International Series in Quantitative Marketing

Peter S.H. Leeflang
Jaap E. Wieringa
Tammo H.A. Bijmolt
Koen H. Pauwels

Modeling Markets

Analyzing Marketing Phenomena and Improving Marketing Decision Making

Springer

International Series in Quantitative Marketing

More information about this series at http://www.springer.com/series/6164

Peter S.H. Leeflang • Jaap E. Wieringa
Tammo H.A. Bijmolt • Koen H. Pauwels

Modeling Markets

Analyzing Marketing Phenomena and Improving Marketing Decision Making

Springer

Peter S.H. Leeflang
Department of Marketing
University of Groningen
Groningen, The Netherlands

Aston Business School
Birmingham UK

University of St. Gallen
Switzerland

Tammo H.A. Bijmolt
Department of Marketing
University of Groningen
Groningen, The Netherlands

Jaap E. Wieringa
Department of Marketing
University of Groningen
Groningen, The Netherlands

Koen H. Pauwels
Department of Marketing
Özyeğin University
Istanbul, Turkey

ISSN 0923-6716 ISSN 2199-1057 (electronic)
ISBN 978-1-4939-2085-3 ISBN 978-1-4939-2086-0 (eBook)
DOI 10.1007/978-1-4939-2086-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014950059

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Dick Wittink (1945–2005),
our colleague, friend,
and one of the founders
of Marketing Science.

Preface

This book is about how models can be developed to represent demand and supply on markets, where the emphasis is on demand models. Our primary focus is on quantitative models that can be used by managers to support marketing decisions. We define a model as a representation of the most important elements of a perceived real-world system. Appropriately constructed models can provide insights about structural relations between marketing variables. Since models explicate the relations, both the process of model building and the model that ultimately results can improve the quality of marketing decisions. This book is our attempt to provide a structure for model building. The content of the book should be of interest to researchers, analysts, managers and students who want to develop, evaluate, and/or use models of marketing phenomena.

Compared with only a few decades ago, marketing models have become important tools for managers in many industries. With technological advances (e.g., the availability of scanner data, improved hardware and software), the opportunity to obtain meaningful estimates of demand models vastly improved. The introduction of the Internet has also changed the landscape of marketing management in the past decade. Internet offers many opportunities to collect data from individual consumers and to use new communication vehicles such as social media.

Managers will particularly benefit from models of marketing phenomena if they understand what these models do and do not capture. With this understanding they can, for example, augment model-based conclusions with their own expertise about complexities that fall outside the modelers' purview. Importantly, the systematic analysis of purchase and other data can provide competitive advantages to managers. Model benefits include cost savings resulting from improvements in resource allocations as we discuss in various applications. And the leaders or first movers in the modeling of marketing phenomena can pursue strategies not available nor transparent to managers lagging in the use of data.

In this book we provide the basics, and we discuss the steps of the model building process. The book is suitable for student use in courses such as "Models in Marketing," "Marketing Science," and "Quantitative Analysis in Marketing" at the

graduate and advanced undergraduate level. The material can be supplemented by articles from journals such as *Journal of Marketing Research*, *Journal of Marketing*, *Marketing Science*, *Quantitative Marketing and Economics*, and *The International Journal of Research in Marketing*.

This book is a revised edition of *Building Models for Marketing Decisions* (Leeflang et al. 2000) and is the first of two volumes.

In this volume we discuss the steps of the model building process:

1. specification (Chap. 2);
2. parameterization, i.e. organizing data and estimation (Chaps. 3 and 4);
3. validation (Chap. 5);
4. re-estimation (Chap. 6).

We also spend attention to Bayesian estimation methods in Sect. 6.8 and we specify models that have been developed for markets in Chaps. 7 (models for aggregate data) and 8 (individual demand models). We observe that there is a growing attention for database marketing. In Chap. 9 we discuss models that have been developed in this area. Finally we discuss issues that deal with the final step of the model building process: Use/implementation (Chap. 10).

The second volume deals with more advanced topics that are used to model markets, such as:

- hierarchical models;
- time series models;
- state space models;
- structural models;
- spatial models;
- diffusion models;
- structural equation models;
- mixture models;
- advanced estimation methods such as nonlinear, non-parametric, and semi-parametric estimation;
- models that represent competitive reactions, including game theoretical models.

Several colleagues have contributed with their comments on various drafts. We thank the former authors: the late Dick Wittink, Michel Wedel, and Philippe Naert and colleagues/assistants of the University of Groningen, viz: Niels Holtrop, Hans Risselada, Chantal Hagen, and Roos Nijzing. We thank Harald van Heerde for providing the data for the Verhouten case.

We dedicate this book to our friend and colleague Dick Wittink who passed away unexpectedly in 2005. Marketing scientists owe much to Dick, one of the founders of Marketing Science.

Groningen, The Netherlands — Peter S.H. Leeflang
Groningen, The Netherlands — Jaap E. Wieringa
Groningen, The Netherlands — Tammo H.A. Bijmolt
Istanbul, Turkey — Koen H. Pauwels
July 2014

Contents

Chapter 1
Building Models for Markets

1.1 Introduction

Managers often use rules of thumb for decisions. For example, a brand manager may have defined a specific set of brands as the competitive set within a product category. Usually this set is based on perceived similarities in brand characteristics, advertising messages, etc. If a new marketing initiative occurs for one of the other brands, the brand manager will have a strong inclination to react. The reaction is partly based on the manager's desire to maintain some competitive parity in the marketing variables. An economic perspective, however, would suggest that the need for a reaction depends on the impact of the marketing activity for the other brand on the demand for the manager's brand. The models we present and discuss in this book are designed to provide managers with such information.

Model building in marketing started in the fifties. It is now a well-developed area with important contributions by academics and practitioners. Models have been developed to advance marketing knowledge and to aid management decision making. Closely related to the process and the activity of model building in marketing is the field of *marketing science*. One interpretation of marketing science is that it represents the scientific approach to the study of marketing phenomena. In its broadest sense this perspective should include the many disciplines relevant to marketing. However, the term marketing science has been appropriated in the early 1980s by researchers who favor quantitative and analytical approaches.

Several state-of-the-art textbooks, for example Lilien et al. (1992), Leeflang et al. (2000), Hanssens et al. (2001), Franses and Paap (2001), Blattberg et al. (2008), Wierenga (2008), Bowman and Gatignon (2010), Diamantopoulos et al. (2012) and Lilien et al. (2013) review and discuss the models developed in the marketing science discipline and applied by practitioners. The current book builds on this tradition, with a specific focus on the steps of the model building process, which should enhance the usefulness of models for marketing practitioners and scientists.

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_1

Throughout the first six chapters of this book, we use a case derived from practice to illustrate the methodology and model building process: the Verhouten case. The case is about a chocolate factory that markets several products. The goal is to develop a sales model (a "predictive" model) for one of the products. This happens in a number of steps. We describe this case in Sect. 1.2. We discuss different model types, such as a predictive model, in Sect. 1.3. This is followed by a discussion about model benefits in Sect. 1.4. The steps that are usually taken to develop a model are discussed in Sect. 1.5. Section 1.6 provides an outline of this book.

1.2 Verhouten Case

The chocolate factory Verhouten is a family company founded in 1889 by Cornelis Verhouten in Haarlem in the Netherlands. Since then, the company has grown continuously. In 1899 there were two employees and at this moment there are 125. The turnover was 6,647 guilders in the first year, whereas last year's (2014) turnover was about 100 million Euro. Currently, a grand-grand child of the founder, Frans Verhouten, is member of the board of directors.

In the first years Verhouten sold only chocolate bars of 75 g. The assortment consists now of four product classes: chocolate bars, candy bars, specialties, and seasonal products. The chocolate bars carry the brand name Verhouten and they are available in three weight classes (100, 200 and 400 g), and four flavors: milk, pure, almond-milk and white. The candy bars are sold under the brand names Tiger and Dream. The specialties are Belgian bonbons and pastilles. Seasonal products are sold in December (chocolate letters, and Christmas chocolates), and at Easter (Chocolate Easter eggs and Chocolate Easter bunnies). Verhouten products are distributed through several distribution channels. The supermarket channel distributes 75 %, candy stores 10 %, sport canteens 8 %, gasoline stations 4 %, and tobacco stores 3 %. The most important supermarkets in the Netherlands are Albert Heijn (33 % of the supermarket channel), and Jumbo. In this case we focus on the most important category for Verhouten: chocolate bars.

There are several usage motives for chocolate bars: the functional motive (to satisfy appetite), the treat-oneself motive, the sharing motive (family, friends), and the give-away motive (as a present). Chocolate bars are typically not listed on the grocery list. Instead, while browsing the shelves in a supermarket, a consumer might notice chocolate and buy a bar on impulse. In many cases, this purchase behavior is influenced by sales promotions. A display containing chocolate products with a temporary price decrease or an enlarged quantity for the same price may be an incentive to purchase. Therefore, sales promotions are used frequently in the chocolate category. Due to frequent promotions consumers adapt their purchase behavior. They purchase chocolate products more frequently during promotional periods and buy less chocolate outside these periods. As a consequence, the promotional sales spikes are very large. This has consequences for production and logistics as well. On the one hand, the distribution channel must have sufficient

stocks to anticipate these promotion-induced spikes. On the other hand, too much inventory is very expensive. To summarize, to anticipate the expected effects of sales promotions, the Verhouten management needs good sales predictions.

Management at Verhouten has shown interest to develop a sales forecasting model for chocolate bars. Such a model is also called a sales prediction model. The predictions should be accurate enough so that production and logistic decisions can be based upon them.

The model should be a test case for this company, and focuses on one chocolate bar variety sold at one supermarket chain: 100 g of milk chocolate sold at Albert Heijn. This is the single most important product for Verhouten since it has the largest turnover of all products in all channels. This test case, if successful, is to be extended to all products in all chains. But before management is willing to take this step, this test case should convince them of the usefulness of a forecasting model. The person who is responsible for this test case is Mrs. Barbara Verhouten. She is marketing director at Verhouten.

A marketing research company has been asked to perform the model building effort. The model will be developed based on "chain level" scanner data from Albert Heijn, contained in the SPSS-file **chocolate.sav**, that can be downloaded from http://www.modelingmarkets.com. "Chain level" means that all variables are at the aggregate chain level, and not at the individual supermarket level. "Scanner data" are obtained by "scanning" items at supermarket checkouts. In this case, we have four 100 g milk chocolate brands: Verhouten (brand 1), Droste (brand 2), Baronie (brand 3), and Delicata (brand 4, Albert Heijn's private label). We have 68 weekly observations. Each row in the data set is a weekly observation. The first row is from week 1 of December 2010. In the columns of the data set are the following variables:

- sales (hundreds of kilos) of Verhouten;
- prices (Euros per unit) of each of the four brands;
- variables for "feature only": weighted distribution figure at the chain level for a "feature only" (no display, see below) in a certain week. A "feature only" means special outside-store attention for the brand, either in the store flier or in an ad in a newspaper or magazine. At the individual store level this variable is measured as a dummy: = 1 for feature only, = 0 else. The weighted distribution figure means that the store-level dummy variables have been weighted with the turnovers of these stores. Hence, this variable varies between 0 (= no store has a feature-only) and 1 (= all stores have a feature-only). This variable is available for all four brands;
- variables for display only: weighted distribution figure for display-only (no feature) in a certain week. A "display only" means special inside-store attention for a brand: a temporary shelf on one of the aisles, or a change in the brand's regular shelf. At the store level this variable is measured as a dummy: = 1 for display-only, = 0 else. At the chain level it varies between 0 and 1. This variable is also available for all brands;

- variables for combined feature and display: again a weighted distribution figure. It is measured as a dummy at the store-level: = 1 for combined use of feature and display, 0 else. Hence, at the store level, we have four situations: no promotion (feature-only, display-only, feature and display all zero), feature-only promotion: only this variable is 1, others are zero, display-only promotion: only this variable is 1, the others zero, and combined feature-and-display promotion: this variable is 1, the others zero. This feature-and-display variable is available for brands 1, 3 and 4 only;
- average weekly temperature in degrees Celsius.

In this book we use this case and these data to illustrate the development of a (forecasting) model. Such a model, also called a *predictive* model, is linked to its purpose: predicting sales. There are, however, also other model purposes. We discuss these and other model typologies in Sect. 1.3.

Management of Verhouten were quite skeptic on the development of a "mathematical/statistical" model. They were not convinced about its benefits. In Sect. 1.4 we will give an overview of possible benefits.

Models are built in a number of steps. In these steps model builder and management may cooperate to build the most adequate, most reliable model. These steps are discussed in Sect. 1.5.

1.3 Typologies of Marketing Models

1.3.1 Introduction

In this section we present typologies of marketing models. First we elaborate on the distinction between decision models and models that advance marketing knowledge (1.3.2). Then we classify models according to their degree of explicitness (1.3.3). This is followed by classification according to the intended use (1.3.4) and the level of demand (1.3.5).

1.3.2 Decision Models Versus Models That Advance Marketing Knowledge

A basic distinction can be made between models that are built for the primary purpose to support decision making of a specific marketing manager versus models that aim to advance general marketing knowledge. The target audience of the former will mostly be practitioners, whereas the target audience of the latter will mostly be marketing scientists. In addition, the former will lead to case-specific insights, whereas the goal of the latter type of modeling will be generalizable insights in marketing phenomena.

Contrary to other parts of this book, in this section, we elaborate on the development of models when the primary purpose is the advancement of knowledge. New knowledge is acquired when generalizable phenomena are found, resulting in laws of marketing or empirical generalizations. In that sense, the approach is more long-term oriented and it transcends the specificity of a particular problem. An empirical generalization

> . . . "is a pattern or regularity that repeats over different circumstances and that can be described simply by mathematical, graphic or symbolic methods. A pattern that repeats but need not be universal over *all* circumstances" (Bass 1995, p. G7).

A well-known advocate to build models that advance our knowledge of models is Ehrenberg. His basic model-building philosophy is described in a book on repeat buying: Ehrenberg (1972), and various other sources (see Lehmann et al. 2011). Ehrenberg distinguishes two kinds of research traditions in marketing: the Theoretical-in-Isolation (TiI) and the Empirical-then-Theoretical (EtT) approaches (Ehrenberg 1994).

The two research traditions can each be characterized by two steps. The steps for TiI or first Theory then Empirical generalization (TE) are:

1. construct a theoretical model or analysis approach;
2. test it on a set of data.

In contrast, the steps for EtT or first Empirical generalization then Theory (ET) are:

1. establish a (generalizable) empirical pattern;
2. develop a (low-level) theoretical model or explanation.

ET seeks to first establish some empirical patterns that exist across a variety of product categories, time periods, geographic areas, etc. If the patterns hold in many different situations, generalizable findings exist and one can establish under what empirical conditions the findings generalize to form a "law". Generalizations that have already been established could, of course, be taken into account in the development of decision models, and in that sense the ET approach becomes complementary to the prevailing approach in building decision models. For example, the development of laws of marketing might lead to theoretical premises about certain model parameters, and restrictions could be imposed a priori on the parameter values. The relevance of such theoretically-based restrictions can be investigated when models are validated.

Generalizable knowledge about market phenomena can be generated in several ways (see Leeflang 2011). One way is to find *regularities* in customer behavior data, such as when Ehrenberg (1972, 1988, 1995) revealed that for most frequently purchased branded goods, market shares of brands related positively to the number of households purchasing the brands and to purchase frequency per brand. Thus, smaller brands have fewer buyers and buyers of smaller brands tend to make fewer purchases in a given period. The combination of these two negatives for brands with

smaller market shares is often referred to as "double jeopardy". Ehrenberg et al. (1990) also found that the buyers of larger brands exhibit unusual high behavioral loyalty: leading to a "triple jeopardy".[1]

Generalizable knowledge also derives from studies that cover *many circumstances* (usually many cross-sectional units, such as brands, markets or countries) and relatively long time periods. Usually *panel data* are used for that purpose. For example, Deleersnyder et al. (2009) investigate the cyclical sensitivity of advertising expenditures in 37 countries covering four key media forms (magazines, newspapers, radio and television). For 85 country-media combinations, the authors use 25 years of data to explain differences between cyclical sensitivity over media and countries. Other examples of this type of studies are Nijs et al. (2001), Steenkamp et al. (2005), Lamey et al. (2007, 2012) and Van Heerde et al. (2013).

Alternatively, *meta analyses* generate generalizable knowledge. Meta-analysis refers to the statistical analysis of results from several individual studies for the purpose of generalizing the individual findings (Bijmolt and Pieters 2001; Borenstein et al. 2009). The primary benefit of meta-analysis in marketing is that it delivers generalized estimates of various *elasticities*, *quantitative characteristics* of *buyer behavior*, and an assessment of the moderators related to the empirical and study context affecting these estimates. The assumption is that different brands and different markets are comparable at a general level but that at the same time model parameters to some extent vary systematically over different brand/model settings in an identifiable manner (Farley et al. 1995, p. G.37).

