2\sigma^2_h}{2} - a\alpha_{t+h}\right) + a\alpha_{t+h} - 1\right]. \tag{6.20}$$

The first-order condition is satisfied by:

$$\alpha_{t+h} = \frac{a}{2}\sigma^2_h$$

so the adjustment to the conditional expectation depends only on h (and the parameter a of the linex loss function). The optimal predictor becomes:

$$\tilde{y}_{t+h|t} = y_{t+h|t} + \frac{a}{2}\sigma^2_h. \tag{6.21}$$

The proof that the covariance between the forecast errors (using the optimal predictor under linex loss) is zero for forecasts made j steps apart, when $j \geq h$, follows immediately. Defining $e_{t+h|t} = y_{t+h} - y_{t+h|t}$, the forecast error using the optimal forecast under squared-error loss, and $\tilde{e}_{t+h|t} = y_{t+h} - \tilde{y}_{t+h|t}$, the forecast error using the optimal predictor under linex loss, then we have by property (3) for squared-error loss that:

$$\text{Cov}(e_{t+h|t}, e_{t+h-j|t-j}) = 0, \quad j \geq h.$$

Because $\tilde{e}_{t+h|t} = e_{t+h|t} - a_h^*$, where $a_h^* = (a/2)\sigma_h^2$, then:

$$\text{Cov}(\tilde{e}_{t+h|t}, \tilde{e}_{t+h-j|t-j}) = 0, \quad j \geq h.$$

6.8.1 General loss functions and the generalized forecast error

Whilst the properties listed at the head of this section will not hold in general for optimal forecast errors under asymmetric loss without restrictions on the higher moments of the data generating process, Granger (1999) establishes that these properties do hold for a 'generalized forecast error', where the generalized forecast error is the derivative of the loss function with respect to the forecast error. Patton and Timmermann (2003) present a slightly more general analysis, where loss depends on Y_{t+h} and $\hat{Y}_{t+h}$ (the forecast), $L(Y_{t+h}, \hat{Y}_{t+h})$, rather than restricting the two arguments to enter as $L(Y_{t+h} - \hat{Y}_{t+h})$. Under some (technical) assumptions on $L(\cdot)$ and the generating process, the generalized forecast error is derived by calculating the optimal forecast ($\hat{Y}^*_{t+h}$) for the general loss function:

$$\hat{Y}^*_{t+h} = \underset{\hat{Y}_{t+h}}{\text{argmin}}\, E_t[L(y, \hat{Y}_{t+h})],$$

$$\hat{Y}^*_{t+h} = \underset{\hat{Y}_{t+h}}{\text{argmin}} \int L(y, \hat{Y}_{t+h}) f_{t+h,t}(y) dy,$$

where $f_{t+h,t}(y)$ is the density of $Y_{t+h} \mid \Omega_t$. Then the first-order condition satisfies:

$$\frac{\partial E_t[L(y, \hat{Y}^*_{t+h})]}{\partial \hat{Y}_{t+h}} = 0$$

or:

$$E_t\left[\frac{\partial L(y,\hat{Y}^*_{t+h})}{\partial \hat{Y}_{t+h}}\right]=\int \frac{\partial L(y,\hat{Y}^*_{t+h})}{\partial \hat{Y}_{t+h}} f_{t+h,t}(y)dy=0. \tag{6.22}$$

The optimal generalized forecast error is defined by:

$$\psi^*_{t+h|t}=\frac{\partial L(y,\hat{Y}^*_{t+h})}{\partial \hat{Y}_{t+h}} \tag{6.23}$$

so that from (6.22) with (6.23):

$$E_t(\psi^*_{t+h|t})=\int \psi^*_{t+h|t} f_{t+h,t}(y)dy=0. \tag{6.24}$$

The generalized forecast error is conditionally (and therefore also unconditionally) unbiased from (6.24) and the other properties can also be established. Note that for squared-error loss, $L(Y_{t+h}-\hat{Y}_{t+h})=(Y_{t+h}-\hat{Y}_{t+h})^2$, the generalized forecast error is:

$$\psi^*_{t+h|t}=-2(Y_{t+h}-\hat{Y}^*_{t+h})$$

that is, twice the standard forecast error.

Artis and Marcellino (2001) use the optimality properties of generalized forecast errors to evaluate fiscal forecasts (assuming particular non-linear loss functions).

6.9 Summary

In this chapter, we assume that forecasts are made in order to guide actions or decisions in an uncertain environment. So actions depend on forecasts, and the payoff or return to a particular action depends on the state of nature that eventuates. For a particular forecast we can calculate the expected economic value (or expected loss), where the expectation is calculated by weighting the returns to that action in the different states of nature by the actual probabilities of those states. This allows a ranking of forecasts. However, users with different loss functions may rank two rival sets of forecasts differently – each user's expected economic value need not be maximized by the same set of forecasts. An important result is that a density forecast that coincides with the true density

will maximize economic value of all users regardless of their loss functions. Furthermore, for general loss functions the whole forecast density is required to calculate economic value – only in exceptional circumstances will the minimum MSFE point forecast be sufficient to generate optimal decisions and maximize economic value.

From the description of the simple two-state, two-action decision problem it is apparent that the decision-based approach to forecast evaluation may often not be feasible: this is illustrated with regard to the Bank of England's inflation targeting and interest-rate setting problem. We consider the relationship between economic value and some simple statistical measures of forecast performance, and calculate these for the Bank's inflation forecast densities.

For general loss functions, a number of standard properties that hold for optimal forecasts in the case of squared-error loss may be violated (e.g., optimal forecasts will be biased for asymmetric loss functions). A generalized forecast error can be defined for which these properties do hold, but for which knowledge of the loss function is required.

7
Postscript

In recent years, forecasts that give a more complete description of the likely future values of economic and financial variables than the expected mean have become increasingly prominent, in academia as well as in government and financial regulation. The focus of this book has been on the evaluation of these forecasts. A number of the issues relevant to the evaluation of point forecasts are equally germane to the evaluation of interval and density forecasts. There are also new problems to be overcome, such as the fact that volatility is unobserved, when forecasts of conditional variance are evaluated.

A key distinction for all the types of forecast is whether the evaluation makes reference to the method of construction of the forecast. We considered evaluation methods for point, interval and density forecasts (in Sections 2.1, 4.4 and 5.2, respectively) that involves only the sequence of forecasts and outcomes.[1] A number of recent papers on density evaluation have considered survey-based forecasts, such as the Survey of Professional Forecasters histograms reviewed in Section 5.6, where the method of construction is typically unknown to the econometrician, so this method of evaluation is warranted. The recent point forecast literature considers the evaluation of model-based forecasts when account is taken of estimation error (Section 2.4.2), either when testing a single set of forecasts for unbiasedness, or when comparing rival sets of forecasts. The approaches reviewed in Section 5.8 evaluate density forecasts which are model-based.

Forecast evaluation that makes reference to the method of construction, that is, to the model underlying the forecasts (as in Section 2.4.1) can become an exercise in the evaluation of the model, rather than of the forecasts from that model. This would be the case, for example, if models were selected and adopted for later use on the basis of their

forecast performance. That a good out-of-sample forecast performance lends credence to an econometric model, and to any economic theory on which it might be based, has a commonsense appeal, and is often expressed, for example:

> any inflation forecasting model based on some hypothesized relationship cannot be considered a useful guide for policy if its forecasts are no more accurate than such a simple atheoretical forecast (namely, next year's inflation will equal last year's). (Atkeson and Ohanian, 2001)

and:

> If a dynamic modeling approach is to be convincing, it needs to say something about the behavior of unemployment out of sample. (Carruth *et al*. 1998, p. 626)

However, Clements and Hendry (2003) suggest that out-of-sample forecast performance may not be a reliable indicator of the validity of an empirical model, nor therefore of the economic theory on which the model is based. One reason is the prevalence of structural breaks and instabilities in macroeconomic time-series relationships (see, e.g., Stock and Watson 1996), such that a model may be a useful partial description of the economic relations of interest even though it forecasts poorly: Clements and Hendry (1999) examine the effects of structural shifts on forecast performance. In addition, Clements and Hendry (2003) draw attention to a number of aspects concerning the conduct of the forecast evaluation exercise which may appear to be incidental but can turn out to be decisive.

In focusing on evaluation, we have had little to say about these wider issues relating to forecasting economic and financial variables, including reasons why forecasts often turn out to be poor (see, e.g., Spanos 1989 and Wallis 1989 for some specific empirical instances), or the relative merits of different models or methods. The wider picture is provided by Clements and Hendry (1998, 1999) and the collection of papers in Clements and Hendry (2002), *inter alia*.

The purpose of forecasting is not viewed as primarily an exercise in evaluating a specific model, or choosing one model from a set of models. But rather to inform decision-making, where actions are taken today which have consequences in the uncertain future, as discussed in Chapter 6. Ideally all forecasts would be made for a specific purpose

and would be evaluated accordingly. But given informational costs and deficiencies, forecasts are often 'general purpose', and may give a far from complete description of the probabilities of the variable falling in certain ranges, necessitating the evaluation techniques in Chapters 2–5.

The last ten years have witnessed significant progress in the development of methods for evaluating forecasts that go beyond the most likely or the expected outcome. New and improved methods of forecast evaluation can be expected in the near future as the trend towards more informative forecasts continues.

8
Computer Code

Some sample code is given and described. The data sets and programs can be downloaded from the Palgrave Macmillan web page `www.palgrave.com/economics/Clements/index.asp`. The Gauss code given below is for illustrative purposes. It is not meant to illustrate good programming technique, or to be especially general. It is hoped that the sample code might encourage the reader to experiment, and some suggestions are given which may be taken up as simple exercises.

8.1 Sample Gauss code for the estimation and forecasting of SETAR models

See Section 2.5.3. The Gauss code consists of two procedures in the file `SETAR.PRG`. `SETARe` estimates a two-regime SETAR model for a given lag order p. The two arguments are the lag order p and the variable y. The second procedure, `SETARf`, generates multi-step forecasts by Monte Carlo from the model estimated by `SETARe`. The two procedures are followed by the main body of the program. Here the data series is loaded from a plain text file, `gnp92.dat`, which contains observations on the first difference of the log of quarterly US GNP, 1959:2 to 1996:2. The data is multiplied by one hundred in the program, to give approximate percentage growth rates. The SETAR estimation procedure is run on the first 127 observations (1959:2 to 1990:4), the results are printed, and the estimated model is then used to forecast 1991:1 to 1996:2. The RMSFE (root mean squared forecast error) is printed, as is the RMSFE of a forecast calculated as the mean of the estimation sample period.

`SETARe` works as follows. Construct the data matrix as $X = (Y_T : \iota : Y_{T-1} : \cdots : Y_{T-p})$, where $Y_T = (y_{p+1}, \ldots, y_T)$, $\iota = (1, \ldots, 1)$, $Y_{T-1} = (y_p, \ldots, y_{T-1})$ and $Y_{T-p} = (y_1, \ldots, y_{T-p})$. The data are then sorted

on the $d + 2$th column, for $d = 1, 2, \ldots$ in turn. For a given d, a grid search is carried out splitting the sample into two for all admissible values of r. That is, separate regressions are run on all the rows of $\mathbf{X}$ for which the corresponding elements in Y_{t-d} are less than r, and on all the rows for which the elements of Y_{t-d} exceed r. The pair of $\{r, d\}$ which minimizes the residual sum of squares (RSS) is then identified. (The program stores the values of the model estimates for the running smallest RSS. For a given $\{r, d\}$, the model estimates are stored only if the associated RSS is smaller than the previous smallest.

The forecast origin is taken by `SETARf` as the last observation of the y vector (passed when the procedure is called). Thus passing the same vector of observations to `SETARe` and `SETARf` results in 'out-of-sample' forecasts based on the estimated model. For $p = 2$, for example, `SETARf` constructs $\mathbf{X}$ as the row vector $(y_T : y_{T-1})$, where y_T is the last element of the vector passed to the procedure. Then the $j = 1$ forecast is calculated using either the lower or upper regime values depending on whether $\mathbf{X}[1, d]$ (which is y_{T+1-d}) is less than or greater than r. The forecast value $y_{T+1|T}$ is then augmented with $\varepsilon_{T+1} = [\sigma_1 1(y_{T+1-d} \leq r) + \sigma_2 1(y_{T+1-d} > r)]\ v_{T+1}, (v_{T+1} \sim N(0, 1))$, before replacing $\mathbf{X}$ by $(y_{T+1|T} + \varepsilon_{T+1} : y_T : y_{T-1})$ and then calculating the 2-step ahead forecast when $j = 2$, etc.

8.1.1 Extensions

The procedure `SETARe` conducts a grid search over the threshold value r and the delay d for a user-input lag order p. A loop could be added to search over $p = 1, 2, \ldots$ up to some pre-set maximum.

A third regime could be included, by partitioning the data into three and performing a grid search over r_1 and r_2, where $-\infty < r_1 < r_2 < \infty$.

Sequences of 1 to h-step ahead forecasts could be produced by moving the forecast origin through the sample. That is, first estimate the SETAR model on observations 1 to 127, and then forecast observations 128 on (as at present), then estimate the model on observations 1 to 128, and forecast observations 129 on, etc. As an alternative to an expanding data estimation window, a fixed window could be used, for example, 1 to 127, then 2 to 128 etc. Either could be easily accomplished by putting the calls to the estimation and forecasting procedures in a loop.

Tests for significant differences in RMSFE between the SETAR forecasts and rival forecasts could be programmed.

```
PROC(1) = SETARe(p, y);
Local d, pp, b1, b2, X1, X2, Y1, Y2, s1, s2,
X, Z, i, lo, hi, e1, e2, T, n1, n2,
Results, JRSS, JRSS0;


JRSS0 = 10000000;
Results = zeros(1,2*(p+1)+6);


X = y[p+1:rows(y),1]~ones(rows(y)-p,1);
pp = 1;                @X contains dependent variable, intercept, @
do while pp <= p;      @and the number of lags specified by p     @
X = X~y[p+1-pp:rows(y)-pp,1];
pp = pp+1;
endo;

T = rows(X); lo = round(.15*T); hi = round(.85*T);

d=1; do while d <= p;   @grid search over d @

Z = SORTC(X,2+d);       @sort data matrix @

@loop over each value of r from lo to hi@
i = lo; do while i <= hi;

X1 = Z[1:i,2:p+2];    Y1 = Z[1:i,1];
X2 = Z[i+1:T,2:p+2];  Y2 = Z[i+1:T,1];

b1 =invpd(X1'X1)*X1'Y1;  b2 =invpd(X2'X2)*X2'Y2;

e1 = Y1 - X1 * b1;
e2 = Y2 - X2 * b2;

JRSS = e1'e1 + e2'e2;

n1 = i; n2 = T - i;

s1 = SQRT((e1'e1)/(n1-(p+1)));
s2 = SQRT((e2'e2)/(n2-(p+1)));

if JRSS < JRSS0; JRSS0 = JRSS;
Results[1,1:p+1] = b1';  Results[1,p+2:2*p+2] = b2';
Results[1,2*p+3:2*p+8] = s1~s2~n1~n2~Z[i,2+d]~d;
endif;

i = i+1; endo;   @end loop over r@

d = d+1; endo;   @end loop over d@

RETP(Results);
ENDP;
```

```
PROC(1) = SETARf(y,a0,a1,b0,b1,r,d,p,s1,s2,f_reps,h);
Local X,pp,e,j,f,f_rep;

e = rndn(h,f_reps);
f = zeros(h,f_reps);

X = y[rows(y),1];
pp=1; do while pp <= p-1;
X = X~y[rows(y)-pp,1];
pp=pp+1; endo;

f_rep = 1; do while f_rep <= f_reps;

j = 1; do while j <= h;
f[j,f_rep]=
  (a0 + X[1,1:p]*a1') * (X[1,d] .<= r)
+ (b0 + X[1,1:p]*b1') * (X[1,d] .> r);

X = ((s1*e[j,f_rep])*(X[1,d] .<= r)
        + (s2*e[j,f_rep])*(X[1,d] .> r) + f[j,f_rep] )~X;
j=j+1; endo;

f_rep = f_rep + 1; endo;
f = meanc(f');

RETP(f);
ENDP;


load y[149,1] = gnp92.dat;   @Diff log US GNP. 59:2 to 96:2@
y = y*100;                   @Approx. growth rates@
p = 2;                       @Two lags@

res = SETARe(p,y[1:127]);    @Estimate 59:2 to 90:4@

@Estimation results@
r = res[1,2*p+7];            d = res[1,2*p+8];
a0 = res[1,1];              a1 = res[1,2:p+1];
b0 = res[1,p+2:p+2];        b1 = res[1,p+2+1:2*p+2];
s1 = res[1,2*p+3:2*p+3];    s2 = res[1,2*p+4:2*p+4];
"Threshold";; r;
"Regime 1 coefficients";
a0~a1;
"Regime 2 coefficients";
b0~b1;
"s1, s2, n1, n2, d";
s1~s2~res[1,2*p+5:2*p+6]~d;
```

```
h = 22; f_reps = 100;
f = SETARf(y[1:127],a0,a1,b0,b1,r,d,p,s1,s2,f_reps,h);
@Forecast results@

"SETAR model RMSFE";; SQRT(meanc((y[128:149,1]-f)^2));

"Mean RMSFE       ";; SQRT(meanc((y[128:149,1]

                        -meanc(y[1:127,1]))^2));

end;

@Results@

Threshold     0.32579019
Regime 1 coefficients
       0.21674205       0.13649007         -0.22823162
Regime 2 coefficients
       0.56852695       0.33262725         0.012922933
s1, s2, n1, n2, d
       1.2172782        0.64313269           31     94     2
SETAR model RMSFE      0.38347886
Mean RMSFE             0.46811554
```