In the marketing literature an increasing number of meta-analyses and empirical generalizations are found. In one of the early studies Leone and Schultz (1980) observed that the elasticity of (selective) advertising on brand sales is positive but small. For frequently purchased branded goods they report elasticities in a range from 0.003 to 0.23. In addition, increasing store shelf space has a positive impact on sales of non-staple grocery items. In meta-analyses conducted by Assmus et al. (1984) and Lodish et al. (1995) somewhat higher advertising elasticities are reported than in the study by Leone and Schultz (1980). Lodish et al. (1995) also found that the advertising elasticities of new products are much higher than the average advertising elasticities.[2] In a meta-analysis on price elasticities, Bijmolt et al. (2005) find an average price elasticity of −2.62. They find that over the past four decades sales elasticities have significantly increased in magnitude whereas share and choice elasticities have remained fairly constant. Kremer et al. (2008) determine the effect of detailing efforts, journal advertising directed to physicians and direct to consumer advertising on sales/market share of pharmaceuticals.

[1] Dyson et al. (1997) have countered this claim though.

[2] Long-term effects of advertising on sales are studied by Ataman et al. (2010) and Sethuraman et al. (2011). Sethuraman et al. report an average short-term advertising elasticity of 0.12. The average long-run advertising elasticity is 0.24.

Empirical generalizations do not always result in "numbers". Kaul and Wittink (1995) describe the relationship between advertising and price sensitivity, and generate a set of three empirical generalizations:

- an increase in *price advertising* leads to higher price sensitivity among consumers;
- the use of *price advertising* leads to lower prices;
- an increase in *non-price advertising* leads to lower price sensitivity among consumers.

The increasing amount of research on the effects of promotions has also led to generalizations such as (Blattberg et al. 1995):

- promotions significantly increase sales;
- higher market share brands are less "promotion" elastic;
- the greater the frequency of promotions, the lower the increase in sales.

Other generalizations refer to the diffusion of new products (Arts et al. 2011), first-mover advantages (Van der Werf and Mahan 1997), repeat buying, the stationarity of market shares, the relation between market share, distribution and sales effort, etc. For a survey of generalizations in marketing, we refer to Hanssens et al. (2001, Chapter 6) and Hanssens (2009).

The generation of generalizable knowledge is not restricted to B2C research. The growth in the number of studies that address B2B marketing situations using models is well illustrated by increasing empirical generalizations about marketing impacts, as collected and summarized by Hanssens (2009). Examples of empirical "B2B generalizations" include:

- factors that affect the strength of the B2B relationships (Tuli et al. 2007);
- pioneer survival rates for industrial goods business (Robinson and Min 2002);
- pioneering in B2B (Kalyanaram et al. 1995);
- order of entry (Robinson 1988);
- industrial trade show effectiveness (Gopalakrishna and Lilien 1995).

1.3.3 Degree of Explicitness

In this section we consider a number of ways of representing the "most important elements of a perceived real world system". We distinguish implicit models, verbal models, formalized models, and numerically specified models, and illustrate different methods of representing systems with an example.

1.3.3.1 Implicit Models

In marketing practice, it is often said that managers approach problems in an intuitively appealing manner and use experience to solve problems in an ad hoc manner. One might be tempted to believe that these decision makers do not use

models. This, however, is not the case. If intuition and experience are the basis for a solution, decision makers have implicitly made use of *a model*. But the model is not necessarily recorded in a communicable form; it is *an implicit model* which so far is present only in the brain of the decision maker.

1.3.3.2 Verbal Models

The first step in making a model explicit is for a manager to state in *words* what he perceives as the important elements surrounding a problem.

We consider a simple example to illustrate the notion of a *verbal* model. Take a monopolist who produces and sells one product. For the last 5 years, price (in real dollars) has remained constant and sales per capita have also been quite stable. He wonders whether the current price is optimal in a profit maximizing sense. He believes that a price reduction will lead to increased unit sales, and a price increase to reduced unit sales. However, a price reduction will also result in a smaller unit contribution (unit price–unit cost), while a price increase will make the unit contribution higher. Thus the monopolist realizes that there is a trade-off between changes in the number of units sold and changes in the contribution per unit, and that there is a price that maximizes his total contribution to fixed costs and profit. What he is really saying is that there exists a demand curve, and by changing the (real) price he can learn how price influences demand. In this manner the monopolist can communicate elements that are the basis for a verbal model.

Our monopolist wants to determine the price leading to optimal profit by trial and error. He is aware of the fact that by using such a procedure, it is unlikely for him to obtain the exactly optimal price and corresponding profit. Specifically, he will continue to change his price until the improvement in profit, $\Delta\pi$, is smaller than a predetermined amount δ. He might also have other insights such as: "I will not increase my price by more than 20 %, because then I would be inviting potential competitors to enter the market. In the long run, my profit figure would shrink". His view of the market environment is now somewhat broader. He now wants to maximize profit, subject to a price constraint p_c, reflecting his belief in limit pricing, although he has probably never heard of this term.[3]

The verbal model can be represented as in Fig. 1.1.

> *"I will change my price in steps with each change equal to plus or minus Δp, until the increase in profit is less than a predetermined amount δ, with the restriction that price stays below the value p_c."*

Fig. 1.1 Verbal model for profit satisficing monopolist

[3] A *limit price* has the property that prices above its value will stimulate entry by competitors, whereas lower prices will discourage entry.

1.3.3.3 Formalized Models

In most marketing problems, there is a variety of variables which can play an important role. These variables may have complex effects on the demand. For a description of the relationships in words it may be difficult or even impossible to keep all relevant characteristics and conditions in mind. In order to make relationships more precise it is necessary to formalize them. This means that we specify which variables influence which other variables and what the directions of causality between these variables are. The representation of a system through formalized relationships between the most important variables of a system is called a formalized model. Within the class of formalized models we make a further distinction between logical flow models and formalized mathematical models.

A *logical flow model* represents an extension of the verbal model by the *use of a diagram*. This diagram shows the sequence of questions and of actions leading to a solution of the problem. This kind of model is also known as a *graphical* or a *conceptual model*. The flow diagram makes clear, or more explicit, what the manager has put in words. Such a diagram can serve as a basis for discussions.[4] The diagram may show discrepancies between the formalized model and the decision maker's thinking. It can also be used to identify possible inconsistencies in the model.

A *formalized mathematical model* represents a part of the real-world system by specifying relations between some explanatory (predictor) variables and some effect (criterion) variable(s).

We now return to the example of the monopolist described before. Our decision maker wants to try a price increase first, and he wants to change price in increments equal to Δp. The *logical flow model* representing the monopolist's problem is shown in Fig. 1.2.

In this figure, the following notation is used:

p_0 = original or current price,

$\Delta\pi$ = change in profit,

δ = a predetermined amount of $\Delta\pi$ used as a termination measure for the trial-and-error procedure, and

p_c = limit price above which competitive entry becomes likely.

If we assume that price is the only variable which determines sales of the product, the *formalized mathematical model* can be represented as in Fig. 1.3.

[4] A good example can be found in Van Oest et al. (2010). We give an example of a conceptual model of the Verhouten case in Sect. 2.2.3.

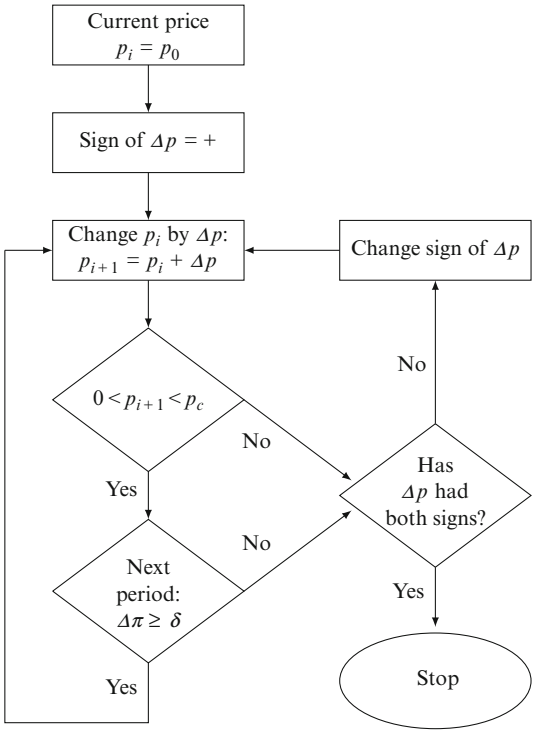

Fig. 1.2 Logical flow model for profit satisficing monopolist

$$\max_{p}(\pi) \tag{1.1}$$

$$\text{subject to } 0 < p < p_c \tag{1.2}$$

$$\text{where } \pi = (p - c)q - FC \tag{1.3}$$

$$q = f(p) \tag{1.4}$$

Fig. 1.3 Formalized mathematical model for profit optimizing monopolist

The mathematical symbols are defined as follows:

π = total profit,

p = price per unit,

c = variable cost per unit (assumed constant),

q = number of units sold,

FC = fixed costs,

$q = f(p)$ indicates that q is a function of the price per unit.

Equation (1.3) states that profit is equal to sales revenue (pq), minus variable production costs (cq), minus fixed costs (FC). This type of model is not very useful from a managerial decision making point of view because nothing is said about *how* demand (q) depends on price. In the logical flow model this relation is *approached* by trial-and-error.

1.3.3.4 Numerically Specified Models

In numerically specified models, the various components and their interrelations are quantified. Numerically specified models are, in many situations, the most appropriate representations of real-world systems.

First of all, a numerically specified model will allow the decision maker to quantify the effects of multiple, and potentially conflicting forces. Consider, for example, the monopolist decision maker who realizes that "there is a trade-off between changes in sales and changes in (unit) contribution". Specifying a model numerically will provide precision to the statements that a price increase results in a sales decrease and an advertising increase results in a sales increase.

Secondly, we may say that if a numerically specified model constitutes a reasonable representation of a real-world system, it can be used to examine the consequences of alternative courses of action and market events. Once various relationships are quantified, the decision maker can contemplate how the demand varies with price and other changes. It should be clear that such experiments are inexpensive, and less risky than the market experiments conducted by the monopolist. Thus, a numerically specified model gives management the opportunity to explore the consequences of a *myriad of actions*, a capability which cannot normally be duplicated in the real world. These considerations lead to the use of simulation models, both in the sense of providing answers to "what-if" type of questions and in the sense of dealing with stochastic (uncertain) elements. Of course, the representation of real-world systems by numerically specified models provides advantages, conditional upon the model being a reasonable representation of reality. How to construct and evaluate such representations is an important objective of this monograph.

Having presented some advantages, we should also consider disadvantages. Building and using models costs money, and the more complicated and the more explicit models become, the more expensive they will be. Thus weighing the costs against the benefits will always be a necessary step in the modeling process.

We conclude this chapter by showing in Fig. 1.4 a numerical specification for the example of the preceding sections. The symbols are the same as in Fig. 1.3. In Fig. 1.4 we assume that the fixed costs (FC) are equal to \$100.

$$\max_{p}(\pi) \tag{1.5}$$

$$\text{subject to} \quad 0 < p < p_c \tag{1.6}$$

$$\text{where} \quad \pi = (p - c)q - 100 \tag{1.7}$$

$$q = 10\, p^{-2} \tag{1.8}$$

Fig. 1.4 Numerically specified model for a profit optimizing monopolist

The difference between Figs. 1.3 and 1.4 is the numerically specified relation[5] between q and p [relation (1.8)]. This relation is of the form:

$$q = \alpha p^{\beta} \tag{1.9}$$

and is generally known as a multiplicative function or relation. The coefficients α and β are referred to as the model parameters, are unknown, but can be estimated (i.e., numerical values can be obtained) in a number of ways. In some cases, an analysis of historical data will be possible. Alternatively, we can use subjective estimation methods.

In order to illustrate how the optimal value for the price variable can be obtained, we assume that the optimal price is smaller than p_c. Once the parameters have been determined [as in Eq. (1.8)], the optimal price can be obtained by differentiating the profit function with respect to price, setting it equal to zero, and solving for price. This is shown below. We start by substituting Eq. (1.8) in Eq. (1.7):

$$\begin{aligned} \pi &= (p-c)10\,p^{-2} - 100 \\ &= 10\,p^{-1} - 10\,cp^{-2} - 100. \end{aligned} \tag{1.10}$$

[5]The terms "numerically specified marketing model" and "marketing model" will be used interchangeably from now on unless otherwise indicated.

Differentiating this equation with respect to p (assuming c to be constant):

$$\frac{d\pi}{dp} = -10\,p^{-2} + 20\,cp^{-3}. \tag{1.11}$$

Setting (1.11) equal to zero and solving for p we obtain:

$$p = 2c \tag{1.12}$$

which implies that the monopolist should use a markup of 100 %. To make sure that $p = 2c$ corresponds to a maximum, second-order conditions should be examined. For reasonable specifications of demand and profit functions, these will generally be satisfied. The expression for the second-order condition in our example is:

$$\frac{d^2\pi}{dp^2} = +20p^{-3} - 60\,cp^{-4} \tag{1.13}$$

which for a maximum should be negative.

Substituting $p = 2$c in (1.13) we get:

$$\frac{d^2\pi}{dp^2} = \frac{20}{8c^3} - \frac{60\,c}{16\,c^4} = \frac{-20}{16\,c^3} < 0 \tag{1.14}$$

which means that $p = 2c$ leads to a maximum value of π, under the reasonable assumption that $c > 0$.

We want to emphasize that the procedure described above has limited real-world applicability. Nevertheless, the example illustrates that part of a real-world system can be represented in different ways, according to the degree of explicitness chosen.

1.3.4 Intended Use: Descriptive, Predictive and Normative Models

Models can be classified according to purpose or intended use, reflecting the reason why a firm might want to engage in a model-building project. Different purposes often lead to different models. We distinguish between descriptive, predictive, and normative models.[6]

Descriptive models are intended to describe decisions or other processes. A decision maker may wonder how particular decisions are arrived at in her organization or by her customers and to that end a descriptive model may be applied.

[6]Franses (2005) discusses diagnostic tests for each of the three types of models.

The decision maker may want to trace the various steps that lead to the decisions, and identify the forces that influence the outcome(s) of the decision processes. Descriptive models also refer to models that describe demand and/or supply relations on markets. These models give answers to questions that may arise in the Verhouten case such as: which marketing instruments at Verhouten affect Verhouten's sales, or which competitive actions affect the sales of Verhouten?

The main purpose of *predictive models* is to forecast or predict future events. For example, Verhouten may want to predict sales for a brand under alternative prices, advertising spending levels, and package sizes.

The final category consists of the *normative* or *prescriptive models*. A normative model has, as one of its outputs, a recommended course of action. This implies that an objective (for example, profit or customer's satisfaction) is defined against which alternative actions can be evaluated and compared.

1.3.5 Level of Demand

The last classification that we want to discuss distinguishes models according to the level of demand. We distinguish between models for individual demand and models for aggregate demand. Aggregate demand may refer to:

1. The total number of units of a product category purchased by the population of all spending units. The corresponding demand model is called an *industry sales*, or *product class sales* model.
2. The total number of units of a particular brand bought by the population of all spending units. The demand model is then a *brand sales model*.
3. The number of units of a particular brand purchased, relative to the total number of units purchased of the product class, in which case the demand model becomes a *market share* model.[7]

We can define the same measures at the segment level and at the level of the individual consumer leading to models with different levels of aggregation: market, store, segment, household and so on. Thus we define, for example:

1. category sales for a given household;
2. brand sales for the household;
3. the proportion of category sales accounted for by the brand, for the household ("share of wallet").

The market share of brand j is equal to the ratio of brand sales of j and product class sales (i.e. brand sales summed over all brands). For example, consider total

[7]The terminology adopted here is not unique. Product class sales, brand sales, and market share models are also respectively referred to as primary demand, secondary demand, and selective (or relative) demand models. See, for example, Fischer and Albers (2010).

sales of beer in period *t*, sales of, say, Heineken in the same period, and the ratio of the latter over the former which is Heineken's market share in period *t*. Note that all these variables can be defined in units or in monetary value (e.g. dollars or euros). In model specifications, it is common that demand variables represent *unit* sales. One reason is that the monetary sales variable is the product of two variables, unit sales and price per unit, and the use of such combinations complicates the interpretation of effects on demand.

We can use two of the three types of models as part of a more complex model. For example, to predict unit sales of a brand, we can develop a product class sales model and a market share model. By multiplying these two performance measures, product class sales and market share, we obtain brand sales.

Given the classifications that we discussed before we are now able to specify what the Verhouten Management wants. They want model builders to develop:

- an *explicit decision* model, which is
- *numerically specified*, which can be used
- to *predict* sales, at
- the *brand sales* level.