Figure 8.1 Gauss code. SETAR model estimation and forecasting

8.2 Estimation and forecasting GARCH(1,1) processes

The monthly observations on three-month US Treasury Bill interest rates and ten-year Treasury bond interest rates (taken from the Federal Reserve of St Louis database, `www.stls.frb.org/fred`) shown in Chapter 3 are contained in `intrates.xls`, for the period 1953:04 to 2001:10. The AR(4)–GARCH(1,1) models, and the forecasts from those models depicted in Figures 3.4 and 3.5, can easily be obtained using GiveWin and PcGive (Doornik and Hendry, 2001). They were produced using the menu-driven 'Volatility models' package of PcGive 10.0.

The main text illustrates the dependence of forecasts of conditional variance on the volatility of the process at the forecast origin for standard GARCH(1,1) models. Possible extensions include estimating alternative GARCH models, such as non-linear threshold models, as well as experimenting with alternative parameter restrictions given that $\alpha + \beta > 1$ freely estimated. What are the consequences of failing to impose $\alpha + \beta = 1$?

These data series have been analysed by Clements and Galvão (2003), *inter alia*, in the context of testing the expectations theory of the term structure.

8.3 Monte Carlo evaluation of interval lengths and coverages

In this section, we present Gauss code to perform a Monte Carlo evaluation of the coverage levels of two methods of calculating interval forecasts for an autoregressive model. See Section 4.2. The two methods are the Box–Jenkins (BJ) method, and a method that allows for parameter estimation uncertainty. The Gauss program is `interval.prg`: see Figure 8.2. It makes use of procedures in the Gauss Time Series library `ARIMA` to estimate and forecast ARIMA models, but as only an AR model is considered, it is a relatively simple task to produce code to carry out these calculations.

We first describe the program, and then suggest extensions.

In addition to the `ARIMA` library procedures `arima` and `forecast`, which estimate and forecast an ARIMA(p, d, q) for given values of p, d and q, a procedure called `coverage` appears at the top of the program. For two h-dimensional vectors defining the upper and lower limits of the intervals for 1 to h-step ahead forecasts, `coverage` calculates coverage levels and interval lengths for an $h \times k$ matrix of continuations, where the k column vectors are the 'continuations'.

The data generation process (DGP) is an AR(2) with coefficients given by *coef*. The model is correctly specified as an AR(2), and includes an intercept (not present in the DGP). The Monte Carlo replications are over $j = 1, \ldots, R$. Within this loop, the following calculations are performed.

The Gauss procedure `recserar` is used to generate an n-dimensional vector $\mathbf{y}$ from an AR(2), where $y_2 = y_1 = 0$, and the disturbances are $N(0, 1)$. The following line uses the same command to generate k vectors of dimension h as continuations (each is constructed from the same y_n and y_{n-1}: the last two elements of $\mathbf{y}$). An AR(2) is then estimated using `arima`, and the upper and lower intervals for the BJ method are generated automatically by `forecast` for a coverage level given by the global variable *_amcritl* (set to 0.80).

The second method of calculating intervals to allow for parameter estimation uncertainty is a little more complex. `recserar` is used to generate b time series of length n, $(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_b^*)$ using the estimated parameter vector for the data $\mathbf{y}$. On each of these, an AR(2)

```
new;
library arima;
format /rd 7,4;

_amfprnt = 0; _am_opts={0,0};
_amcritl = 0.80; covlev = _amcritl;

proc coverage (w,ly,uy);
  @w  : h by #continuations     @
 @ly : lower interval, h by 1. @
 @uy : upper interval, h by 1. @
local  length,below,above,covg,n;
  n =  cols(w);
  length = uy - ly;
  below  =  sumc(w' .lt ly');
  above  =  sumc(w' .gt uy');
   covg   =  cols(w) - (below + above);
retp (length~below/n~above/n~covg/n);
endp;

n = 25;                  @ # of observations           @
h = 10;                  @ # of step-ahead forecasts   @
R = 10000;               @ # of data replications      @
k = 1000;                @ # continuations to calculate coverages@
b = 100;                 @ # bootstrap replications@


pm = 2; dm = 0; qm = 0;          @AR Model@
coef={0.9,-.7};                  @AR DGP@

resBJ =  zeros(h,4);
resBS =  zeros(h,4);

j = 1;  do while j <= R;

y    =  recserar(rndn(n,1),  zeros  (2,1), coef);
ysim =  recserar(rndn(h+2,k),y[n-1:n,1].*ones(1,k),coef.*ones(1,k));
ysim = ysim[3:h+2,.];

{coeffs,ll,resids,vcb,aic,sbc} = arima(0,y,pm,dm,qm,1);

f = forecast(coeffs,y,pm,dm,qm,1,resids,h);
ly = f [.,1]; uy = f[.,3];                @Box-Jenkins intervals@
resBJ = resBJ + coverage(ysim,ly,uy);

ystar =  recserar(rndn(n,b)+coeffs[3,1] *ones(n,b),
         zeros(2,1). *ones(1,b),coeffs[1:2,1]. *ones(1,b));
fstar =  zeros(b,h);
i = 1;  do while i <= b;
{coeffs,ll,resids,vcb,aic,sbc} = arima(0,ystar[.,i],pm,dm,qm,1);
f =  recserar(rndn(h+2,1)+coeffs [3,1] *ones(h+2,1),
             y[n-1:n,1],coeffs[1:2,1]);
fstar [i,.] = f[3:h+2]';
i = i+1;  endo;
```

```
ly = zeros(h,1); uy = zeros(h,1);
s = 1; do while s <= h;
fsort = sortc(fstar [.,s],1);
ly[s,1]=fsort[round(b*(0.5*(1-covlev)))];
uy[s,1]=fsort[round(b*(0.5*(1+covlev)))];
s = s + 1; endo;

resBS = resBS + coverage(ysim,ly,uy);

j=j+1; endo;

"************************************************"
        "AR coefficients=" coef' " n=" n " c= " covlev;"";

resBJ/r~resBS/r;

end;

*****************************************
AR coefficients= 0.9000 -0.7000 n=25.0000 c= 0.8000


 2.4326 0.1308 0.1330 0.7362 2.7558 0.1023 0.1153 0.7824
 3.1552 0.1461 0.1483 0.7057 3.5697 0.1193 0.1328 0.7479
 3.2025 0.1407 0.1417 0.7177 3.6345 0.1173 0.1297 0.7530
 3.4166 0.1384 0.1388 0.7228 3.7772 0.1174 0.1282 0.7544
 3.5832 0.1410 0.1417 0.7173 3.8432 0.1202 0.1306 0.7492
 3.6202 0.1355 0.1372 0.7273 3.8641 0.1160 0.1274 0.7567
 3.6803 0.1369 0.1388 0.7243 3.9178 0.1185 0.1310 0.7506
 3.7341 0.1410 0.1424 0.7167 3.9629 0.1228 0.1352 0.7420
 3.7558 0.1393 0.1396 0.7210 3.9813 0.1227 0.1347 0.7426
 3.7782 0.1378 0.1375 0.7247 3.9889 0.1221 0.1335 0.7444
```

Figure 8.2 Gauss code. Monte Carlo evaluation of interval forecasts

is estimated, and continuations are generated, conditioned on y_n and y_{n-1}. The sample of b continuations are then sorted and the appropriate percentiles of the empirical distributions are read off.

At the bottom of Figure 8.2 the results of running the program are given. The first four columns are the estimates of interval length, the proportion of actuals below, above, and within, the interval, for the BJ method. The last four columns record the same information for the other method. The rows relate to the forecast horizon (1 to 10).

8.3.1 Extensions

Obvious extensions are to assess the dependence of the results on the sample size n (as given, $n = 25$), and on the assumption that disturbances

are normal. The code also assumes that the process starts at $y_2 = y_1 = 0$ on each of the Monte Carlo replications. Instead of this 'fixed startup' assumption, the first two values of the process could be drawn from the joint distribution of a Gaussian AR(2).

As coded, the $\mathbf{y}_i^*$ vectors are generated using random draws from $N(0, 1)$ for the errors. In the event that the variance of the AR(2) disturbances is neither unity nor normal, the errors should be bootstrapped as described in the main text. This can be achieved by replacing `rndn`(n, b) by an $n \times b$ matrix of randomly drawn (with replacement) errors from *resids*.

Further, code could be added to bias-correct the estimated parameters, one could allow the model to be mis-specified for the DGP, and look at the consequences of selecting the order of the AR (or ARIMA). Finally, ARCH errors could be investigated, as discussed in Section 4.2.6.

8.4 Forecast density evaluation

This section provides a sketch of the application of the density evaluation techniques discussed in Section 5.2. The probability integral transform was defined as:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(y_t), \tag{8.1}$$

where $p_{Y,t-1}(y)$, $t = 1, \ldots, n$, is a sequence of 1-step forecast densities of period t conditional on $t-1$, and $\{y_t\}_{t=1}^n$ are the corresponding outcomes. In Section 8.1 the procedure `SETARf` was used to estimate by Monte Carlo the point forecast as the average across a number of simulated continuations. The first row of the matrix f (calculated in `SETARf`) contains the 1-step continuations (which are averaged to obtain an estimate of the 1-step ahead point forecast) can be used to estimate the empirical distribution function (EDF) as:

$$\hat{P}_{t-1}(u) = \frac{1}{R}\sum_{i=1}^{R} 1(y_{t|t-1}^i < u),$$

where $y_{t|t-1}^i$, $i = 1, \ldots, R$ are the elements of the first row of f. The probability integral transform can be calculated as $z_t = \hat{P}_{t-1}(y_t)$, where y_t is the actual value. Assuming f is returned from the procedure `SETARf`, the

following line of GAUSS code would calculate z_t (where y is the actual):

$$z = meanc(f[1, .]' . < y)$$

If the forecast origin were moved through the data, giving rise to a sequence of 1-step EDFs, then the resulting sample $\{z_t\}_{t=1}^n$ could be tested to see if it constitutes a random sample from a $U(0, 1)$. Suppose $\{z_t\}_{t=1}^n$ is sorted by size, small to large. The theoretical CDF for a $U(0, 1)$ evaluated at the points $\{z_t\}_{t=1}^n$ is just $\{z_t\}_{t=1}^n$, and the empirical CDF has probability $1/n$ attached to z_1, $2/n$ attached to z_2, etc. The following line of Gauss code calculates the maximum absolute difference between the two distributions (and forms the KS statistic):

$$diff = \max c(abs(z - \text{seqa}(1/n, 1/n, n))),$$

where z is the vector of sorted z_t's.

Notes

2 Point Forecasts

1. As an alternative to 'powering-up' the OLS or 1-step estimator $\hat{\phi}$, the relationship between y_t and y_{t-h} can be estimated directly: Clements and Hendry (1996) and Bhansali (2002) discuss 'multi-step' estimation.
2. Approximations to the variances of powers of estimated parameters are given in Schmidt (1977), Baillie (1979a,b) and Chong and Hendry (1986) for general models, and are reviewed in Ericsson and Marquez (1989, 1998) and Campos (1992).
3. The combination of forecasts can be extended to an arbitrary number of forecasts, and weights can be allowed to vary over time: see, for example, Diebold and Pauly (1987), Deutsch *et al*. (1994) and Donaldson and Kamstra (1996).
4. Monte Carlo is used extensively in econometrics. One use of Monte Carlo is to obtain the small-sample distributions of test statistics, as here, where we calculate the actual sizes of test statistics corresponding to the 5% nominal size critical value of the asymptotic distribution (or, as in the case of the 'Standard' test with student t errors, of an *assumed* distribution). General references to Monte Carlo methods include Hammersley and Handscomb (1964), Hendry (1984), Ripley (1987), Davidson and MacKinnon (1993) and Clements and Hendry (1998, Ch. 5).
5. See, for example, Diebold *et al*. (1993, 1994), Filardo (1994), Lahiri and Wang (1994) and Durland and McCurdy (1994).

3 Volatility Forecasts

1. The terms 'time-varying conditional variance' and volatility will be used interchangeably.
2. There is a vast literature on ARCH and related models: Engle and Bollerslev (1987), Bollerslev *et al*. (1992), Bera and Higgins (1993) and Shephard (1996) provide good reviews; and Engle (1995) is an edited selection of some of the key papers. Franses and van Dijk (2000) provide a good recent treatment of linear and non-linear ARCH-type models. Our focus is on forecasting: see in particular Bera and Higgins (1993), Baillie and Bollerslev (1992) and Poon and Granger (2003) for a recent review of volatility forecasting.
3. A cross-plot of the quantiles of an empirical distribution of the data against the quantiles of an hypothesized distribution (here a normal distribution). Departures from the 45° line indicate a poor match between the two distributions.
4. As for modelling the conditional mean, in any specific instance there may be useful extraneous variables that could be brought to bear, but the history of the process is always available. Multivariate models are available but will not be considered here.

5. The Taylor-series expansion of $C(x)$ about $x = 0$ is given by:

$$C(x) = C(0) + (x-0)C'(x)\,|_{x=0} + \frac{(x-0)^2}{2}C''(x)\,|_{x=0} + \cdots$$
$$= \frac{ba^2}{2}x^2$$

after substituting $C'(x) = ab\exp(ax) - ab$, and $C''(x) = a^2b\exp(ax)$.
6. For a textbook treatment of unit roots and (non-)stationarity see, for example, Banerjee *et al*. (1993).
7. Maddala and Li (1996) show that the bootstrap tests of Lamoureux and Lastrapes (1990) are not valid, and outline appropriate bootstrap tests.
8. Anderson and Bollerslev (1998, p. 891) list a large number of studies reporting low R^2's for daily and intra-daily returns, many of which are less than 0.05.
9. $\kappa = 3$ when $\{z_t\}$ is normal. For fatter-tailed distributions $\kappa > 3$, making for an even noisier proxy.
10. Assessments of volatility forecasts based on the quality of derived interval and density forecasts are not subject to the problems that arise from using noisy proxies for actual volatility.