1.4 Benefits from Using Marketing Decision Models

Using a marketing decision model may lead to direct and indirect benefits. Although the line between these two types of benefits is not always easy to draw, we define indirect benefits to be those that are not related directly to the reasons for which the model was built in the first place.

1.4.1 Direct Benefits

Companies invest in model building presumably because it leads to better decisions. "Better" is understood here as contributing to the fulfillment of the company's goals. For example, if the firm's single objective is to maximize profit, the benefits of a model can be defined as the incremental discounted profit generated by having the model as opposed to not having it. This requires knowledge of the amount of incremental profit over time, or of some proxy measure. Furthermore, the relevant time horizon has to be determined, and a discount rate defined.

We provide a few examples to suggest how direct benefits may materialize:

1. Suppose a model indicates that a firm is overspending on advertising, i.e. the marginal cost from advertising exceeds the marginal revenue. Adjusting the spending level will result in higher profitability.

2. A promotion budget can be allocated over different instruments such as displays, featuring (support of retailers' advertising by manufacturers), bonuses, refunds, samples, etc. A model can help in this allocation process by showing how the different instruments contribute to the profit resulting from any possible allocation.
3. Marketing managers are often faced with the question whether they should increase their advertising budget *or* whether they should decrease their prices. A marketing decision model may incorporate the empirical generalization that the optimal price depends on the advertising expenditures.[8]
4. In sealed competitive bidding, suppliers submit a price and the lowest bidder wins. Systematizing information on past bidding behavior into a model may result in a pricing strategy that will lead to an increase in expected profit.

In some cases it is difficult to measure the benefits directly while in other cases it is straightforward. The measurement is complicated by the fact that a cost–benefit evaluation should be carried out before (1) the model is built and (2) before it is implemented.

1.4.2 Indirect Benefits

A number of indirect benefits can be distinguished:

1. A marketing manager may not have to be explicit about his understanding of the environment in which he operates. Hence, he may decide on a multi-million dollar advertising budget without detailed knowledge about the effectiveness of advertising in influencing sales. A model would force him to explicate how the market works. This explication alone will often lead to an *improved understanding* of the role of advertising and how advertising effectiveness might depend on a variety of other marketing and environmental conditions. Managers may force themselves to specify how marketing activities affect consumer demand. This confrontation shows what managers believe they know well and what they are uncertain about. If the confrontation occurs in a group, the discussion can provide valuable reasons why marketing activities should provide specified types of effects. Such an exercise is also very useful prior to model estimation. It puts management beliefs on record, and this allows for a comparison between model results with those beliefs.
2. Models may work as *problem-finding* instruments. That is, problems may emerge after a model has been developed if the model outcomes are contrary to expectations. Managers may identify problems by discovering differences between their perception of the environment and a model of that environment. As

[8] See Bemmaor and Mouchoux (1991); Kaul and Wittink (1995).

an example we mention a study[9] in the Netherlands in which a negative partial relation has been discovered between the performance of bars and restaurants and the strength of the relation between beer brewers and these bars/restaurants. The stronger the relation in terms of assisting bars/restaurants by the beer brewer through effective beer tapping installations and particularly premiums, the lower the profit of the bar/restaurant.

3. Information is often available but not used. There are many examples of decisions which would have been *reversed* if available information had been used. Management may not know that data exist, or may lack methods for handling the information. Models can be instrumental *in improving the process by which decision makers deal with existing information*.
4. Models can help managers decide *what information should be collected*. Thus models may lead to improved data collection, and their use may avoid the collection and storage of large amounts of data without apparent purpose. This issue is becoming more and more relevant in the 'Big Data' era, see also Sect. 3.5.6. Clearly, model development should usually go hand-in-hand with data collection.
5. Models can also *guide research* by identifying areas in which information is lacking, and by pointing out the kinds of experiments that can provide useful information. By using models, managers have a better understanding of what they need to know and how experiments should be designed to obtain that information. To illustrate, suppose that *by estimation* of a model we learn that the average effect of advertising on sales in some periods differs from the average effect in other time periods. To explain such differences we need additional information about changes in advertising messages, the use of media, etc.
6. A model *often allows management to pinpoint changes in the environment faster* than is possible otherwise. Assume that the model-based forecasts of sales and actual sales are very close up to a certain moment in time: t. After period t, there are substantial differences between forecasted and observed values that persist in subsequent periods. This points to a very useful aspect of models namely, their *diagnostic capacity*. Since the deviation after t is larger than in previous periods, the managers may conclude that something has changed in the environment. It remains to be determined exactly what has changed, but the model warns the manager faster than is usually possible without it.
7. Models provide a *framework for discussion*. If a relevant performance measure (such as market share) is decreasing, the model user may be able to defend himself to point to the effects of changes in the environment that are beyond his control, such as new product introductions by the competition. Of course, a top manager may also employ a model to identify poor decisions by lower-level managers. Models are also used in this context to analyze the sources of profit contributions (Albers 1998).

[9] See Pleijster et al. (2011).

8. Finally, a model may result in a *beneficial reallocation of management time*, which means less time spent on programmable, structured, or routine and recurring activities, and more time on less structured ones. Examples of structured activities are media-allocation decisions, inventory decisions, decisions on the allocation of the sales force over areas, customers or products, promotion calendars, decisions on how many mailings to send to each of many (potential) customers, but also decisions on the offering of personalized products (Zhang 2011) etc.[10]

By confronting the management team of Verhouten with these direct and indirect benefits of a model, the team was convinced to proceed. Consequently, we are now ready to take the next steps in the model building process. These steps are discussed in Sect. 1.5.

1.5 The Model Building Process

Experience in model building led to the formulation of a sequence of steps for the development of mathematical models. We propose a model-building process in which model implementation has a central role: see Fig. 1.5, and we distinguish the following steps:

1. *Opportunity Identification*
   In this stage a model builder has to evaluate whether the development/use of a model can improve managerial decision making. The model builder will often be invited to consider the opportunity by a manager who is overwhelmed by demands on his/her time or who believes that the effectiveness or efficiency of decision making can be improved. Ideally the model builder and manager work together to define the problem, to agree on an approach and to determine that the expected benefits exceed the costs of model building. In the Verhouten case, the model should produce predictions that are accurate enough to base production and logistic decisions on it.
2. *Model Purpose*
   The intended use of the model should be defined as precisely as possible. For example, the manager may need a model to obtain accurate sales forecasts. The model builder needs to know the level of demand for which forecasts are required. The model builder also needs to learn what the manager believes to be the relevant determinants of demand so that model-based forecasts can be developed. In the Verhouten case the model focuses on the prediction of sales of the 100 g milk chocolate bar sold at Albert Heijn.

[10] See also Wierenga (2011).

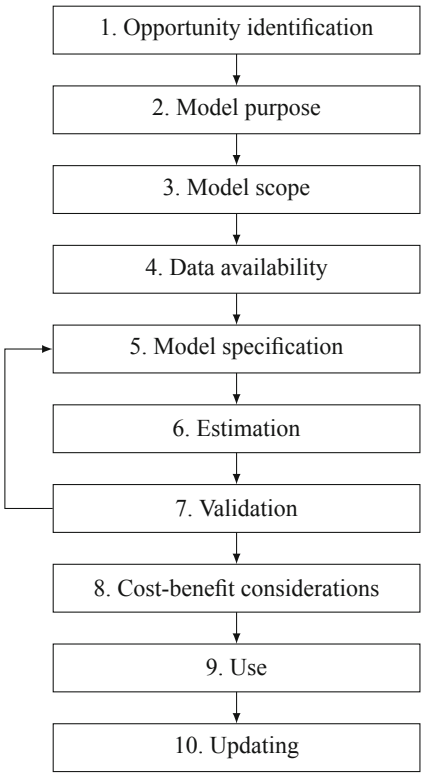

Fig. 1.5 Stages in the model-building process with an implementation focus

3. *Model Scope*
   Model building can take place for a specific type of decision or for a broader set of decisions. The manager may want the model-building effort to focus on a single decision variable. Thus, the desired outcome may be a model of advertising effects. Alternatively, the manager may desire to have a model that includes the effects of all relevant marketing activities, like is the case in the model that is developed for Verhouten.

   Similar arguments may apply to other decision variables pertaining to non-marketing activities. For example, promotions such as temporary price cuts often have strong effects on sales which may require intensive adjustments in production and distribution. The financial consequences of those adjustments need to be taken into account by the manager to determine the total profit impact of the marketing activity. Hence in such a case we need a model with a broader scope than when the model is restricted to decompose the effects of marketing variables on sales.
4. *Data Availability*
   One reason a manager may ask for a model-building effort is the increasing availability of large amounts of data (see also the discussion on "Big Data" in Sect. 3.5.6). With the introduction of scanners into supermarkets and the

increased use of customer databases, managers obtain much more detailed market feedback much more frequently.

5. *Specification* (or *Representation* or *Structure*) is the expression of the most important elements of a real-world system in mathematical terms. This involves two major steps:

   (a) Specifying the variables to be included in the model, and making a distinction between those to be explained (the dependent or criterion variables), and those providing the explanation (the explanatory, independent or predictor variables).
   (b) Specifying the functional relationship between the variables. For example, the effects of the explanatory variables can be linear or non-linear, immediate and/or lagged, additive or multiplicative, etc.[11]

6. *Estimation* is the determination of parameter estimates for a model. For this, data are needed. These data are sometimes available or can be obtained without much effort. Apart from data collection issues, we need to identify techniques to be applied for extracting estimates of the model parameters from the data collected. The choice of a technique depends on:

   - the kind of data available and/or needed;
   - the kind of variables (observable/unobservable) in the model;
   - the assumptions (of a statistical nature) that are necessary and/or acceptable;
   - the computational effort and expense considered to be reasonable.

   We note that there is often a trade-off between statistical qualities of the estimators, and flexibility (and realism) in the specification. Based on data availability, we consider: data-based parameterization (parameter estimation from historical data) and subjective estimation (judgment-based parameter estimation).

7. *Validation* (or *Verification* or Evaluation) of a model and its parameters implies assessing the quality or the success of the model. Possible criteria are:

   (a) the degree to which the results are in accordance with theoretical expectations or well-known empirical facts;
   (b) the degree to which the results satisfy statistical criteria or tests;
   (c) the degree to which the result is relevant to the original purpose:

      - is the model useful for clarifying and describing market phenomena?
      - does the model provide an acceptable degree of predictive accuracy?
      - are the model results suitable for the determination of optimal marketing-policies?

8. *Cost–Benefit Considerations*

[11] For a definition of these terms, see Sects. 2.3, 2.4 and 2.8.

We mentioned that in stage 1 *(opportunity identification)* one should establish that the expected benefits exceed the (expected) costs of model building. At this point both benefits and cost should be known with a fair amount of precision. Before the model is implemented and incorporated in a manager's decision-making process, it is appropriate to re-examine the cost–benefit trade-off. The question now is not whether the model-building effort is worthwhile. Instead, it is useful to determine if the insights gained appear to be more beneficial than the costs. One way to make this practical is to compare, in the next stage, the decisions that will be made with the benefit of the model to the decisions that would otherwise occur. In this manner it is possible to determine whether the model should in fact be used. If the model fails on this criterion, we can return to earlier stages if there is sufficient promise for a modified approach to be successful.

9. *Use*
   Use of the model requires that the manager fully understands both its strengths and its weaknesses.
10. *Updating*
    Over time, the manager may develop a better understanding of the marketplace, and this could require modifications in the model. Even without this, the continued comparison of actual outcomes with those predicted by the model may suggest that the model needs to be expanded (e.g. an additional variable or greater complexity in effects) or that the parameters need to be updated. Thus, the updating in this stage refers to updating of both the model specification and the estimation.

    The continued comparison of actual outcomes with predictions requires that differences (errors) be analyzed so that one can distinguish between errors due to e.g. model specification, measurement error, aggregation, and changes in the environment.

The "implementable model-building process" is an iterative procedure. The procedure, in its entirety or with only a few stages, can be repeated until an acceptable model is specified and estimated.

1.6 Outline

In this book we discuss in detail the steps of the model building process. In Chap. 2, we discuss model specification, in Chap. 3 the data organization, and in Chap. 4 parameter estimation in the so-called general linear model. We discuss the basic assumptions for this model, some basic test statistics and we introduce pooling issues. Testing and validation methods for detecting violations of the basic assumptions of the general linear model are considered in Chap. 5. In Chap. 6 we discuss more advanced estimation methods which account for violations of the basic assumptions of the general linear model. Chapters 7 and 8 discuss examples of

well-known models in marketing that are specified at the aggregate and individual demand level respectively. In Chap. 9 we discuss a number of specific models which have been developed for database marketing decision-making. Finally, we discuss some implementation issues in Chap. 10. In this monograph we use matrix algebra. The most important concepts are introduced in Appendix A.

References

Albers, S.: A framework for analysis of sources of profit contribution variance between actual and plan. Int. J. Res. Mark. **15**, 109–122 (1998)

Arts, J.W.C., Frambach, R.T., Bijmolt, T.H.A.: Generalizations on consumer innovation adoption: A meta-analysis on the drivers of intention and behavior. Int. J. Res. Mark. **28**, 134–144 (2011)

Assmus, G., Farley, J.U., Lehmann, D.R.: How advertising affects sales: Meta analysis of econometric results. J. Mark. Res. **21**, 65–74 (1984)

Ataman, B.M., Van Heerde, H.J., Mela, C.F.: The long-term effect of marketing strategy on brand performance. J. Mark. Res. **47**, 866–882 (2010)

Bass, F.M.: Empirical generalizations and marketing science: A personal view. Mark. Sci. **14**, G6–G19 (1995)

Bemmaor, A.C., Mouchoux, D.: Measuring the short term effect of in-store promotion and retail advertising on brand sales: A factorial experiment. J. Mark. Res. **28**, 202–214 (1991)

Bijmolt, T.H.A., Pieters, F.G.M.: Meta-analysis in marketing when studies contain multiple measurements. Mark. Lett. **12**, 157–169 (2001)

Bijmolt, T.H.A., Van Heerde, H.J., Pieters, F.G.M.: Determinants of price elasticity: New empirical generalizations. J. Mark. Res. **42**, 141–156 (2005)

Blattberg, R.C., Briesch, R., Fox, E.J.: How promotions work. Mark. Sci. **14**, G122–G132 (1995)

Blattberg, R.C., Kim, B.D., Neslin, S.A.: Database Marketing: Analyzing and Managing Customers. Springer Science, New York (2008)

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R.: Introduction to Meta-Analysis. Wiley, New York (2009)

Bowman, D., Gatignon, H.: Market Response and Marketing Mix Models: Trends and Research Opportunities. Now Publishers Inc, Delft (2010)

Deleersnyder, B., Dekimpe, M.G., Steenkamp, J.B.E.M., Leeflang, P.S.H.: The role of national culture in advertising's sensitivity to business cycles: An investigation across continents. J. Mark. Res. **46**, 623–636 (2009)

Diamantopoulos, A., Fritz, W., Hildebrandt, L. (eds.): Quantitative Marketing and Marketing Management: Marketing Models and Methods in Theory and Practice. Springer Gabler, Wiesbaden (2012)

Dyson, P., Farr, A., Hollis, N.: What does the marketing team need, description or prescription. J. Advert. Res. **37**(1), 13–17 (1997)

Ehrenberg, A.S.C.: Repeat Buying, Theory and Applications. North-Holland Publishing Company, Amsterdam (1972)

Ehrenberg, A.S.C.: Repeat-buying: Facts, Theory and Applications, 2nd edn. Oxford University Press, New York (1988)

Ehrenberg, A.S.C.: Theory or well-based results: Which comes first? In: G. Laurent, G.L. Lilien, B. Pras (eds.) Research Traditions in Marketing. Kluwer Academic, Boston (1994)

Ehrenberg, A.S.C.: Empirical generalizations, theory and method. Mark. Sci. **14**, G.20–G.28 (1995)

Ehrenberg, A.S.C., Goodhardt, G.J., Barwise, T.P.: Double jeopardy revisited. J. Mark. **54**(3), 82–91 (1990)