4 Interval Forecasts

1. See Lopez (1996) for a discussion of the relationship between VaR analysis and interval forecasting.
2. The empirical distribution function assigns to each value $\{\hat{\varepsilon}_t\}, t = p+1, \ldots, T$, a measure equal to $1/(T-p)$. So independent draws from $F_{\hat{\varepsilon}}$ correspond to sampling the errors with replacement.
3. The realized values of Y^*_{T+k} are $\{y^*_{b,T+k}, b = 1, \ldots, B\}$.
4. See, for example, Thombs and Schucany (1990), Pascual *et al*. (2001) and Clements and Taylor (2001) for results for a range of distributional assumptions for the disturbance term $\{\varepsilon_t\}$ and other AR models.
5. In the Monte Carlo described in Section 4.2.5 we simulate R future values at Step 1 to perform this calculation.
6. I_t is a binary variable, taking the values of 1 and 0 with probabilities p and $(1-p)$, under correct specification.
7. The likelihood is the product of the n individual densities given independence. Assuming $f(I_j = 1) = \pi$ and $f(I_j = 0) = 1 - \pi$, we obtain:

$$L(\pi; I_1, I_2, \ldots, I_n) = f(I_1) \times \cdots \times f(I_n) = (1-\pi)^{n_0}\pi^{n_1}.$$

8. The likelihood is derived from decomposing the joint density into the product of the 1-step conditionals $f(I_t \mid I_{t-1}), t = 2, \ldots, n$, and $f(I_1)$, and then ignoring the first observation I_1:

$$L(\Pi_1) = f(I_2 \mid I_1)f(I_3 \mid I_2) \times f(I_n \mid I_{n-1}).$$

Then note that $f(I_j = 1 \mid I_{j-1} = 1) = \pi_{11}$, and this occurs n_{11} times, $f(I_j = 0 \mid I_{j-1} = 0) = \pi_{00} = 1 - \pi_{01}$, and this occurs n_{00} times, $f(I_j = 1 \mid I_{j-1} = 0) = \pi_{01}$, and this occurs n_{01} times, and $f(I_j = 0 \mid I_{j-1} = 1) = \pi_{10} = 1 - \pi_{11}$, and this occurs n_{10} times.

9. 'Independence' is a misnomer because only second-order properties are being considered.
10. See Taylor and Bunn (1998) for extensions.
11. The standard errors are calculated using the Bollerslev and Wooldridge (1992) method.
12. For details of this runs test in an interval forecast evaluation setting see Christoffersen and Diebold (2000).

5 Density Forecasts

1. See, for example, the recent survey article Tay and Wallis (2000).
2. Diebold *et al*. (1999) discuss errors in the specification of the sequence of $p_{Y,t-1}(\cdot)$'s that would lead to either the i.i.d. or uniformity aspects failing to hold.
3. Detailed information on the survey as well as the survey results are available at the URL `www.phil.frb.org/econ/spf`. An academic bibliography of articles that either discuss or use data generated by the SPF is also maintained online.
4. Giordani and Söderlined (2003) consider a number of different ways of obtaining measures of inflation uncertainty from the SPF, and in particular the practice of aggregating individual respondents' histograms.
5. The series were taken from the Federal Reserve Bank of St Louis database (FRED), available at the URL `www.stls.frb.org/fred/data/` and have the codes GNPDEF, GDPDEF and GDPCTPI.
6. As an example, suppose we wish to calculate the forecast probability of observing a value less than $Y = 3.5$. Suppose Pr $(Y < 2)$ is 0.5, and the bin defined by $[2, 4)$ has a probability of 0.2. Then:

$$\begin{aligned}\Pr(Y < 3.5) &= \Pr(Y < 2) + \frac{1.5}{2}\Pr(Y \in [2, 4)) \\ &= 0.5 + \frac{1.5}{2}0.2 = 0.65.\end{aligned}$$

   Linear interpolation follows the assumption implicit in the histogram – that probability mass is uniform within a bin. If a bin is bordered by a high probability bin and a relatively low probability bin, it might be desirable to attach higher probabilities to points near the boundary with the high probability bin. See Giordani and Söderlined (2003) for a discussion of this point and an alternative approach.
7. Available for each quarter from 1997: 3 to the present from the Bank of England web site (homepage `www.bankofengland.co.uk`). The forecasts reported here were downloaded 6 August 2003.
8. See Britton *et al*. (1998) and Wallis (1999) for a discussion.

9. Note that for the US, Atkeson and Ohanian (2001) find that no-change forecasts are more accurate on MSFE than Phillips Curve forecasts, but see also Stock and Watson (1999b). In a study of forecasting inflation in the G7 countries Canova (2002) finds that bivariate and trivariate models are little better than univariate models.
10. For i.i.d. observations Andrews (1997) takes the supremum over points in the sample, $\{y_t, y_{t-1}; t = 1, \ldots, T\}$. The exposition here follows Corradi and Swanson (2003).

6 Decision-based Evaluation

1. See also Granger and Pesaran (2000a).
2. Whether the roads are gritted overnight will not affect the probability of temperatures falling below freezing, for example. Of course, phenomena of the sort typified by Douglas Adams' Rain God, Rob McKenna (in '*So Long, and Thanks for All the Fish*', 1984, London: Pan Books Ltd.) are possible.
3. Because s_{t+1} is binary, $\hat{\pi}_{t+1}$ and $(1 - \hat{\pi}_{t+1})$ constitute the complete forecast density (the probabilities that $s_{t+1} = 1$ and 0, respectively). Moreover, $\hat{\pi}_{t+1}$ is also the forecast of the conditional expectation of s_{t+1} – the point forecast and forecast density coincide. When s_{t+1} is discrete but takes on a range of values, or is a continuous random variable, a full elucidation of the probabilities associated with each value is required, which will go beyond the point forecast.
4. $q_{t+1} > 0$ because $\delta_{t+1,b} > 0$ and $\delta_{t+1,g} > 0$ by assumption. $U_{.,by} > U_{.,bn}$ because 'action' will alleviate the costs incurred in the bad state (having gritted the roads when there is a frost will reduce the number of road traffic accidents). $U_{.,gn} > U_{.,gy}$ because gritting is costly and unnecessary when there is not a frost. q_{t+1} will approach unity as $\delta_{t+1,g}$ gets large relative to $\delta_{t+1,b}$, that is, the costs of prevention ('gritting') are prohibitively expensive relative to the savings in terms of a reduced number of accidents.
5. Consistent with this view, the forecast probabilities attached to the event that inflation falls outside the target range of $1\frac{1}{2}$ to $3\frac{1}{2}$ are generally small.
6. From which it is evident that if the bad state is forecast to occur in each period, for example, $\hat{\pi}_t = 1$, all t, ensuring $1(\hat{\pi}_{t+1} > q_{t+1}) = 1$, all t, then Ks = 0 because $H = F = 1$. However, QPS will be small to the extent that the bad state happens to occur each period.
7. Where $\{v_t\}$ is an i.i.d. zero-mean series $(0, \sigma^2)$ with $E[v_t \mid y_{t-1}, y_{t-2}, \ldots] = 0$, and $|\phi| < 1$.
8. Complications arise from parameter estimation error, but these are another issue and are ignored here to sharpen the arguments.

7 Postscript

1. For the volatility forecasts, Section 3.6 noted a possible proxy for 'actual' volatility.

References

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003). Modelling and forecasting realized volatility. *Econometrica*, **71**, 579–625.

Anderson, G. A. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accuate forecasts. *International Economic Review*, **39**, 885–905.

Anderson, H. M. (1997). Transaction costs and non-linear adjustment towards equilibrium in the US treasury bill market. *Oxford Bulletin of Economics and Statistics*, **59**, 465–484.

Andrews, D. W. K. (1997). A Conditional Kolmogorov test. *Econometrica*, **65**, 1097–1128.

Artis, M. and Marcellino, M. (2001). Fiscal forecasting: The track record of the IMF, OECD and EC. *Econometrics Journal*, **4**, S20–S36.

Atkeson, A. and Ohanian, L. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis, Quarterly Review*, **25**(1), 2–11.

Baillie, R. and Bollerslev, T. (1989). The message in daily exchange rates: A conditional variance tale. *Journal of Business and Economic Statistics*, **7**, 297–305.

Baillie, R. T. (1979a). The asymptotic mean squared error of multistep prediction from the regression model with autoregressive errors. *Journal of the American Statistical Association*, **74**, 175–184.

Baillie, R. T. (1979b). Asymptotic prediction mean squared error for vector autoregressive models. *Biometrika*, **66**, 675–678.

Baillie, R. T. and Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics*, **52**, 91–113.

Banerjee, A., Dolado, J. J., Galbraith, J. W. and Hendry, D. F. (1993). *Co-integration, Error Correction and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bera, A. K. and Higgins, M. L. (1993). ARCH models: Properties, estimation and testing. *Journal of Economic Surveys*, **7**, 305–366.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, **19**, 465–474.

Bhansali, R. J. (2002). Multi-step forecasting. In Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 206–221. Oxford: Blackwells.

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, **3**, 167–179.

Blair, B. J., Poon, S. H. and Taylor, S. J. (2001). Forecasting S&P 100 volatility: The incremental information content of implied volatilities and high frequency returns. *Journal of Econometrics*, **105**, 5–26.

Bollen, B. and Inder, B. (2002). Estimating daily volatility in financial markets utilizing intraday data. *Journal of Empirical Finance*, **9**, 551–562.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **51**, 307–327.

Bollerslev, T., Chou, R. S. and Kroner, K. F. (1992). ARCH modelling in finance – A review of the theory and empirical evidence. *Journal of Econometrics*, **52**, 5–59.
Bollerslev, T. and Ghysels, E. (1996). Periodic autoregressive conditional heteroscedasticity. *Journal of Business and Economic Statistics*, **14**, 139–151.
Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in finance: A review of the theory and empirical evidence. *Econometric Reviews*, **11**, 143–172.
Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day. First published, 1970.
Box, G. E. P. and Tiao, G. C. (1976). Comparison of forecast and actuality. *Applied Statistics*, **25**, 195–200.
Breidt, F. J., Davis, R. A. and Dunsmuir, W. T. (1995). Improved bootstrap prediction intervals for autoregressions. *Journal of Time Series Analysis*, **16**, 177–200.
Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **75**, 1–3.
Britton, E., Fisher, P. and Whitley, J. (1998). February 1998 Quarterly Bulletin. *Bank of England*, 30–37.
Campos, J. (1992). Confidence intervals for linear combinations of forecasts from dynamic econometric models. *Journal of Policy Modeling*, **14**, 535–560.
Canova, F. (2002). G-7 Inflation forecasts. Mimeo, Universitat Pompeu Fabra.
Carruth, A. A., Hooker, M. A. and Oswald, A. J. (1998). Unemployment equilibria and input prices: Theory and evidence from the United States. *Review of Economics and Statistics*, **80**, 621–628.
Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics*, **11**, 121–135.
Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, A*, **158**, 419–466. With discussion.
Chong, Y. Y. and Hendry, D. F. (1986). Econometric evaluation of linear macroeconomic models. *Review of Economic Studies*, **53**, 671–690. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press.
Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.
Christ, C. F. (1966). *Econometric Models and Methods*. New York: John Wiley.
Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**, 841–862.
Christoffersen, P. F. and Diebold, F. X. (1997). Optimal prediction under asymmetric loss. *Econometric Theory*, **13**, 808–817.
Christoffersen, P. F. and Diebold, F. X. (2000). How relevant is volatility forecasting for financial risk management. *Review of Economics and Statistics*, **82**, 12–22.
Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, **105**, 85–110.
Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583.
Clements, M. P. (2003). Evaluating the survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts. Mimeo, Department of Economics, University of Warwick.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *Economic Journal*, **114**, 855–877.
Clements, M. P. and Galvão, A. B. (2003). Testing the expectations theory of the term structure of interest rates in threshold models. *Macroeconomic Dynamics*, **7**, 567–585.
Clements, M. P. and Hendry, D. F. (1996). Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics*, **58**, 657–684.
Clements, M. P. and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press. The Marshall Lectures on Economic Forecasting.
Clements, M. P. and Hendry, D. F. (1999). *Forecasting Non-Stationary Economic Time Series*. Cambridge, MA: MIT Press. The Zeuthen Lectures on Economic Forecasting.
Clements, M. P. and Hendry, D. F. (eds) (2002). *A Companion to Economic Forecasting*. Oxford: Blackwells.
Clements, M. P. and Hendry, D. F. (2003). Evaluating a model by forecast performance. Mimeo, Department of Economics, University of Warwick.
Clements, M. P. and Hendry, D. F. (2004). Pooling of forecasts. *The Econometrics Journal*, **7**, 1–31.
Clements, M. P. and Smith, J. (1999). A Monte Carlo study of the forecasting performance of empirical SETAR models. *Journal of Applied Econometrics*, **14**, 124–141.
Clements, M. P. and Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment. *Journal of Forecasting*, **19**, 255–276.
Clements, M. P. and Smith, J. (2002). Evaluating multivariate forecast densities: A comparison of two approaches. *International Journal of Forecasting*, **18**, 397–407.
Clements, M. P. and Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models. *International Journal of Forecasting*, **17**, 247–267.
Clements, M. P. and Taylor, N. (2003). Evaluating prediction intervals for high-frequency data. *Journal of Applied Econometrics*, **18**, 445–456.
Corradi, V. and Swanson, N. R. (2003). Bootstrap conditional distribution tests in the presence of dynamic misspecification. Mimeo, Department of Economics, Queen Mary and Westfield College, London.
Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.
Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of The Royal Statistical Society, ser. A*, **147**, 278–292.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**, 1–38. Series B.
Deutsch, M., Granger, C. W. J. and Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, **10**, 47–57.
Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts: With applications to financial risk management. *International Economic Review*, **39**, 863–883.
Diebold, F. X., Hahn, J. Y. and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High

frequency returns on foreign exchange. *Review of Economics and Statistics*, **81**, 661–673.
Diebold, F. X., Lee, J. H. and Weinbach, G. C. (1994). Regime switching with time-varying transition probabilities. In Hargreaves, C. (ed.), *Non-stationary Time-series Analyses and Cointegration*, pp. 283–302. Oxford: Oxford University Press.
Diebold, F. X. and Lopez, J. A. (1996). Forecast evaluation and combination. In Maddala, G. S. and Rao, C. R. (eds), *Handbook of Statistics*, Vol. 14, pp. 241–268. Amsterdam: North Holland.
Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.
Diebold, F. X. and Pauly, R. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, **6**, 21–40.
Diebold, F. X., Rudebusch, G. D. and Sichel, D. E. (1993). Further evidence on business cycle duration dependence. In Stock, J. and Watson, M. (eds), *Business Cycles, Indicators, and Forecasting*, pp. 255–280. Chicago, IL: University of Chicago Press and NBER.
Diebold, F. X., Tay, A. S. and Wallis, K. F. (1999). Evaluating density forecasts of inflation: The Survey of Professional Forecasters. In Engle, R. F. and White, H. (eds), *Festschrift in Honor of C. W. J. Granger*, pp. 76–90. Oxford: Oxford University Press.
Donaldson, R. G. and Kamstra, M. (1996). Forecast combining with neural networks. *Journal of Forecasting*, **15**, 49–61.
Doornik, J. A. and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.
Doornik, J. A. and Hendry, D. F. (2001). *GiveWin: An Interface to Empirical Modelling*. London: Timberlake Consultants Press.
Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, B*, **57**, 45–97. With discussion.
Durbin, J. (1969). Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals. *Biometrika*, **56**, 1–15.
Durland, J. M. and McCurdy, T. H. (1994). Duration dependent transitions in a Markov model of U.S. GNP growth. *Journal of Business and Economic Statistics*, **12**, 279–288.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.
Engle, R. F. (1995). *ARCH*. Oxford: Oxford University Press.
Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, **5**, 1–50.
Engle, R. F. and Bollerslev, T. (1987). Modelling the persistence of conditional variances. *Econometric Reviews*, **5**, 1–50.
Engle, R. F. and Manganelli, S. (1999). CAViaR: Conditional Autoregressive Value-at-Risk by regression quantiles. UCSD Discussion Paper 99-20, Department of Economics, UCSD.
Ericsson, N. R. and Marquez, J. R. (1989). Exact and approximate multi-period mean-square forecast errors for dynamic econometric models. International finance discussion paper 348, Federal Reserve Board.
Ericsson, N. R. and Marquez, J. R. (1998). A framework for economic forecasting. *Econometrics Journal*, **1**(1), C228–C266.
Figlewski, S. and Wachtel, P. (1981). The formation of inflationary expectations. *Review of Economics and Statistics*, **63**, 1–10.