Farley, J.U., Lehmann, D.R., Sawyer, A.: Empirical marketing generalization using meta-analysis. Mark. Sci. **14**, G36–G46 (1995)
Fischer, M., Albers, S.: Patient- or physician-oriented marketing: What drives primary demand for prescription drugs? J. Mark. Res. **47**, 103–121 (2010)
Franses, P.H.: On the use of econometric models for policy simulation in marketing. J. Mark. Res. **42**, 4–14 (2005)
Franses, P.H., Paap, R.: Quantitative Models in Marketing Research. Cambridge University Press, Cambridge (2001)
Gopalakrishna, S., Lilien, G.L.: A three-stage model of industrial trade show performance. Mark. Sci. **14**, 22–42 (1995)
Hanssens, D.M.: Empirical Generalizations about Marketing Impact. Marketing Science Institute, Relevant Knowledge Series (2009)
Hanssens, D.M., Parsons, L.J., Schultz, R.L.: Market Response Models: Econometric and Time Series Analysis. Kluwer Academic, Boston (2001)
Kalyanaram, G., Robinson, W.T., Urban, G.L.: Order of entry: Established empirical generalizations, emerging empirical generalizations, and future research. Mark. Sci. **14**, G212–G221 (1995)
Kaul, A., Wittink, D.R.: Empirical generalizations about the impact of advertising on price sensitivity and price. Mark. Sci. **14**, G151–G160 (1995)
Kremer, S.T.M., Bijmolt, T.H.A., Leeflang, P.S.H., Wieringa, J.E.: Generalizations on the effectiveness of pharmaceutical promotional expenditures. Int. J. Res. Mark. **25**, 234–246 (2008)
Lamey, L., Deleersnyder, B., Dekimpe, M.G., Steenkamp, J.B.E.M.: How business cycles contribute to private-label success: Evidence from the United States and Europe. J. Mark. **71**(1), 1–15 (2007)
Lamey, L., Deleersnyder, B., Dekimpe, M.G., Steenkamp, J.B.E.M.: Effect of business cycle fluctuations on private-label share: What has marketing conduct got to do with it? J. Mark. **76**(1), 1–19 (2012)
Leeflang, P.S.H.: Paving the way for distinguished marketing. Int. J. Res. Mark. **28**, 76–88 (2011)
Leeflang, P.S.H., Wittink, D.R., Wedel, M., Naert, P.A.: Building Models for Marketing Decisions. Kluwer Academic, Dordrecht (2000)
Lehmann, D.R., McAlister, L., Staelin, R.: Sophistication in research in marketing. J. Mark. **75**(4), 155–165 (2011)
Leone, R.P., Schultz, R.L.: A study of marketing generalizations. J. Mark. **44**(1), 10–18 (1980)
Lilien, G.L., Kotler, P., Moorthy, K.S.: Marketing Models. Prentice-Hall, Englewood Cliffs (1992)
Lilien, G.L., Rangaswamy, A., De Bruyn, A.: Principles of Marketing Engineering, 2nd edn. DecisionPro, Inc. (2013)
Lodish, L.M., Abraham, M.M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., Stevens, M.E.: How T.V. advertising works: A meta analysis of 389 real world split cable T.V. advertising experiments. J. Mark. Res. **32**, 125–139 (1995)
Nijs, V.R., Dekimpe, M.G., Steenkamp, J.B.E.M., Hanssens, D.M.: The category-demand effects of price promotions. Mark. Sci. **20**, 1–22 (2001)
Pleijster, F., Van Der Zeijden, P., Ruis, A., Overweel, M.: Rendement en Relatie. EIM, Zoetermeer (2011)
Robinson, W.T.: Marketing mix reactions to entry. Mark. Sci. **7**, 368–385 (1988)
Robinson, W.T., Min, S.: Is the first to market the first to fail? Empirical evidence for industrial goods businesses. J. Mark. Res. **39**, 120–128 (2002)
Sethuraman, R., Tellis, G.J., Briesch, R.: How well does advertising work? Generalizations from a meta-analysis of brand advertising elasticity. J. Mark. Res. **48**, 457–471 (2011)
Steenkamp, J.B.E.M., Nijs, V., Hanssens, D.M., Dekimpe, M.G.: Competitive reactions to advertising and promotion attacks. Mark. Sci. **24**, 35–54 (2005)
Tuli, K.R., Kohli, A.K., Bharadwaj, S.G.: Rethinking customer solutions: From product bundles to relational processes. J. Mark. **71**(3), 1–17 (2007)
Van der Werf, P.A., Mahan, J.F.: Meta analysis of the impact of research methods on findings of first-mover advantage. Manag. Sci. **43**, 1510–1519 (1997)

Van Heerde, H.J., Gijsenberg, M.J., Dekimpe, M.G., Steenkamp, J.B.E.M.: Price and advertising effectiveness over the business cycle. J. Mark. Res. **50**, 177–193 (2013)

Van Oest, R.D., Van Heerde, H.J., Dekimpe, M.G.: Return on roller coasters: A model to guide investments in theme park attractions. Mark. Sci. **29**, 721–737 (2010)

Wierenga, B. (ed.): Handbook of Marketing Decision Models. Springer, New York (2008)

Wierenga, B.: The ROI of marketing management support systems. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)

Zhang, J.: The perils of behavior-based personalization. Mark. Sci. **30**, 170–186 (2011)

Chapter 2
Model Specification

2.1 Introduction

Specification is an important step in the model building process. As discussed in Chap. 1, the goal of this step is to express the most important elements of a real-world system in one or more mathematical equations. In other words: the outcome of this step is a formula that summarizes the most important relationships of the phenomenon that we are studying.

Models should satisfy certain criteria in order to stand a good chance at being implemented. In Sect. 2.2 we discuss these so-called implementation criteria with respect to model structure.

Before we turn to the discussion of the model specification process, we introduce the basic terminology related to models in Sect. 2.3. In Sect. 2.4 we discuss an important decision that a model builder needs to take in the specification step: the choice of the functional form that formalizes how the variables in the model are related. Section 2.5 discusses moderation and mediation effects. We specify formalized models for Verhouten in Sect. 2.6. In this chapter, the main focus is on models that capture marketing effects over time for a single entity (e.g. for one brand, one supermarket or one market). In Sect. 2.7, we discuss how to specify appropriate models when data for multiple entities are available. We return to this issue in Sects. 4.5 and 5.4.

Marketing is in essence dynamic. For example, we expect different sales levels over time, depending on whether the brand is in the introduction phase, in the maturity phase or in the decline phase of the product life cycle. A brand's sales will also change when e.g. more and better packaging is introduced, when competitors come and go, and so on. Another source of dynamic effects stems from the fact that the effects of, for example, advertising expenditures do not end when the campaign is over. The effects, or parts of it, remain perceptible for some future periods. In Sect. 2.8 we discuss how these dynamic effects can be incorporated in the specification of a model.

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_2

2.2 Model Criteria

2.2.1 Implementation Criteria Related to Model Structure

In this section we pay attention to the question when a model can be considered to be a 'good' model. Little (1970) proposed criteria for appropriate modeling, where he takes the model user's point of view. These criteria consider two aspects of marketing models: *model structure* and *ease of use*. In this chapter we focus on criteria that relate to *model structure*. Models should be:

1. simple;
2. complete;
3. adaptive;
4. robust.

To link the criteria "simple" and "complete", Urban and Karash (1971) introduced the notion of evolutionary model-building. This criterion is also added by Little (1975a,b) in his later work. The evolutionary criterion of model-building in fact does not relate to model structure, but to the implementation process. Evolutionary model-building is one way to reconcile simplicity with completeness.

We do not claim that any model that fails on one of the criteria is unacceptable. Rather, the criteria, described in detail below, are intended to illustrate desirable model characteristics. Thus, the more a given model satisfies each criterion, the greater its likelihood of acceptance.

2.2.2 Models Should Be Simple

All models are simplified representations of real-world phenomena. One way in which model simplicity can be achieved by the model builder is by keeping the number of variables small, and only to include the important phenomena in the model. This can be achieved in one or more of the following ways:

(a) *Eliminating Less Relevant Variables*
In situations where many explanatory variables are available for possible inclusion in a marketing model, it is tempting to include all variables. However, this interferes with the simplicity criterion, and possibly also introduces estimation problems due to multicollinearity (see Sect. 5.2.6) or lack of degrees of freedom (see "Quantity" in Sect. 3.3). An obvious first reduction of the number of variables can be realized if the least relevant are not included in the model. Potentially, this can lead to omitted variable bias (see Sect. 5.2.1), but if the variables are truly not very relevant, this bias will be small.

(b) *Clustering of Variables*
Clustering of variables is often done in econometric studies. For example, a large number of brands assumed to influence the performance of the brand

under study, may be aggregated into one variable "competition". Or marketing instruments are aggregated into a small number of classes, such as product, distribution, advertising, promotions, and price. For example, advertising expenditures separated by media such as television, radio, newspapers, magazines and billboards would be aggregated into total advertising expenditures. We note that this aggregation implicitly assumes that the marginal effect of an extra investment in advertising does not differ across the media. If the data are sufficient, this assumption should be tested.

(c) *Introducing Relative Variables*
Imagine an equation that specifies total product category expenditures (over time) in current dollars as a function of total disposable income, also in current dollars, an inflation index and total number of individuals in the population. This equation can be simplified as follows. Both product category expenditures and total disposable income can be expressed *per capita*, and in *constant dollars*. This reduces the number of predictor variables from three to one.

(d) *Phasing Variables over Different Levels*
In demand models, variables can be divided into classes according to the various levels of demand that can be distinguished. Fluctuations in product class sales per capita can be explained by fluctuations in environmental variables, such as disposable income per capita, a weather index and fluctuations in marketing instruments such as average price, product class advertising expenditures per capita, etc. And variations in market share can be explained by variations in relative or share values of the various classes of marketing instruments. The phasing of variables over different levels can be accomplished by decomposition of a dependent variable. For example, revenue for a given time period and a given territory can be decomposed as[1]:

$$\text{Revenue} = \text{Price} \times \text{Quantity} \tag{2.1}$$

and, for the same period and territory:

$$\begin{aligned}\text{Quantity} = {} & \text{Number of Buyers} \times \text{Average Size/Purchase} \\ & \times \text{Frequency of Purchase.}\end{aligned} \tag{2.2}$$

Other examples are models in which the impact of marketing variables on components such as (1) category purchase (product class sales) (2) brand choice and (3) purchase quantity decisions of households for frequently purchased goods are determined.[2]

[1] Farris et al. (1992); Lam et al. (2001); Van Heerde et al. (2003); Van Heerde and Bijmolt (2005).

[2] See, e.g., Chintagunta (1993, 1999); Van Heerde et al. (2004). Other examples of decompositions are found in Gupta (1988); Krishnamurthi and Raj (1988); Bucklin and Lattin (1991); Albers (1998); Pauwels et al. (2011) and Van Nierop et al. (2011). See also Chap. 8.

(e) *Constraining Parameter Values*
If the sales response to advertising expenditures is such that the immediate effect is largest, and that the effects die out gradually in later periods, the Koyck model is a useful model. It imposes this structure on the dynamic sales response coefficients in a parsimonious way, see Sect. 2.8.2.

Points (a)–(e) above represent different ways of obtaining a simple structure, often through a reduction in the number of variables/parameters. This relates to the concern of many model builders, especially those with a background in statistics and econometrics, that models must be manageable and estimable. This calls for *parsimony* of the models, i.e. there should be a modest number of parameters, and for a simple structure, which might mean that linear or linearizable models are preferred to non-linearizable ones.

The notions of simplicity favored by the model builder will not always be agreeable to the user. Consider Barbara Verhouten, the marketing director of Verhouten. One of her marketing mix elements is a temporary price cut or price promotion offered multiple times per year. The brand is supported by a limited amount of featuring throughout the year.

During the times of promotion, however, featuring support is usually increased, primarily to make consumers aware of the price cut. If the model builder wants to estimate separate effects of promotion and featuring, he is likely to experience difficulties because these effects are confounded, since heavy featuring spending coincides with price cut campaigns. Thus he may have to combine the two variables and measure their joint effect. This, however, may not be acceptable to the model user. For her, promotion and featuring are separate instruments, even though in some periods, one might be used to support or complement the other. Combining them for estimation purposes may result in a loss of quality and prestige of both model and model builder in the eyes of the user. In such a situation, the model builder may have to educate the model user about some statistical difficulties and how those difficulties can be reduced.

We now define more clearly what "simple" means for the user. We cannot expect managers to be experts in mathematics, statistics, econometrics, operations research and computer science. They are not, they do not pretend to be, nor do they want to be. The manager is often not interested in the detailed intricacies of the model. What she wants is a basic understanding of the logic of the model and of what it can do for her. For the user, a model will be simple if this basic understanding is provided. Communication and involvement are two means of achieving this.

Models have become easier to communicate with as a result of the widespread availability of on-line computer systems and the development of decision support systems (Lilien and Rangaswamy 2004; Lilien et al. 2013; Wierenga and Van Bruggen 2000; Wierenga et al. 1999).

To stimulate management involvement the model builder can compile a comprehensive list of important factors bearing on the problem in close cooperation with decision makers. The model structure should also represent the decision makers' view of how the market works.

Market research firms, such as AC Nielsen, GfK and IRI, and consulting firms, such as Accenture, Bain & Company, Booz & Company, the Boston Consulting Group, Forrester Research, McKinsey & Company and Roland Berger Strategy Consultants have become heavily involved in the development and estimation of market response models. Lilien (2011) argues that these "intermediaries serve as a vital boundary-spanning role in this domain". These consultants assist many companies in model-based decision-making through the development of models, customer friendly dashboards (Pauwels and Weiss 2008; Wiesel et al. 2011) and providing measures for value-based (marketing) management. Intermediaries are segmented by Lilien (2011) as:

- infrastructure vendors (SPSS);
- vendors of model solutions (Management Decision Systems);
- large generalist firms (Boston Consulting Group, McKinsey);
- implementation oriented firms (Accenture);
- accounting firms (Deloitte), and
- market research suppliers (Gallup Consulting).

Much of the development is based on client needs, and this requires that implementation of model results plays a large role. Thus, the model builders have to take into account how new models fit into the decision-making environment.

2.2.3 Models Should Be Built in an Evolutionary Way

We should, of course, realize that the real world is not simple, and that when a model represents the most important elements of a system, it will often look uncomfortably complicated. It is for this reason that Urban and Karash (1971) suggested building models in an evolutionary way, i.e. starting simple and expanding in detail as time goes on.[3]

The basic idea is that one does not build a model with all ramifications from the start. Manager and model builder begin by defining the important elements of the problem, and how these elements are related. Based on an initial meeting or set of meetings, the primary elements are specified. The manager should be fully involved, so that she is likely to understand what the model does, and is interested in this tool, because it should represent her view of the world. As the manager uses the model, and builds up experience with this decision aid, she will realize its shortcomings. The model will then be expanded to incorporate additional elements. The model is now becoming more complex, but the manager still understands it, because it is her realization that something was missing which led to the increase in complexity. In a sense the model becomes difficult, yet by using an evolutionary approach, it also remains simple because the manager has a clear understanding of what the model is supposed to do.

[3] See, for example, Van Heerde et al. (2002); Leeflang (2008).

This approach can be accompanied by a two-step presentation. First, a formalized or *conceptual* model is presented to management. This model ultimately reflects management's own views about the nature of market response. Second, an empirical model or *statistical model* is used to convey to management, how much of the overall response is captured in the model. We return to a discussion about this issue in Sect. 10.4.

Figure 2.1 is an example of a conceptual model that is constructed for the Verhouten case. It shows which variables might influence Verhouten sales.

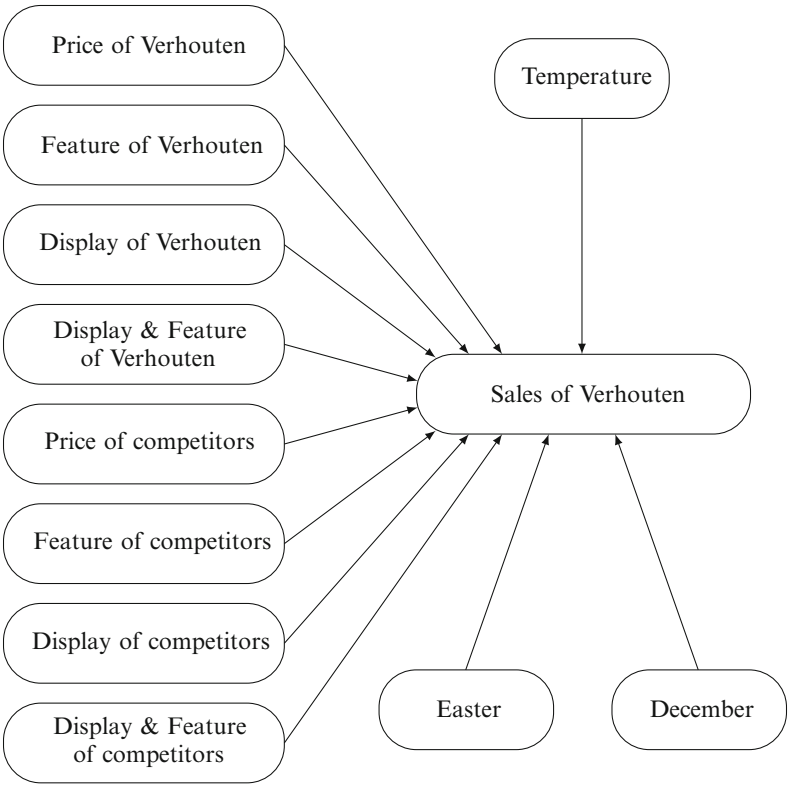

Fig. 2.1 Conceptual model for the Verhouten case

2.2.4 Models Should Be Complete on Important Issues

For a model to be a useful decision-support tool, it has to represent all relevant elements of the problem being studied. This means that a model should account for all important variables: the model should be complete. If competitors matter, then the effects of their actions on a brand under study should be incorporated. The marketing dynamics should be built in, and so on.

It should be clear that completeness on all important issues is a criterion which may conflict with simplicity. As long as simple is roughly synonymous with understandable, then the model builder can resolve the conflict at least partially by adopting an evolutionary approach.

There may be other conflicts between the interests of the model user and the needs of the model builder. Suppose, for example, that the regular price is not an active marketing instrument, a typical situation in many oligopolistic markets. It will then be difficult, or impossible, to assess its impact on sales because it does not show sufficient variation. As a result, the regular price does not appear in the specification of an econometric model. In that case, the implication is not that price does not affect sales, but that its effect cannot be measured by analyzing historical data. To have a "complete" model of demand, the effect of price has to be assessed through other means.[4]

Completeness is, of course, a relative concept. It is relative to the problem, to the organization, and to the user.

Completeness relative to the *problem* can be illustrated as follows. In modeling the effects of advertising, we may wonder whether we should focus on total advertising, i.e. estimate its overall effectiveness, or whether we should differentiate between the various media vehicles available for the communication of advertising messages. An initial answer to this question is that it depends on the problem definition. If the model is intended to aid marketing mix planning at the brand management level, then the total effect of advertising is what is needed. In that case it may be appropriate to use the aggregate of all advertising expenditures. However, if the advertising manager wants to know how to allocate expenditures to the different media vehicles, data on each of these media are required. The advertising manager needs detailed understanding of the effects of alternative media vehicles as well as of different advertising copies and specific advertising campaigns.

This delineation of data needs corresponding to the needs of different managers may, however, not be sufficient. Even if the marketing manager is only concerned with the determination of total advertising expenditures, and not its breakdown, it is possible that the estimation of the effect of total advertising is estimated with greater validity and precision from a model in which the advertising components are separated.

At the same time, the size of the *organization* may influence the desired degree of completeness. A marketing mix problem for a small firm operating at the regional level will not be the same as that of a firm selling a national brand.

The desired level of completeness will also depend on the *user*. Larréché (1974) has observed that one manager's integrative complexity, i.e. her ability to integrate pieces of information into an organized pattern, is not the same as that of another

[4]Here one may use the outcomes of laboratory experiment to determine price elasticity. See Nies et al. (2014). Other (field) experiments for this purpose are Gabor-Granger procedures and Brand-Price Trade-Off analyses. See, for example, Leeflang and Wedel (1993); Kalyanam and Shively (1998); Wedel and Leeflang (1998). Survey experiments, such as provided by conjoint analysis, offer additional opportunities; see, for example, Mahajan et al. (1982).

manager. The desired level of complexity and completeness will, therefore, vary according to the user. This amplifies the need for intense involvement of the user in the model-building process.

2.2.5 Models Should Be Adaptive

Market change and market behavior are dynamic. Thus, it is not possible to think of model building as a one-time affair. Instead, models need to be adapted more or less continuously. This implies that either the structure and/or the parameters have to be adapted. For example, the entry or exit of a competitor may imply that certain model parameters change. In addition, the specification of the model may have to change.

The changes that require model adaptation can take many forms. The true values of model parameters can change if the set of consumers that makes product category purchases changes. We discuss this issue in detail in Volume II (State Space Modeling, Dynamic Linear Models and Kalman Filtering). Brand parameters may also change if the amount of advertising for all brands or the average price in the category changes. Modifications in product characteristics, which are rarely included in demand models, can change brand-level parameters as well. These examples suggest why model parameters may vary, even if the structure of the model can remain as it is. All observable market changes that can be related to model characteristics give reason for the model builder to respecify the structure and/or reestimate the parameters. For example, if a firm has traditionally sold its products through independent distributors but has designed a wholly-owned distribution network, then the change in the structure of the selling process will require that a new model be created.

Knowledge of the marketplace, and the changes that require a new model structure, is the primary determinant of adaptation. A secondary determinant is the difference between actual and predicted values. The greater this difference, to the extent that it cannot be attributed to statistical uncertainty, the more reason there is to adapt. Thus, the greater the prediction error the greater the need to respecify the model. Is there a missing variable? Should the functional form be modified? Do consumers respond differently?

It should be clear that both the use of logical arguments ("the model needs to change because the market environment is different") and a careful analysis of prediction errors are critical for continued model use. Alternatively, one can update parameter estimates routinely, by reestimation each time new data become available. Routine reestimation is advisable if there are, for example, gradual changes in the consumer population. It is conceivable that by tracking how the parameter estimates change over time, short-term projections can be made of how the parameters will change in the future. Little (1975b, p. 662) stresses the importance of adaptive control by continually measuring programs and monitoring systems to detect change.

2.2.6 *Models Should Be Robust*

Little defines model *robustness*[5] as a quality characteristic which makes it difficult for a user to obtain bad answers. He suggest that robustness is achievable through the specification of a structure that constrains answers to a meaningful range of values. Application of the robustness criterion requires that the model builder has an understanding of marketplace behavior. This understanding is necessary for the model builder to identify relevant constructs, to define valid measures of the constructs, to specify meaningful functional forms (e.g. nonlinear effects), and to accommodate appropriate interaction effects.

Empirically, model robustness exists if the model results reflect:

(a) Correct marginal effects and changes therein. Broadly speaking, each marginal effect should be plausible over a wide range of possible values for the corresponding predictor variable (appropriate functional form).
(b) Meaningful interaction effects. The marginal effect of a given predictor variable may depend on the value of another predictor variable: for example, there is a large body of evidence to support the view that the price elasticity of a brand depends on the amount of advertising support (Kaul and Wittink 1995).
(c) The endogeneity of variables for which the models are intended to produce marginal effects. If a predictor variable is manipulated by management based on realized values of the criterion variable then this "reverse causality" must be taken into account in the model-building process (Kadiyali et al. 1999).

In practical applications, the argument is often made that it is sufficient to obtain an estimated equation that produces plausible predicted values over the range of observed values that occur in the sample for the predictor variables. This is a myopic perspective. The objective of model building is to enrich our understanding of relationships beyond historical practices, as much as possible. The best opportunity to do this is to use a model specification that can provide plausible predictions outside the range of sample values as well as within. In this regard it is useful to note that empirical researchers seldom indicate that an estimated equation can only be used for restricted ranges of values for the predictor variables.

For some criterion variables, it is possible to specify constraints. For example, if the criterion variable represents a brand's market share, robustness is violated if predicted values are less than zero or more than one. Since market share is bounded by zero and one, it is appropriate to require that *predicted* market shares satisfy the same constraints. And, if all brands belonging to a product category are modeled together, we may insist on an additional constraint: the predicted market shares should sum to one. Thus, when actual values are subject to certain constraints, their model counterparts should satisfy the same constraints. Such models are called

[5]Robustness has a different meaning in statistics and econometrics. According to Theil (1971, p. 615), a statistical test is called robust if it is insensitive to departures from the assumptions under which it is derived.

logically consistent (see, for example, Naert and Bultez 1973) or *consistent sum-constrained*. Consistent sum-constrained models are models, specified in such a way that the sum-constraint is automatically satisfied, which means that the *estimated* values of the dependent variables (e.g. market shares) or a subset of the dependent variables sum to a known number. Let q_{jt} be the demand for brand j in period t. Product class sales in period t can then be written as: $restr_{\cdot t} = \sum_{j=1}^{n} q_{jt}$, where n is the number of brands in the market. For sum-constrained models we require:

$$\sum_{j=1}^{n} \hat{q}_{jt} = \sum_{j=1}^{n} q_{jt} = restr_{\cdot t}. \tag{2.3}$$

Consistent sum-constrained models have to satisfy a number of conditions which have implications for their structure.[6]

The probability that a model will be implemented also depends on other criteria as is discussed in Chap. 10. One of the model building criteria which deserves specific attention in this respect is whether the decision model is "standardized" or not (idiosyncratic). Standardized models have a larger probability of being implemented than idiosyncratic models (Hanssens et al. 2005; Vriens 2012).

2.3 Model Elements

In this section we introduce the components or elements of a model, using a simple linear model for the Verhouten application:

$$S_t = \alpha + \beta_1 P_t + \beta_2 F_t + \beta_3 D_t + \beta_4 FD_t + \varepsilon_t, \quad t = 1, 2, \ldots, T \tag{2.4}$$

where

S_t = sales of Verhouten in week t, measured in hundreds of kilos,
α = an unknown constant (intercept),
$\beta_1, \ldots, \beta_4$ = unknown slope (effect) parameters,
P_t = price of Verhouten (in euros) in week t,
F_t = use of feature-only in week t,
D_t = use of display-only in week t,
FD_t = use of feature and display in week t,
ε_t = an error term, and
T = the number of observations/weeks.

[6] See, for example, Leeflang and Reuyl (1979).

The objective of the specification of relation (2.4) is to explain variation in the unit sales of Verhouten. Thus, S_t is the *variable to be explained, the dependent variable* or the *criterion variable*. In a model, the dependent variable is usually situated on the left-hand side of the mathematical equation.

Selection of an appropriate dependent variable is essential for successful model building. It is important that the dependent variable aligns with the intended use of the model (Sect. 1.3.4) and is measured at the desired level of demand (Sect. 1.3.5). Common examples of dependent variables for market models are brand level sales, industry or category level sales, market shares and profits. However, there are many more possibilities. For example, when a model is used to explain pricing decisions, price might be the appropriate dependent variable.

The right-hand side of the equation explains the variation in the dependent variable, and consists of two parts. The first part contains the intercept and the explanatory variables and their corresponding parameters, the second part is the error term.

The first part of the right-hand side of the equation is also referred to as the *systematic* part of the model, because it indicates how the dependent variable systematically varies with the variables that are included in the model [in model (2.4), these are P_t, F_t, D_t, and FD_t]. The latter are referred to as *independent variables, explanatory variables*, *predictors, control variables* or *regressors*.

Each independent variable is associated with a *parameter* [$\beta_1, \ldots, \beta_4$ in Eq. (2.4)], which indicates how strongly the dependent variable responds to a one-unit change in that independent variable. Therefore, they are sometimes referred to as *response parameters* or *effect parameters*. For example in Eq. (2.4), β_1 indicates the change in sales of Verhouten if a price change of one euro occurs.

The response parameters depend on the scaling of the associated independent variable and of the dependent variable: if we measure price in euro-cents instead of euros in Eq. (2.4), the associated parameter will be 100 times as small as β_1. Conversely, if we measure S_t in kilos instead of hundreds of kilos, β_1 will increase with a factor 100. Note that in the specification stage it is neither necessary to specify values for the parameters nor to indicate their sign. The most appropriate values and signs for the model parameters will be determined in the estimation step.

Selection of appropriate independent variables is very important for any model building exercise. Similar to the selection of the dependent variable, the set of independent variables should align with the intended use of the model and be measured at the right level of demand. In addition, the collection of independent variables should jointly represent the most important factors that affect the dependent variable. When important independent variables are omitted, their effect may be absorbed by (or reflected in) the estimated effects of the variables that are included in the model (omitted variable bias). This relates to the discussion about completeness in Sect. 2.2.

A sometimes overlooked consequence of omitted variable bias is that also when the model builder is interested in the effect of only a few factors, it is necessary to specify a complete model. In such situations, the model builder should not give in to the temptation to specify a too simple model with a very limited number of independent variables.

The foregoing makes clear that for the selection of an appropriate set of independent variables, a model builder needs profound knowledge of the market under study. This knowledge can be obtained from earlier studies, expert opinions or from preliminary data analyses.

The remaining element of the systematic part of an equation is the *constant term* or the *intercept* [α in Eq. (2.4)]. The constant term reflects the average value of the dependent variable if all the independent variables are zero. In Eq. (2.4) this means that α reflects the average sales level of Verhouten if price is zero, and features and displays are not used at all. Because a situation where price is zero will not occur easily, it is quite hard to interpret α in this case. This changes if the price variable and the other variables are *mean-centered* (subtract from each value of a variable the mean of that variable): then α can somewhat loosely be interpreted as the "baseline level of sales"; at least it indicates the average sales when the independent variables are at their mean level. In general, the interpretation of the constant term in a linear model depends on the centering and the scaling of the variables included in the model, and does not always have a straightforward interpretation unless all variables are properly centered and scaled.

The second part of the right-hand side of Eq. (2.4) consists of the error term and is referred to as the *stochastic* part of the model. The error term captures the variation in the dependent variable that is left after all systematic variation due to changes in the independent variables is removed, and therefore is usually characterized by a probability distribution, which assumes that this variation can be considered as being random (or *stochastic*). Although the disturbance term seems a relatively unimportant part of the model, it is actually very relevant, and the assumption for the distribution has important consequences which we will discuss further in Chap. 4.

The fact that a probability distribution is a useful way to characterize the variation in the error term does not mean that variation in the disturbances is caused by just random variation:

1. The disturbance term also represents *the error due to missing* or *omitted variables*. Obviously, the sales of Verhouten also depend on variables other than the four on the right-hand side of Eq. (2.4), such as e.g. competitor's marketing actions, environmental variables, etc. Excluding such variables from the deterministic part of the model means that their effects become part of the disturbance term. Despite the fact that strictly speaking, their individual effects on the dependent variable cannot be considered as random, a probability distribution is a useful way to characterize their joint effect on the dependent variable. Possible reasons for omitting variables are that no data are available, or that neither the manager nor the model builder imagines their relevance.
2. The disturbance term is also affected by *error in the functional relationship*. For example, in Eq. (2.4) the relation between sales and the independent variables is assumed to be linear. If the price response is actually nonlinear, a linear model is inappropriate, and deviations between the linear model and the observed nonlinearity in the relationship between price and sales is reflected in the disturbance term.

3. The variation in the disturbance term is also affected by *errors in measurement* of the variables. One reason for this type of error is sampling. For example, in the Verhouten case, the value for D_t that is recorded in a given week (which is a weighted distribution figure, see Sect. 1.2) may be based on an audit of a selection of stores. This can be considered random error if a random sample of the total set of stores is used, but there may also be systematic measurement error if some outlets are unavailable for inclusion (e.g. a retailer may refuse the market research company access to her store).

   Errors in measurement may also result from poor measurement instruments or from the use of approximate (or proxy) variables. For example, when studying the effect of price sensitivity, a researcher may measure this by an item like: "I often buy items on promotions" in a survey or by the proportion of items bought that are on promotion, both being non-perfect representations of true price sensitivity.

All elements of Eq. (2.4) have now been defined.

2.4 Specification of the Functional Form

Once the dependent and the independent variables are selected, a model builder should decide on the type of mathematical relationship between the variables. In this section we focus on the following most common types of mathematical forms:

1. linear in both parameters and variables;
2. nonlinear in the variables, but linear in the parameters;
3. nonlinear in the parameters and linearizable;
4. nonlinear in the parameters and not linearizable.

The distinction is important from the point of view of estimation. Forms 1–3 are estimable by classical econometric methods, whereas 4 is not.

2.4.1 Models Linear in Parameters and Variables

Models linear in parameters and variables have the following structure:

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t \tag{2.5}$$

where

y_t = value of the dependent variable in period t,

x_{kt} = value of independent variable k in period t, $(k = 1, \ldots, K)$, and

$\alpha, \beta_1, \ldots, \beta_K$ = model parameters.

A model that is linear in parameters and in variables is sometimes also referred to as a linear *additive* model. It is additive in the sense that the joint effect of the predictor variables is equal to the sum of their separate effects, as can be seen in Eq. (2.5).

While this is the simplest functional form that can be used for specification of the relationships between y_t and the x-variables, it also has serious drawbacks.

The linearity assumption implies constant returns to scale with respect to each of the independent variables. This can be seen by taking the first-order partial derivative of y_t with respect to any of the independent variables x_{kt}:

$$\frac{\partial y_t}{\partial x_{kt}} = \beta_k, \quad k = 1, \ldots, K \tag{2.6}$$

which means that increasing x_{kt} by one unit results in an increase of y_t by β_k units.

This assumption of constant returns to scale is unreasonable in most applications. For example, if x_{kt} is advertising and y_t is sales, we might expect an increment in x_{kt} to have more effect when x_{kt} itself is lower than when it is higher. This means that we expect advertising to have decreasing returns to scale.

Another drawback of the linear additive model is that it assumes no *interactions* between the variables. Interaction between two x-variables occurs when the effect of one x-variable on the y-variable depends on the level of the other x-variable. In such cases, the latter x-variable is said to *moderate* the effect of the first x-variable on y.