Filardo, A. J. (1994). Business–cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, **12**, 299–308.

Fildes, R. and Ord, K. (2002). Forecasting competitions – their role in improving forecasting practice and research. In Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 322–353. Oxford: Blackwells.

Fildes, R. A. and Stekler, H. O. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics*, **24**.

Findley, D. F. (1986). On bootstrap estimates of forecast mean square errors for autoregressive processes. In Allen, D. M. (ed.), *Computer Science and Statistics: The Interface*: North Holland: Elsevier Science Publishers B. V.

Franses, P. H. and van Dijk, D. (2000). *Non-linear time series models in empirical finance*. Cambridge: Cambridge University Press.

Gallant, A. R., Rossi, P. E. and Tauchen, G. (1992). Stock prices and volume. *Review of Financial Studies*, **5**, 199–242.

Garman, M. and Klass, M. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, **53**, 67–78.

Giordani, P. and Söderlined, P. (2003). Inflation forecast uncertainty. *European Economic Review*, **74**, 1037–1060.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance*, **48**, 1779–1801.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Granger, C. W. J. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, **1**, 161–173.

Granger, C. W. J. and Newbold, P. (1973). Some comments on the evaluation of economic forecasts. *Applied Economics*, **5**, 35–47.

Granger, C. W. J. and Newbold, P. (1977). *Forecasting Economic Time Series*. New York: Academic Press.

Granger, C. W. J. and Pesaran, M. H. (2000a). A decision-based approach to forecast evaluation. In Chan, W. S., Li, W. K. and Tong, H. (eds), *Statistics and Finance: An Interface*. London: Imperial College Press.

Granger, C. W. J. and Pesaran, M. H. (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, **19**, 537–560.

Granger, C. W. J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Granger, C. W. J., White, H. and Kamstra, M. (1989). Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, **40**, 87–96.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, **16**, 927–953.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, **45**, 39–70.

Hamilton, J. D. and Lin, G. (1996). Stock market volatility and the business cycle. *Journal of Applied Econometrics*, **11**, 573–593.

Hamilton, J. D. and Susmel, R. (1994). Autoregressive heteroskedasticity and changes in regime. *Journal of Econometrics*, **64**, 307–333.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. London: Chapman and Hall.

Harvey, D. I., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13**, 281–291.

Harvey, D. I., Leybourne, S. and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, **16**, 254–259.

Hendry, D. F. (1974). Stochastic specification in an aggregate demand model of the United Kingdom. *Econometrica*, **42**, 559–578. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993 and Oxford University Press, 2000.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macroeconomics: The transactions demand for money. In Ormerod, P. (ed.), *Economic Modelling*, pp. 217–242. London: Heinemann. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993 and Oxford University Press, 2000.

Hendry, D. F. (1984). Monte Carlo experimentation in econometrics. In Griliches, Z. and Intriligator, M. D. (eds), *Handbook of Econometrics*, Vol. 2, ch. 16, pp. 937–976. Amsterdam: North Holland.

Hendry, D. F. and Richard, J.-F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, **20**, 3–33. Reprinted in Granger, C. W. J. (ed.) (1990), *Modelling Economic Series*. Oxford: Clarendon Press and in Hendry D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers 1993 and Oxford University Press, 2000.

Hendry, D. F. and Richard, J.-F. (1989). Recent developments in the theory of encompassing. In Cornet, B. and Tulkens, H. (eds), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, MA: MIT Press.

Henriksson, R. D. and Merton, R. C. (1981). On market timing and investment performance. II Statistical procedures for evaluating forecast skills. *Journal of Business*, **54**, 513–533.

Holden, K. and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts. *Manchester School*, **58**, 120–127.

Hong, Y. (2001). Evaluation of out-of-sample probability density forecasts with applications to stock prices. Manuscript, Department of Economics, Cornell University.

Katz, R. W. and Murphy, A. H. (eds) (1997). *Economic Value of Weather and Climate Forecasts*. Cambridge: Cambridge University Press.

Keane, M. P. and Runkle, D. L. (1990). Testing the rationality of price forecasts: New evidence from panel data. *American Economic Review*, **80**, 714–735.

Kilian, L. (1998a). Pitfalls in constructing bootstrap confidence intervals for asymptotically pivotal statistics. Mimeo, Department of Economics, University of Michigan.

Kilian, L. (1998b). Small-sample confidence intervals for impulse response functions. *The Review of Economics and Statistics*, **80**, 218–230.

Kim, J. H. (2001). Bootstrap-after-Bootstrap prediction intervals for autoregressive models. *Journal of Business and Economic Statistics*, **19**, 117–128.

Kim, J. H. (2003). Forecasting autoregressive time series with bias-corrected parameter estimators. *International Journal of Forecasting*, **19**, 493–502.

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility : Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **81**, 361–393.

Kiviet, J. F. (1986). On the rigor of some mis-specification tests for modelling dynamic relationships. *Review of Economic Studies*, **53**, 241–261.

Kling, J. L. and Bessler, D. A. (1989). Calibration-based predictive distributions: An application of prequential analysis to interest rates, money, prices and output. *Journal of Business*, **62**, 477–499.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33–55.

Koenker, R. and Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **50**, 43–62.

Koopman, S. J., Jungbacker, B. and Hol, E. (2004). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measures. Mimeo, Department of Econometrics, Free University Amsterdam.

Lahiri, K. and Wang, J. G. (1994). Predicting cyclical turning points with leading index in a Markov switching model. *Journal of Forecasting*, **13**, 245–263.

Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the garch model. *Journal of Business and Economics Statistics*, **8**, 225–234.

Leitch, G. and Tanner, J. E. (1991). Economic forecast evaluation: Profits versus the conventional error measures. *American Economic Review*, **81**, 580–590.

Leitch, G. and Tanner, J. E. (1995). Professional economic forecasts: Are they worth their costs? *Journal of Forecasting*, **14**, 143–157.

Li, F. and Tkacz, G. (2002). A consistent bootstrap test for conditional density functions with time-series data. Discussion paper, Department of Monetary and Financial Analysis, Bank of Canada.

Lopez, J. (1996). Regulatory evaluation of Value-at-Risk models. Discussion paper 95–6, Federal Reserve Bank of New York.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. New York: Springer-Verlag.

Maddala, G. S. and Li, H. (1996). Bootstrap based tests in financial models. In Maddala, G. S. and Rao, C. R. (eds), *Handbook of Statistics*, Vol. 14, pp. 463–488. Amsterdam: North Holland.

Marris, R. L. (1954). The position of economics and economists in the Government Machine: A comparative critique of the United Kingdom and the Netherlands. *Economic Journal*, **64**, 759–783.

Masarotto, G. (1994). Bootstrapping prediction intervals for autoregressions. *International Journal of Forecasting*, **6**, 229–239.

McCracken, M. W. (2000). Robust out-of-sample inference. *Journal of Econometrics*, **99**, 195–223.

McCullough, B. D. (1994). Bootstrapping forecast intervals: An application to AR(p) models. *Journal of Forecasting*, **13**, 51–66.

McNees, S. K. (1995). Forecast uncertainty: Can it be measured? Federal reserve Bank of New York Discussion Paper.