The absence of interaction effects in Eq. (2.5) can again be seen by looking at the first-order derivative in (2.6). Since it is constant, the effect of x_{it} on y_t does not depend on the values of other independent variables. This assumption is also often unreasonable. For example, price promotions will have a greater effect on sales if the brand is available in more rather than in fewer retail stores. At decreasing levels of availability, price promotions should have increasingly smaller effects. In Sect. 2.5 we discuss how the linear additive model can be extended to include interaction effects.

2.4.2 Models Linear in Parameters But Not in Variables

A second class of models consists of models which are nonlinear in the variables, but linear in the parameters.[7] They are also called *nonlinear additive models*. Equation (2.7) is an example of a model in which three variables (x_{1t}, x_{2t}, and x_{4t}) are assumed to have nonlinear effects:

$$y_t = \alpha + \beta_1 e^{x_{1t}} + \beta_2 \sqrt{x_{2t}} + \beta_3 x_{3t} + \beta_4 \ln x_{4t} + \varepsilon_t. \tag{2.7}$$

[7] See for an in-depth analysis of the shape of (advertising) response functions, for example, Vakratsas et al. (2004).

Equation (2.7) can be turned into a linear additive model relatively easy by transforming the nonlinear variables as follows before they enter the model:

$$\begin{aligned} x_{1t}^* &= e^{x_{1t}} \\ x_{2t}^* &= \sqrt{x_{2t}} \\ x_{4t}^* &= \ln x_{4t}. \end{aligned}$$

After substituting the transformed variables in Eq. (2.7) a linear additive model emerges:

$$y_t = \alpha + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + \beta_3 x_{3t} + \beta_4 x_{4t}^* + \varepsilon_t. \tag{2.8}$$

The proposed relation between each independent variable and the dependent variable should be based on *theory* or *experience*. If we know that advertising shows decreasing returns to scale we can focus on appropriate mathematical formulations. In addition, as we discussed in Sect. 2.2, model-building criteria can provide direction with regard to an appropriate model specification.

We next discuss a few formulations of nonlinear additive models, with their characteristics, advantages, and disadvantages. Consider the following relation:

$$q_t = \alpha + \beta_1 a_t + \beta_2 a_t^2 + \varepsilon_t \tag{2.9}$$

where

q_t = sales in units of the focal brand in period t,

a_t = advertising expenditures for the focal brand in period t.

To understand the nature of effects, we investigate the first derivative of q_t with respect to a_t:

$$\frac{\partial q_t}{\partial a_t} = \beta_1 + 2\beta_2 a_t.$$

If $\beta_1 > 0$ and $\beta_2 < 0$, we have decreasing returns to scale, because due to $\beta_2 < 0$, the marginal sales are smaller for larger values of a_t. But if

$$a_t > \frac{\beta_1}{-2\beta_2}$$

sales would decline with further increases in advertising. This phenomenon is known as *supersaturation* (Hanssens et al. 2001, p. 42). Supersaturation results if excessive marketing effort causes a negative response. If this phenomenon runs counter to our prior beliefs, we could reject the model specified in Eq. (2.9). Nevertheless, the model may perform well within a certain range of values of a_t.

Indeed, if we have no confidence in the estimated model outside a specific range of variation, we should make this explicit. We discussed this point in more detail in Sect. 2.2.

There are other ways to represent decreasing returns to scale. In (2.10) sales are a function of the *square root* of the independent variable, say advertising,

$$q_t = \alpha + \beta \sqrt{a_t} + \varepsilon_t. \tag{2.10}$$

The first-order derivative of q_t with respect to a_t is:

$$\frac{\partial q_t}{\partial a_t} = \frac{\beta}{2\sqrt{a_t}}$$

which shows decreasing returns to scale, and tends to zero when a_t is very large, indicating that the marginal sales effect approaches zero for very high levels of advertising.

Another frequently used mathematical form is the *semi-logarithmic* specification:

$$q_t = \alpha + \beta \ln a_t + \varepsilon_t \tag{2.11}$$

where $\ln a_t$ is the natural logarithm of a_t.

Equation (2.11) shows decreasing returns to scale over the whole range of a_{jt}, since:

$$\frac{\partial q_t}{\partial a_t} = \frac{\beta}{a_t}$$

which decreases with a_t. Again, returns to advertising tend to zero for high levels of advertising.

The sales-advertising relations (2.9), (2.10), and (2.11) all represent decreasing returns to scale. All three however, are deficient for high values of advertising: the first because for high values of a_t, q_t starts to decline; the second and the third because q_t tends to infinity when a_t tends to infinity. Since we know that sales potential is a finite quantity, we prefer sales-advertising models in which sales approaches a saturation level as advertising grows large.

A simple example of a model with a saturation level is the *reciprocal relation*:

$$q_t = \alpha + \beta \frac{1}{a_t} + \varepsilon_t, \qquad \text{with } \alpha > 0,\ \beta < 0. \tag{2.12}$$

As a_t increases, q_t approaches α asymptotically. Note that if $a_t < -\beta/\alpha$, q_t is negative. Thus, while a reciprocal relation leads to a finite asymptote for q_t when a_t increases, it can still be problematic for very low values of a_t.

Another example of a reciprocal relation is:

$$q_t = \alpha + \beta \frac{1}{p_t} + \varepsilon_t, \qquad \text{with } \alpha, \beta > 0 \tag{2.13}$$

where

$$p_t = \text{price of the focal brand in period } t.$$

It is clear that (2.13) may not be meaningful for extremely low values of p_t: as p_t goes to zero, q_t goes to infinity. Equation (2.13) may, therefore, only be a reasonable approximation of reality within a restricted range of values for p_t.

If one wants a response function to show increasing returns to scale first, then decreasing returns, the *logarithmic reciprocal* relation may be used:

$$\ln q_t = \alpha + \beta \frac{1}{a_t} + \varepsilon_t. \tag{2.14}$$

This relation shows increasing returns to scale for $a_t < -\beta/2$, and decreasing returns for $a_t > -\beta/2$, if we assume that $\alpha > 0$ and $\beta < 0$.

2.4.3 *Models That Are Nonlinear in Parameters, But Linearizable*

One of the most frequently encountered marketing response functions that are nonlinear in the parameters (they appear as exponents), is the so-called *multiplicative model*:

$$y_t = \alpha x_{1t}^{\beta_1} x_{2t}^{\beta_2} \cdots x_{Kt}^{\beta_K} \varepsilon_t \tag{2.15}$$

or more compactly:

$$y_t = \alpha \left(\prod_{k=1}^{K} x_{kt}^{\beta_k} \right) \varepsilon_t. \tag{2.16}$$

Model (2.15) can be linearized using a simple transformation. Taking the logarithm of (2.15) we obtain:

$$\ln y_t = \ln \alpha + \beta_1 \ln x_{1t} + \beta_2 \ln x_{2t} + \cdots + \beta_K \ln x_{Kt} + \ln \varepsilon_t. \tag{2.17}$$

Equation (2.17) is linear in the parameters $\alpha^*, \beta_1, \beta_2, \ldots, \beta_K$, where $\alpha^* = \ln \alpha$, and is referred to as a *double-logarithmic* or a *log-log* model.

The response function (2.15) has several desirable properties. For ease of interpretation it is convenient to assume throughout that the dependent variable y_t represents sales in units of the focal brand in period t.

Firstly, Eq. (2.15) allows for a specific form of interaction between the various instruments. This can easily be seen by looking at the first-order derivative with respect to, say, variable x_{kt}:

$$\frac{\partial y_t}{\partial x_{kt}} = \alpha\beta_k x_{1t}^{\beta_1} x_{2t}^{\beta_2} \cdots x_{kt}^{\beta_k - 1} \cdots x_{Kt}^{\beta_K} \varepsilon_t$$

which can be written as:

$$\frac{\partial y_t}{\partial x_{kt}} = \frac{\beta_k y_t}{x_{kt}}. \tag{2.18}$$

The impact of a change in x_{kt} on y_t is therefore a function of y_t itself, which means that it depends not only on the value of x_{kt} but on all the other variables as well.

Secondly, model (2.15) has a simple economic interpretation. Let us consider η_k, the *elasticity* of y_t with respect to variable x_{kt}, which is defined as:

$$\eta_k = \frac{\partial y_t}{\partial x_{kt}} \frac{x_{kt}}{y_t}.$$

Using (2.18) we find $\eta_k = \beta_k$, which means that the exponents in a multiplicative response model are constant elasticities. This is a disadvantage if one wants η_k to depend on one or more of the independent variables.

To see the functional form of relations in the multiplicative model, consider the case with only one explanatory variable:

$$y_t = \alpha x_{1t}^{\beta_1} \varepsilon_t. \tag{2.19}$$

Figure 2.2 shows (2.19) for various values of β_1, relative to the case where $\beta_1 = 1$ (dashed upward sloping line). Curve I represents the case $\beta_1 > 1$, i.e. increasing returns to scale. Curve II is typical for $0 < \beta_1 < 1$, i.e. decreasing returns to scale. This is what we might expect if x_{1t} were advertising. Curve III illustrates the case $-1 < \beta_1 < 0$, and finally curve IV, $\beta_1 < -1$. The latter two might apply when x_{1t} is a price variable, curve III representing inelastic demand and curve IV elastic.

There are many other nonlinear models that can linearized besides the multiplicative model. One is the *exponential model*:

$$y_t = \alpha e^{\beta x_t} \varepsilon_t \tag{2.20}$$

which, after taking logarithms, becomes linear in the parameters $\alpha^* (= \ln \alpha)$ and β:

$$\ln y_t = \alpha^* + \beta x_t + \ln \varepsilon_t.$$

This model may, also, with β negative, be appropriate for a sales–price relation.

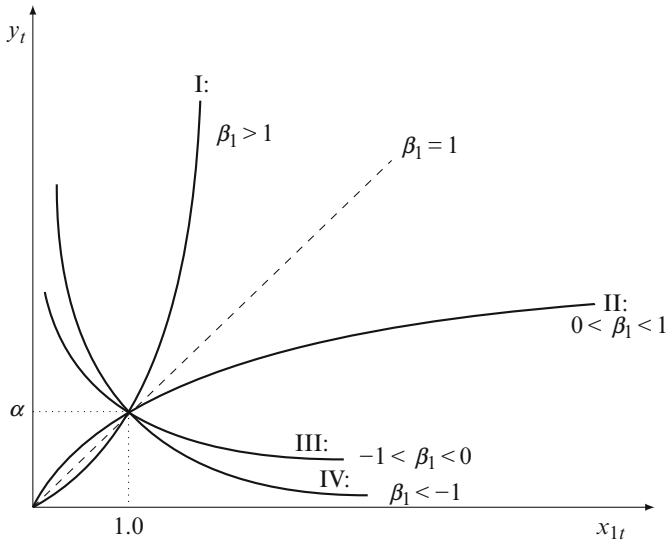

Fig. 2.2 Examples of multiplicative relations

For price (x_t) equal to zero, sales equal α, whereas for price going to infinity, sales tend to zero. However, for $\beta > 0$, (2.20) has no saturation level. Yet for almost all products in virtually any market it holds that no matter how much marketing effort is spent, there is a finite upper limit for sales. We show later how the saturation level is a finite quantity in the modified exponential model (2.21), which is an intrinsically nonlinear model.

2.4.4 Models That Are Nonlinear in Parameters and Not Linearizable

Marketing relations are generally nonlinear in either variables or parameters or both. In some cases these relations are linearizable while in other cases they are not. If the model is not linearizable, the model is called *intrinsically nonlinear* or *intractable*. In the past, model builders often went to great efforts to make their models linearizable. This was primarily due to the fact that estimation methods in econometrics generally assumed models to be linear in the parameters. In the mean time, powerful nonlinear estimation techniques have been developed, and are nowadays available in statistical software packages. Thus, from an estimation point of view, intrinsic nonlinearity is no longer problematic in a purely technical sense. It remains true, however, that the statistical properties of nonlinear estimation techniques are not as well known as those of linear models.

As an example of an *intrinsically nonlinear* model, consider the *modified exponential model*:

$$y_t = \alpha(1 - e^{-\beta x_t})\varepsilon_t, \quad \text{with } \alpha > 0, \text{ and } \beta > 0. \tag{2.21}$$

If x_t equals zero, y_t also equals zero. As x_t goes to infinity, y_t approaches α asymptotically. An interesting characteristic of the model is that the marginal sales response is proportional to the level of untapped potential $(\alpha - y_t)$. This is easily demonstrated as follows. The first-order derivative of y_t with respect to x_t is:

$$\frac{\partial y_t}{\partial x_t} = \alpha\beta e^{-\beta x_t}. \tag{2.22}$$

Untapped potential is $\alpha - y_t = \alpha e^{-\beta x_t}$, and it follows from (2.22) that the marginal sales response is proportional to $(\alpha - y_t)$, with β serving as the proportionality factor.

2.5 Moderation and Mediation Effects

In Sect. 2.4.1 we briefly introduced interaction or moderation effects. Schematically, an interaction effect between x_{it} and x_{2t} on y_t can be depicted as in Fig. 2.3.

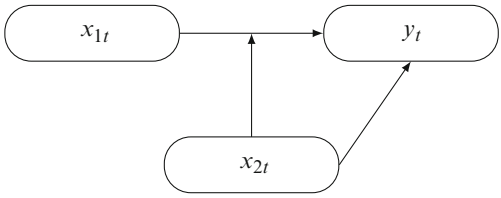

Fig. 2.3 Schematical representation of an interaction effect

Congruent with the definition of interaction that we discussed in Sect. 2.4.1, Fig. 2.3 illustrates that the effect of x_{1t} on y_t depends on x_{2t}. Alternatively, one can say that x_{2t} affects the strength of (or moderates) the effect of x_{1t} on y_t.

It is possible to extend the linear additive model so that it accommodates interaction effects by adding the product of the variables that are likely to have interaction effects. For example, with two predictors x_{1t} and x_{2t} we can add the product $x_{1t}x_{2t}$:

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 (x_{1t}x_{2t}) + \varepsilon_t. \tag{2.23}$$

The effect of a marginal change in x_{1t} on y_t is now:

$$\frac{\partial y_t}{\partial x_{1t}} = \beta_1 + \beta_3 x_{2t}. \tag{2.24}$$

Compared to Eq. (2.6), we see that by adding the product of x_{1t} and x_{2t} as an additional variable to the model the marginal effect of x_{1t} on y_t is no longer constant, but changes with x_{2t} (assuming nonzero β_3, i.e. there is a relevant interaction effect). For example, if x_{1t} is advertising, and x_{2t} distribution, measured by the number of retail stores carrying the brand, (2.23) allows for advertising to have a larger effect if more stores sell the brand.

Note that we cannot distinguish between the situation where the effect of x_{1t} on y_t is moderated by x_{2t} from the case where the effect of x_{2t} on y_t is moderated by x_{1t} (i.e. x_{1t} and x_{2t} can be reversed in Fig. 2.3).

A disadvantage of the interaction term formulation becomes apparent when the number of predictor variables exceeds two. For example, with three predictors, a full interaction model becomes:

$$\begin{aligned} y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} \\ + \beta_4 (x_{1t} x_{2t}) + \beta_5 (x_{1t} x_{3t}) + \beta_6 (x_{2t} x_{3t}) + \beta_7 (x_{1t} x_{2t} x_{3t}) + \varepsilon_t \end{aligned} \tag{2.25}$$

and

$$\frac{\partial y_t}{\partial x_{1t}} = \beta_1 + \beta_4 x_{2t} + \beta_5 x_{3t} + \beta_7 (x_{2t} x_{3t}) .$$

In general, with K predictor variables, a full interaction model contains 2^K terms. It can easily be seen that both estimation and interpretation will become problematic, even for fairly small values of K. Thus, it is often necessary for a model builder to specify in advance which of many possible interaction variables to include. For further in-depth discussion of moderated multiple regression models, we refer to Irwin and McClelland (2001).

A recent example of a model that accounts for moderating effects is Koschate-Fischer et al. (2014).[8] They demonstrate that the relationship between private label share and store loyalty is moderated by factors such as:

1. customers' price orientation;
2. the commodization of the product category;
3. product category involvement and
4. the retailer's price positioning;

Moderated models are sometimes confused with *mediation models*. Mediation models are models that aim to explain the relationship between a independent variable, say x_{1t} and a dependent variable y_t via a third explanatory variable, say x_{2t}, known as the mediator variable. For instance, the effect of marketing efforts (x_{1t}) on sales (y_t) may be mediated by consumer attitudes/brand health indicators (x_{2t}), as demonstrated in Hanssens et al. (2014) . Baron and Kenny (1986) is the seminal paper on mediation, with over 40,000 citations. They offer the diagrams in Fig. 2.4 to understand mediation.

[8] See also Horváth and Fok (2013).

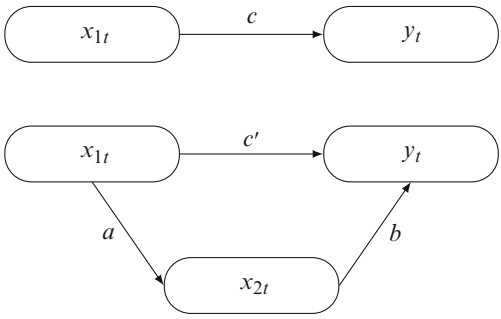

Fig. 2.4 Schematical representation of a mediation model

The unmediated model in the top diagram shows the total effect (path c) of x_{1t} on y_t. In the mediated model (bottom diagram), path a is the effect of x_{1t} on mediator x_{2t}, path b is the effect of mediator x_{2t} on y_t, and path c' is the direct effect of x_{1t} on y_t. Complete mediation occurs when path c' is 0; i.e. x_{1t} no longer affects y_t after x_{2t} has been controlled for. Partial mediation occurs when c' is not zero but smaller than c; i.e. the direct effect of x_{1t} on y_t becomes smaller after including the mediator variable.

Continuing our marketing example, marketing may have both a direct, unmediated effect on sales (e.g. price promotions and reminder advertising increase transactions with current brand customers) and an indirect effect through improving awareness, consideration and/or liking among category consumers. Hanssens et al. (2014) call path a the "responsiveness" of brand health indicators to marketing actions, path b the conversion of these brand health indicators into sales, and path c' the "transactions route" of marketing influence on brand sales. Interestingly, market-response modeling literature has focused on the unmediated model, while advertising and attitude literature has focused on responsiveness (path a) and marketing metrics literature mostly on sales conversion (path b). Only recently, these elements were integrated by Srinivasan et al. (2010); Pauwels et al. (2013) and Hanssens et al. (2014).

2.6 Formalized Models for the Verhouten Case

At the end of Sect. 1.3 we concluded that Verhouten Management wants model builders to develop:

- an *explicit decision* model, which is
- *numerically specified*, which can be used
- to *predict* sales, at
- the *brand sales* level.

This means that the first three steps of the model building process as discussed in Sect. 1.5 are concluded. Also, the data is available (Sect. 1.2), so that we are now

ready to proceed with step 5. We specify two formalized static models that satisfy the implementation criteria as well as possible. The first model (2.26) is a linear additive model, the second one (2.27) has a multiplicative specification.

Linear additive sales model:

$$\begin{aligned} S_{1t} = \alpha + \sum_{j=1}^{n} \beta_{1j} P_{jt} + \sum_{j=1}^{n} \beta_{2j} F_{jt} + \sum_{j=1}^{n} \beta_{3j} D_{jt} \\ + \sum_{j=1}^{n} \beta_{4j} FD_{jt} + \beta_5 Temp_t + \beta_6 \delta_{1t} + \beta_7 \delta_{2t} + \varepsilon_{1t} \end{aligned} \tag{2.26}$$

where, in week t,

S_{jt} = sales of brand j (Verhouten is brand 1) at Albert Heijn,

P_{jt} = price of brand j,

F_{jt} = use of feature-only of brand j,

D_{jt} = use of display-only of brand j,

FD_{jt} = combined use of feature and display of brand j,

$Temp_t$ = average weekly temperature in Celsius,

δ_{1t} = a dummy variable to account for weeks in December,

δ_{2t} = a dummy variable to account for the week before Easter,

ε_{1t} = a disturbance term.

Multiplicative model:

$$S_{1t} = \theta \left[\prod_{j=1}^{n} \left(P_{jt}^{\gamma_{1j}} \gamma_{2j}^{F_{jt}} \gamma_{3j}^{D_{jt}} \gamma_{4j}^{FD_{jt}} \right) \right] Temp_t^{*\gamma_5} \gamma_6^{\delta_{1t}} \gamma_7^{\delta_{2t}} \varepsilon_{1t}. \tag{2.27}$$

The parameter θ affects the overall level of the model outcomes across all periods: its function is comparable to that of α in the linear additive model. We usually do not interpret its value. The γ_{1j} $(j = 1, \ldots, n)$ parameters in the model can be interpreted as the own (γ_{11}) and cross-elasticities $(\gamma_{1j}, j \neq 1)$. The variables F_{jt}, D_{jt}, FD_{jt}, δ_{1t} and δ_{2t} are used as exponents given that these variables can attain zero values in some weeks. In such cases, S_{1t} will be zero at t, which is not realistic. The associated parameters, $\gamma_{2j}, \gamma_{3j}, \gamma_{4j}, \gamma_6$ and γ_7 are so-called multipliers. To avoid these issues for the temperature variable we transformed it to $Temp_t^*$, which is defined as the temperature measured in degrees Kelvin.

Both *static* models (2.26) and (2.27) are simple. They do not account for leads and lags and nonlinearities of the variables (2.26). Hence, they are not complete, yet.

There is no guarantee that (2.26) is a robust model given that there is no guarantee that the estimated value at S_{1t} ($\hat{S}_{1t}$) are always larger than or equal to zero. The multiplicative model (2.27) only specifies sales predictions that are larger than zero and is robust in this sense. In a broader sense, robustness also means that the parameters have face validity. Here, it means that the own price elasticities are expected to be negative and the cross-price elasticities to be positive. We also expect own multipliers to be larger than one, and cross-multipliers to lie in the range between zero and one.

By adding or deleting variables in an evolutionary way and accounting for varying parameters over time, the model allows for adaptiveness.

2.7 Including Heterogeneity

Thus far, we have only considered models for a single entity in this chapter. For example, in the Verhouten case, we have considered a sales model for the Verhouten brand only. However, in many applications, data of multiple entities is available, such as data for different stores, regions or brands. This has two consequences for the model specification step.

The first consequence is that we need to carefully indicate which entity we are considering. In many cases, this has consequences for the mathematical notation that is used. Often, additional indices are required to indicate precisely which entity is referred to in a model. For example, up until Eq. (2.26) the variable S_t did not have an brand index, because we were focusing on the sales of Verhouten only, and there was no ambiguity when referring to the sales variable. If we would include sales variables of the other brands as well, we would replace S_t with S_{it}, where the index i refers to the different brands in our data set ($i = 1,\ldots,4$). This allows us to precisely indicate which sales observations we are referring to.

The second consequence is that we need to make decisions how to treat the entities in our model specification. We consider four approaches:

1. aggregation across the entities;
2. specify a different model for each entity (unit-by-unit models);
3. specify a model that is the same for each entity (pooled models);
4. specify a model where some parameters are pooled, and others are entity-specific (partially pooled models).

Ad 1: *Aggregated* models.
A first option that can be considered is to aggregate across the entities, and specify a model on a higher aggregation level. For example, in the Verhouten case, we might aggregate all variables in the data set across brands and specify a model for the category sales of chocolate bars at Albert Heijn. Advantages of such an approach are that we do not need to account for differences between entities, and that model outcomes are less affected by potential entity-specific irregularities in the data: due to the aggregation process, such irregularities are 'averaged' with data from the other

entities. A disadvantage is that aggregation changes the demand level, so that the parameters are not specified at the right level of demand. This potentially interferes with Step 3 of the model building process (Fig. 1.5), where the scope of the model is determined. In the Verhouten case, the intended level of demand was set at the brand level. If we aggregate across brands, this changes the level of demand to primary demand. This will have consequences for the selection of independent variables, and the decisions that a manager can make based on the outcomes of the model. Another disadvantage is that a model builder loses the opportunity to accommodate differences across entities.

Ad 2: *Unit-by-unit* models.
A second approach is to specify a model for each entity. This generates a separate set of parameter estimates for each entity, and provides maximum flexibility/heterogeneity for accommodating all potential differences between the entities. This is the preferred approach when there are enough data points available for estimating all separate models reliably, and when there is interest in obtaining entity-specific parameter estimates. However, in cases where there is a limited amount of data, this approach does not exploit the benefits of combining data across the entities, and does not provide statistical efficiency benefits in estimating the parameter values. In other cases, it is not interesting to obtain individual-level parameters, or it is impossible to interpret all parameters. For example, if we would estimate separate store-level models for all Albert Heijn stores in the Netherlands, we would need to interpret more than 800 sets of model outcomes, one for each store or for each store/brand combination, which is an unduly task, especially for managers that suffer from scarcity of time.

Ad 3: *Pooled* models.
A third approach is to specify exactly the same model for each entity, and to utilize the observations of all entities to estimate the parameters in the model. This approach is quite restrictive because it requires that all parameters do not differ across the entities. For example, specifying exactly the same model for each of the four brands in the Verhouten case requires that the intercepts of all brands have the same value. This is quite unrealistic because not all brands will have the same base line level of sales. On the positive side, if it is reasonable to assume that all parameters are the same this strategy ensures that we use all available data points as efficiently as possible.

Ad 4: *Partially pooled* models.
The last option that we consider here is when some parameters are restricted to be the same across entities, whereas others are allowed to differ between entities. This approach provides statistical efficiency, because it combines data from multiple categories wherever possible, and accommodates differences across entities wherever needed. Specifying exactly the same model for each of the four brands in the Verhouten case requires that customers are equally price sensitive, irrespective of the brand they typically buy. This might not be very realistic as each of the brands target different segments in the market, as reflected by their price positioning.

Figure 2.5 illustrates the last three options for the Verhouten case. For clarity of presentation, we assume that we are specifying a simple model, where sales of a brand is only explained by price. Also, we assume that we only have data of two brands. In the top panel of Fig. 2.5, the price response and the intercept appear to be different. Therefore, a unit-by-unit model is a reasonable model for these data. The corresponding formula indicates that both the intercept and the price response parameter are brand specific. The middle panel of Fig. 2.5 represents a situation where a pooled model appears to be reasonable. The formula shows that both the intercept and the price response are assumed to be the same for both brands. The bottom panel of Fig. 2.5 depicts a typical situation where partial pooling can be applied: the intercepts differ between the two brands, whereas the price responses (slopes) are identical. This is reflected in the corresponding formula: the intercepts are brand specific, whereas the price response is the same for both brands.

Choosing the right level of pooling in the specification step can be based on prior research or earlier model building efforts, or on subjective knowledge of the market, or on exploratory data analysis. We return to the issue of pooling in Sect. 4.5, where we discuss how to estimate pooled models. In Sect. 5.4 we consider statistical pooling tests that allow us to validate the pooling choices that are made in the model specification step.

2.8 Marketing Dynamics

2.8.1 Introduction

Marketing is dynamic in essence. For example, new products are developed, tested, and launched, more and better packaging is introduced, competitors come and go, and so on. These effects can be accounted for by adding appropriate variables to the model that control for these developments. Another type of marketing dynamics occurs when the effects of, for example, advertising expenditures do not end when the campaign is over. The effects, or parts of it, remain noticeable for several future periods. Or, looked at somewhat differently, sales in period t will be affected by advertising in t, but also by expenditures in $t-1, t-2, \ldots$. Thus, one can refer to the *lagged effects* of advertising. Similarly, one can also observe a dynamic sales response when consumers or competitors anticipate a marketing stimulus and adjust their behavior before it actually occurs. This results in *lead effects* of marketing expenditures. Together, these effects are known as dynamic effects of marketing variables, or simply *marketing dynamics*.[9] In Sect. 2.8.2 we first discuss the modeling of marketing dynamics in the case of one explanatory variable. Issues

[9]The term 'Marketing dynamics' is also used to indicate a much more broader area of time-series modeling; see Volume II. This topic has received much attention in the past decade: See for some surveys for example Pauwels et al. (2004) and Leeflang et al. (2009).

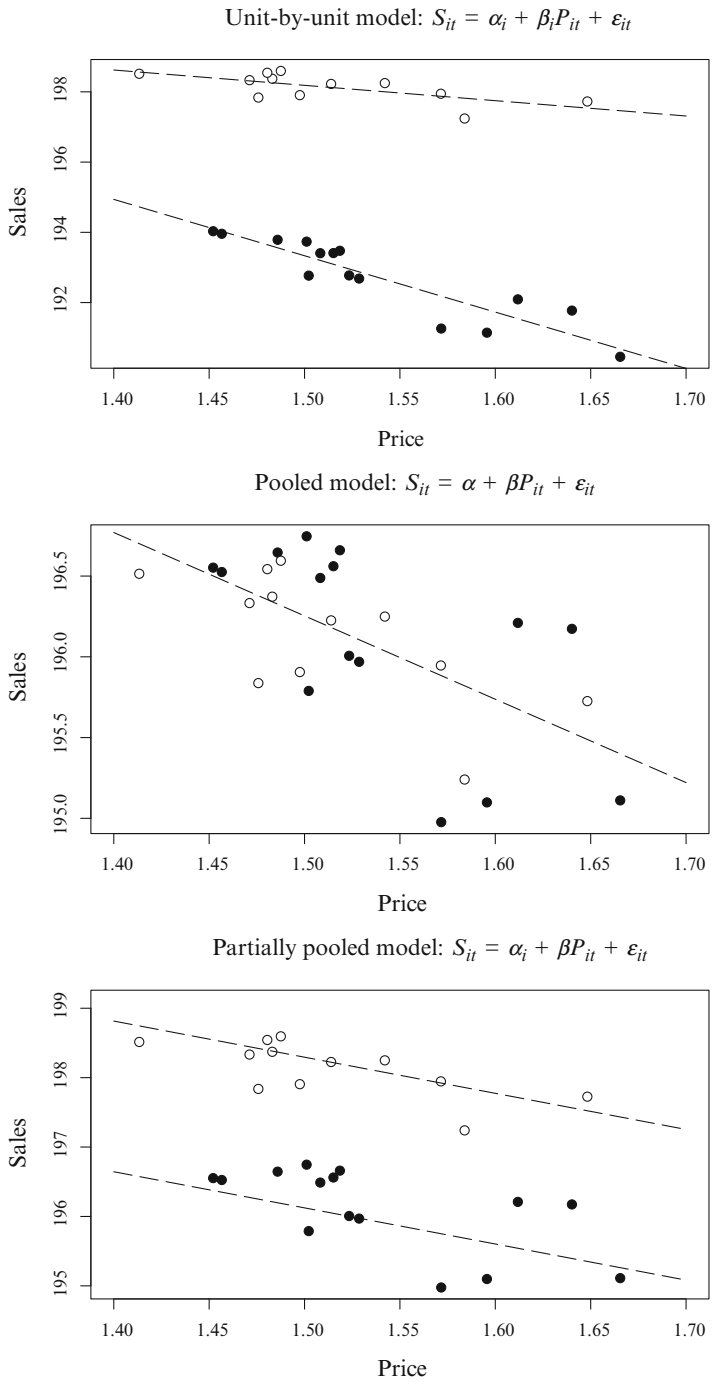

Fig. 2.5 Three different options to treat brand-level heterogeneity

related to having more than one independent variable are dealt with in Sect. 2.8.3. In Sect. 2.8.4 we give attention to lead effects. These dynamic effects are known to occur for all marketing variables, but we take the relationship between advertising and sales as a typical example throughout this section.

2.8.2 *Modeling Lagged Effects: One Explanatory Variable*

The sales effects of an advertising campaign in period t are generally not limited to that period, but will also have an effect in future periods. These lagged effects of marketing activities on sales have long been recognized as one of the complicating features of market measurement. Its existence means that promotional programs should be evaluated over a longer period than that of the campaign.

We make a distinction between delayed-response effects and customer-holdover effects.

Delayed-response effects arise from the delay between marketing money being spent and sales occurring.[10] It can happen because of:

- execution delay: the time between management spending money or preparing an ad and its appearance;
- noting delay: the time between a magazine being published and read;
- purchasing delay: the time between a consumer receiving the stimulus and a purchase being made.

The *delayed-response* effect can be simply represented by a relation such as:

$$q_t = \alpha + \beta a_{t-s} + \varepsilon_t \tag{2.28}$$

where

$$\begin{aligned} q_t &= \text{sales in period } t, \\ a_{t-s} &= \text{advertising expenditures in period } t-s, \\ \varepsilon_t &= \text{a disturbance term,} \end{aligned}$$

and s is the number of time periods between, say, the time the advertising money is spent and the sales that result from this expenditure.

Customer-holdover effects occur because customers sometimes make repeated purchases for some time, e.g. at times $t, t+1, t+2, \ldots, t+s$ after the initial stimulus in t, either because of:

- new buyer holdover effects, where marketing activity attracts new customers who make repeat purchases, or

[10] See, Lilien and Kotler (1983, p. 80), and Leeflang et al. (1992).

- increased purchase holdover effects, where the marketing stimulus increases the average quantity purchased per period for some time.

Taking delayed response and customer holdover effects together, the following demand function represents a reasonable specification of the dynamic effects of advertising:

$$q_t = \alpha + \beta_1 a_t + \beta_2 a_{t-1} + \cdots + \beta_{s+1} a_{t-s} + \varepsilon_t. \tag{2.29}$$

Equation (2.29) indicates that advertising has an effect up to s periods into the future. Given that s is a finite number, (2.29) is called a *finite distributed lag model* of *order s*.