McQueen, G. and Thorley, S. (1993). Asymmetric business cycle turning points. *Journal of Monetary Economics*, **31**, 341–362.

Mincer, J. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. (ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Mizon, G. E. (1984). The encompassing approach in econometrics. In Hendry, D. F. and Wallis, K. F. (eds), *Econometrics and Quantitative Economics*, pp. 135–172. Oxford: Basil Blackwell.

Mizon, G. E. and Richard, J.-F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica*, **54**, 657–678.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S. and Tiao, G. C. (1998). Forecasting the U.S. unemployment rate. *Journal of the American Statistical Association*, **93**, 478–493.

Morgan, W. A. (1939–1940). A test for significance of the difference between the two variances in a sample form a normal bivariate population. *Biometrika*, **31**, 13–19.

Nelson, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of the US economy. *American Economic Review*, **62**, 902–917.

Newbold, P. and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society, A*, **137**, 131–146.

Newbold, P. and Harvey, D. I. (2002). Forecasting combination and encompassing. In Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 268–283. Oxford: Blackwells.

Newey, W. K. and West, K. D. (1987). A simple positive semi-definite heteroskedasticity and autocorrelation-consistent covariance matrix. *Econometrica*, **55**, 703–708.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *Review of Economics and Statistics*, **69**, 667–674.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, **53**, 61–65.

Pascual, L., Romo, J. and Ruiz, E. (2000). Forecasting returns and volatilities in GARCH processes using the Bootstrap. Working paper 00–68, Universidad Carlos III de Madrid.

Pascual, L., Romo, J. and Ruiz, E. (2001). Effects of parameter estimation uncertainty on prediction densities: A Bootstrap approach. *International Journal of Forecasting*, **17**, 83–103.

Patton, A. J. and Timmermann, A. (2003). Properties of optimal forecasts. Mimeo, London School of Economics.

Pesaran, M. H. and Skouras, S. (2002). Decision-based methods for forecast evaluation. In Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 241–267. Oxford: Blackwells.

Pesaran, M. H. and Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business and Econometic Statistics*, **10**, 461–465.

Poon, S. H. and Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, **41**, 478–539.

Potter, S. (1995). A nonlinear approach to US GNP. *Journal of Applied Econometrics*, **10**, 109–125.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: John Wiley & Sons.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–472.

Schmidt, P. (1977). Some small sample evidence on the distribution of dynamic simulation forecasts. *Econometrica*, **45**, 97–105.

Schnader, M. H. and Stekler, H. O. (1990). Evaluating predictions of change. *Journal of Business*, **63**, 99–107.

Shenton, L. R. and Bowman, K. O. (1977). A bivariate model for the distribution of $\sqrt{b_1}$ and b_2. *Journal of the American Statistical Association*, **72**, 206–211.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.

Shephard, N. G. (1996). Statistical aspects of ARCH and stochastic volatility. In Cox, D. R., Hinkley, D. V. and Barndorff-Nielsen, O. E. (eds), *Time Series Models: In Econometrics, Finance and Other Fields*. London: Chapman and Hall.

Spanos, A. (1989). Early empirical findings on the consumption function, stylized facts or fiction: A retrospective view. *Oxford Economic Papers*, **41**, 150–169.

Stekler, H. O. (1994). Are economic forecasts valuable? *Journal of Forecasting*, **13**, 495–505.

Stekler, H. O. (2002). The rationality and efficiency of individuals' forecasts. In Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 222–240. Oxford: Blackwells.

Stine, R. A. (1987). Estimating properties of autoregressive forecasts. *Journal of The American Statistical Association*, **82**, 1072–1078.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics*, **14**, 11–30.

Stock, J. H. and Watson, M. W. (1999a). A comparison of linear and nonlinear models for forecasting macroeconomic time series. In Engle, R. F. and White, H. (eds), *Cointegration, Causality and Forecasting*, pp. 1–44. Oxford: Oxford University Press.

Stock, J. H. and Watson, M. W. (1999b). Forecasting inflation. *Journal of Monetary Economics*, **44**, 293–335.

Tay, A. S. and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, **19**, 235–254. Reprinted in Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 45–68. Oxford: Blackwells (2002).

Taylor, J. W. and Bunn, D. W. (1998). Combining forecast quantiles using quantile regression: Investigating the derived weights, estimator bias and imposing constraints. *Journal of Applied Statistics*, **25**, 193–206.

Teräsvirta, T. (1994). Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208–218.

Teräsvirta, T. and Anderson, H. M. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, **7**, 119–139.

Thombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, **85**, 486–492.

Thompson, S. B. (2002). Evaluating the goodness of fit of conditional distributions, with an application to affine term structure models. Manuscript, Harvard University.

Tiao, G. C. and Tsay, R. S. (1994). Some advances in non-linear and adaptive modelling in time-series. *Journal of Forecasting*, **13**, 109–131.

Tong, H. (1978). On a threshold model. In Chen, C. H. (ed.), *Pattern Recognition and Signal Processing*, pp. 101–141. Amsterdam: Sijhoff and Noordoff.

Tong, H. (1983). *Threshold Models in Non-Linear Time Series Analysis*. New York: Springer-Verlag.

Tong, H. (1995a). *Non-linear Time Series. A Dynamical System Approach*. Oxford: Clarendon Press. First published 1990.

Tong, H. (1995b). A personal overview of non-linear time series analysis from a chaos perspective. *Scandinavian Journal of Statistics*, **22**, 399–445.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of The Royal Statistical Society*, **B 42**, 245–292.

Tse, Y. (1999). Market microstructure of FTSE100 index futures: An intraday empirical analysis. *Journal of Futures Markets*, **19**, 31–58.

Varian, H. R. (1975). A Bayesian approach to real estate assessment. In Fienberg, S. E. and Zellner, A. (eds), *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, pp. 195–208. Amsterdam: North Holland.

Verbeek, M. (2000). *A Guide to Modern Econometrics*. John wiley & Sons, Inc.

Wallis, K. F. (1989). Macroeconomic forecasting: A survey. *Economic Journal*, **99**, 28–61. Reprinted in Mills, T. C. (ed.), *Economic Forecasting*. Cheltenham: Edward Elgar (1999).

Wallis, K. F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart. *National Institute Economic Review*, 106–112.

Wallis, K. F. (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting*, **19**, 165–176.

Wallis, K. F., Andrews, M. J., Fisher, P. G., Longbottom, J. and Whitley, J. D. (1986). *Models of the UK Economy: A Third Review by the ESRC Macroeconomic Modelling Bureau*. Oxford: Oxford University Press.

Wallis, K. F., Fisher, P. G., Longbottom, J., Turner, D. S. and Whitley, J. D. (1987). *Models of the UK Economy: A Fourth Review by the ESRC Macroeconomic Modelling Bureau*. Oxford: Oxford University Press.

Wallis, K. F. and Whitley, J. D. (1991). Sources of error in forecasts and expectations: U.K. economic models 1984–8. *Journal of Forecasting*, **10**, 231–253.

Werner, I. and Kleidon, A. (1996). UK and US trading of British cross-listed stocks: An intraday analysis of market integration. *Review of Financial Studies*, **9**, 619–664.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, **64**, 1067–1084.

West, K. D. (2001). Tests for forecast encompassing when forecasts depend on estimated regression parameters. *Journal of Business and Economic Statistics*, **19**, 29–33.

West, K. D., Edison, H. J. and Cho, D. (1993). A utility based comparison of some models of exchange rate volatility. *Journal of International Economics*, **35**, 23–45.

West, K. D. and McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review*, **39**, 817–840.

West, K. D. and McCracken, M. W. (2002). Inference about predictive ability. In Clements, M. P. and Hendry, D. F. (eds), *A Companion to Economic Forecasting*, pp. 299–321. Oxford: Blackwells.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.

Zarnowitz, V. (1985). Rational expectations and macroeconomic forecasts. *Journal of Business and Economic Statistics*, **3**, 293–311.

Zellner, A. (1986). Biased predictors, rationality and the evaluation of forecasts. *Economics Letters*, **21**, 45–48.

Index