The specification of lagged effects in the demand function (2.29) is difficult for at least three reasons. First, one has to decide how many lagged terms to include in the specification, i.e. one has to decide how long the advertising effect lasts. In what follows, we refer to this as the *duration interval*, and determining the correct value of s is called the *truncation problem*. Secondly, with s lagged terms in the model, the number of parameters becomes $s+2$, which might be a large number in relation to the number of observations. Since data over time may not be available, statistical problems are likely to arise because of loss of *degrees of freedom* when estimating the model in the next model building step. This results both from an increase in the number of parameters and from a decrease in the number of usable observations. If s lagged effects are included in the model, the first s observations cannot be used for estimation because $q_1, q_2, \ldots, q_s$ all require one or more observations of advertising expenditures before $t = 1$. These are not available, because we assume that $t = 1$ represents the start of the data collection period. A third difficulty with direct specification of lags is that the larger s, the higher the chance that the explanatory variables become collinear.[11]

The three reasons above have led researchers to consider relations between the various β_i in (2.29) in order to arrive at a simpler model. A simpler model will have fewer parameters and will be less troublesome to estimate. Model (2.29) is known as an *unrestricted model*.

Many alternative specifications of a relation between the β_i's have been proposed in literature. We consider the most common ones.[12] The most popular lag structure in marketing studies is the *Geometric lag model*, *Koyck model* or *exponential decay model*.

[11] Collinearity or multicollinearity relates to the correlation between explanatory variables. A more exact definition will be given in Chap. 4. We limit ourselves here to stating that collinearity has the disadvantage of making the coefficients less reliable.

[12] For extensive surveys, we refer to Judge et al. (1985, Chapters 9 and 10) at the more theoretical level, and Leeflang et al. (1992); Hanssens et al. (2001, Chapter 7) and Van Heerde et al. (2000) in a marketing context.

First Eq. (2.29) is rewritten with an infinite number of terms, which means that the entire advertising history is taken into account:

$$q_t = \alpha + \sum_{\ell=0}^{\infty} \beta_{\ell+1} a_{t-\ell} + \varepsilon_t. \tag{2.30}$$

It is now assumed that:

$$\begin{aligned} \beta_2/\beta_1 &= \lambda, \\ \beta_3/\beta_2 &= \lambda, \text{ or} \\ \beta_3/\beta_1 &= \lambda^2, \\ &\vdots \\ \beta_{\ell+1}/\beta_1 &= \lambda^{\ell}, \\ &\vdots \end{aligned} \tag{2.31}$$

where $0 \leq \lambda < 1$. The lag structure specified in (2.31) assumes that the advertising effect is geometrically (or exponentially in the continuous case) decreasing over time.

Substituting (2.31) in (2.30) we obtain:

$$q_t = \alpha + \beta_1 a_t + \beta_1 \lambda a_{t-1} + \beta_1 \lambda^2 a_{t-2} + \cdots + \beta_1 \lambda^{\ell} a_{t-\ell} + \cdots + \varepsilon_t. \tag{2.32}$$

Now we lag equation (2.32) by one period and multiply by λ:

$$\lambda q_{t-1} = \lambda\alpha + \beta_1 \lambda a_{t-1} + \beta_1 \lambda^2 a_{t-2} + \cdots + \beta_1 \lambda^{\ell+1} a_{t-\ell-1} + \cdots + \lambda\varepsilon_{t-1}. \tag{2.33}$$

Subtracting (2.33) from (2.32) gives:

$$q_t - \lambda q_{t-1} = \alpha(1-\lambda) + \beta_1 a_t + \varepsilon_t - \lambda\varepsilon_{t-1}. \tag{2.34}$$

With $\alpha^* = \alpha(1-\lambda)$, and $u_t = \varepsilon_t - \lambda\varepsilon_{t-1}$, we obtain:

$$q_t = \alpha^* + \lambda q_{t-1} + \beta_1 a_t + u_t. \tag{2.35}$$

The estimation problems are now greatly reduced since only three parameters remain. The procedure is due to Koyck (1954), and is generally referred to as the Koyck transformation.

The direct (*short term*) effect of advertising is represented by β_1, while λ measures how much of the advertising effect in one period is retained in the next. One often refers to λ as the *retention rate*. The *long-term* advertising effect is obtained as follows. If the advertising investment is kept constant ($a_t = a$ for all t),

Table 2.1 Implied duration interval as a function of the data interval

Data interval	$\hat{\lambda}$	90 % duration interval in months	Number of studies
Weekly	0.537	0.9	2
Monthly	0.440	3.0	10
Bimonthly	0.493	9.0	10
Quarterly	0.599	25.1	10
Annual	0.560	56.5	27

Source: Clarke (1976, p. 351)

in equilibrium we have $q_t = q_{t-\ell} = q$ for all t and all ℓ. Or from (2.35), and omitting the error term:

$$q = \alpha^* + \lambda q + \beta_1 a \qquad \text{or:}$$

$$q = \alpha + \frac{\beta_1}{1-\lambda} a.$$

Thus $\beta_1/(1-\lambda)$ measures the (total) long term effect of advertising.[13]

The Koyck model, however, is not without problems. If ε_t satisfies the assumptions of the classical linear regression model, then u_t does not.[14] We do not discuss this issue here, but refer to Chaps. 5 and 6.

The duration interval of advertising implied by the Koyck model has been studied by Clarke (1976) on the basis of a survey of 59 cases. Since (2.35) is an additive model, the same relation holds for, say, $q_t^* = q_t + q_{t-1}$ and $a_t^* = a_t + a_{t-1}$. Hence the implied duration interval should not vary with the periodicity of the data, expressed by the data interval. However, from the figures in Table 2.1 it is clear that the implied duration interval is a function of the periodicity of the data. The table shows the average value of $\hat{\lambda}$, and the average 90 % duration interval, which means the time it takes for advertising to reach 90 % of its total effect, for each of the five data intervals. The results indicate a large increase in the implied duration time as the data interval increases, pointing to a data interval bias.[15]

Leone (1995) provides a theoretical explanation for the inconsistent findings from previous econometric analyses of aggregated data concerning the duration

[13] In an applied setting, estimated values $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\lambda}$ will be used. These estimates are obtained from the numerical specification of (2.35). Since $\hat{\beta}_1$ and $\hat{\lambda}$ are estimated, $\hat{\beta}_1/(1-\hat{\lambda})$ will be estimated as well. One difficulty is that the distributional properties of that ratio are not well known.

[14] In particular, if the residuals ε_t of the original model are uncorrelated, then the u_t must be autocorrelated.

[15] The problem is even more striking when very different implied duration intervals are obtained from one and the *same* data set. A case in point is the relation between industry sales and industry advertising expenditures for the West-German cigarette market, which has been estimated by Leeflang and Reuyl (1985) using annual, bimonthly and monthly data covering the same time period (1960–1975).

of advertising carry-over effects. He also adjusted the lagged sales parameter in models with a data interval/aggregation bias. He found that the average carry-over effect is between 6 and 9 months. Similar conclusions are based on a meta-analysis performed by Assmus et al. (1984) and correspond with the findings of Lodish et al. (1995) and Sethuraman et al. (2011).

We now turn our attention to the *shape of the lagged effects*. A geometrically decaying lag structure implies that a campaign in period t has its greatest effect in the same period. This may or may not be realistic, depending, among other things, on the periodicity of the data. For example, decreasing lagged effects might be quite reasonable for annual or even for quarterly data. With monthly data, it is possible that advertising will reach its peak effect after a few periods. For example, in their study of the dynamic effects of a communications mix for an ethical drug, Montgomery and Silk (1972) found that direct mail had its peak effect in the month after the mailing. Samples and literature similarly peaked after a 1-month lag. Journal advertising showed a peak, although a modest 1, 3 months after the advertising appeared in the medical journals.

There are various ways of dealing with such more complex lag structures. The most obvious way is to include a number of direct lags, and let the geometric decay take effect after a few periods. In that case, the sales-advertising equation becomes, for example:

$$q_t = \alpha + \beta_1 a_t + \beta_2 a_{t-1} + \beta_3 a_{t-2} + \beta_4 a_{t-3} + \beta_4 \lambda a_{t-4} + \beta_4 \lambda^2 a_{t-5} + \cdots + \varepsilon_t. \tag{2.36}$$

Applying the Koyck transformation to (2.36), we obtain after rearranging terms:

$$\begin{aligned} q_t = {} & \alpha(1-\lambda) + \lambda q_{t-1} + \beta_1 a_t + (\beta_2 - \lambda\beta_1) a_{t-1} \\ & + (\beta_3 - \lambda\beta_2) a_{t-2} + (\beta_4 - \lambda\beta_3) a_{t-3} + \varepsilon_t - \lambda\varepsilon_{t-1} \end{aligned} \tag{2.37}$$

and the relation to be estimated is:

$$q_t = \alpha^* + \lambda q_{t-1} + \beta_1 a_t + \beta_2^* a_{t-1} + \beta_3^* a_{t-2} + \beta_4^* a_{t-3} + u_t \tag{2.38}$$

where

$$\begin{aligned} \alpha^* &= \alpha(1-\lambda) \\ \beta_\ell^* &= \beta_\ell - \lambda\beta_{\ell-1}\,, \qquad \ell = 2,3,4 \\ u_t &= \varepsilon_t - \lambda\varepsilon_{t-1}. \end{aligned}$$

Equation (2.38) contains six parameters. Their estimates suffice to obtain estimates of the parameters in the original model (2.37). Although this formulation allows for more flexibility in the nature of lagged effects, it reintroduces some of the difficulties (loss of degrees of freedom and multicollinearity) which we want to avoid.

The combination of direct lags with geometrically declining lags has been applied by a number of authors, see e.g. Doyle and Saunders (1985), Leeflang et al. (1992), or Van Heerde et al. (2000). See for an advanced approach that accounts for joint lagged effects of several advertising themes: Bruce (2008).

Almon (1965) proposed a method to estimate a model with finite distributed lags. The relation between the β_ℓ's ($\ell = 1, \ldots, s$) in (2.29) and the lag length is approximated by a continuous function of the form:

$$\beta_\ell = \phi_0 + \phi_1 \ell + \phi_2 \ell^2 + \cdots + \phi_r \ell^r, \qquad r \leq s. \tag{2.39}$$

This equation is a polynomial in ℓ, and if r is strictly less than s, the use of this approximation imposes restrictions on β_ℓ, $\ell = 1, \ldots, s$. The parameters ϕ_i may be estimated by substituting (2.39) into (2.29). This model has been widely used in applied econometric work because of the flexibility of the polynomial lag shape, the decrease in the number of parameters that must be estimated, the ease of estimation, and the reduction in multicollinearity. In marketing this model has been applied by Van Heerde et al. (2000) to model dynamic effects of promotions.

In the marketing literature several other distributed lag models have been developed and applied. We discuss two of them. We take as a starting point a simple linear additive model, which is in this context also referred to as the *current-effects model*:

$$q_t = \alpha + \beta a_t + u_t. \tag{2.40}$$

In the current-effects model it is assumed that the disturbances u_t are *not* correlated with the disturbances in preceding periods. If u_t is correlated with earlier disturbances, for example, with u_{t-1} for $t = 1, 2, \ldots$, the residuals are said to be autocorrelated. In that case we have:

$$u_t = \rho u_{t-1} + \varepsilon_t \tag{2.41}$$

where

$$\rho = \text{the autocorrelation coefficient, } \rho \neq 0, \text{ and}$$

$$\varepsilon_t = \text{an error term, which is not autocorrelated.}$$

Subtracting ρq_{t-1} from (2.40), we get:

$$q_t - \rho q_{t-1} = \alpha(1-\rho) + \beta(a_t - \rho a_{t-1}) + \varepsilon_t \tag{2.42}$$

which is a linear relationship between a change in the level of demand and a change in the level of advertising expenditure. If there is *no* autocorrelation and $\rho \approx 0$, (2.42) reduces to (2.40). However if $\rho \neq 0$ we have a *dynamic specification*. This *autoregressive current-effects model* (2.42) is often used by researchers (Kanetkar et al. 1986).

Another well-known dynamic model is the *partial adjustment model* which is specified as follows:

$$q_t = \alpha + \beta a_t + \lambda q_{t-1} + \varepsilon_t. \tag{2.43}$$

The partial-adjustment model (2.43) and the geometric lag model (2.35) have similar structures, which indicates that it may be very difficult to discriminate the different distributed lag models. Testing procedures have been proposed by e.g. Bass and Clarke (1972) and Weiss and Windal (1980). We return to these tests in Sects. 5.6.2 and 5.6.3.

2.8.3 Modeling Lagged Effects: Several Explanatory Variables

So far, we have just considered the dynamics of advertising. In this subsection we discuss how lagged effects of multiple marketing variables can be included in a sales model. As a start, price (p_t) and distribution (d_t) can be included as additional variables in Eq. (2.35):

$$q_t = \alpha^* + \lambda q_{t-1} + \beta_1 a_t + \beta_2 p_t + \beta_3 d_t + u_t. \tag{2.44}$$

This may seem quite plausible. It implies, however, that price, distribution and advertising all have the same decay in their lagged effects. This can be seen by realizing that Eq. (2.44) was obtained by applying the Koyck transformation to:

$$\begin{aligned} q_t = {} & \alpha + \beta_1 a_t + \beta_1 \lambda a_{t-1} + \beta_1 \lambda^2 a_{t-2} + \cdots + \beta_2 p_t + \beta_2 \lambda p_{t-1} \\ & + \beta_2 \lambda^2 p_{t-2} + \cdots + \beta_3 d_t + \beta_3 \lambda d_{t-1} + \beta_3 \lambda^2 d_{t-2} + \cdots + \varepsilon_t, \end{aligned} \tag{2.45}$$

which shows that for all three variables in the model, the effect of the first lag equals $\lambda\times$ (immediate effect), the effect of the second lag equals $\lambda^2\times$ (immediate effect), and so on.

To state that price and advertising have the same lag structure is a *heroic* assumption. It is generally accepted that price responses occur much faster. Suppose that price and advertising are the only explanatory variables, and that both follow a geometrically declining lag structure, but with different parameters λ_1 and λ_2. The basic model becomes:

$$\begin{aligned} q_t = {} & \alpha + \beta_1 a_t + \beta_1 \lambda_1 a_{t-1} + \beta_1 \lambda_1^2 a_{t-2} + \cdots \\ & + \beta_2 p_t + \beta_2 \lambda_2 p_{t-1} + \beta_2 \lambda_2^2 p_{t-2} + \cdots + \varepsilon_t. \end{aligned} \tag{2.46}$$

The model transformation now requires *two* steps. First, Eq. (2.46) is lagged one period, multiplied by λ_1 and subtracted from (2.46), yielding:

$$q_t - \lambda_1 q_{t-1} = (1-\lambda_1)\alpha + \beta_1 a_t + \beta_2 p_t + (\lambda_2 - \lambda_1)(\beta_2 p_{t-1} + \lambda_2 \beta_2 p_{t-2} + \cdots) + \varepsilon_t - \lambda_1 \varepsilon_{t-1}. \tag{2.47}$$

Now (2.47) is lagged one period, multiplied by λ_2, and subtracted from (2.47). After rearranging terms, one obtains:

$$q_t = (1-\lambda_1)(1-\lambda_2)\alpha + \beta_1(a_t - \lambda_2 a_{t-1}) + (\lambda_2 - \lambda_1)\beta_2 p_{t-1} + \beta_2(p_t - \lambda_1 p_{t-1}) + (\lambda_1 + \lambda_2) q_{t-1} - \lambda_1 \lambda_2 q_{t-2} + u_t \tag{2.48}$$

with

$$u_t = \varepsilon_t - (\lambda_1 + \lambda_2)\varepsilon_{t-1} + \lambda_1 \lambda_2 \varepsilon_{t-2}.$$

We observe that (2.48) is not only much more complex than (2.44), but is also nonlinear in the parameters. The unknown parameters λ_1 and λ_2 appear both in the relation and in the expression for the disturbance terms (u_t).

In a number of models (advertising) dynamics are modeled through the creation of an advertising goodwill variable.[16] This variable is a weighted average of advertising expenditure over time. Hence it is assumed that advertising expenditures create a (goodwill) stock variable which grows and declines ("depreciates") over time. Such a variable is also called a *stock variable*. The stock variable can be defined as:

$$mes_{it} = \rho_d mes_{i,t-1} + mef_{it} \tag{2.49}$$

where

mes_{it} = marketing expenditure stock of brand i at t,

mef_{it} = marketing expenditure flow of i at t, i.e. the marketing expenditures in t,

ρ_d = the discount rate.

The discount rate usually is determined by a grid search. In a study performed by Leeflang and Wieringa (2010) of the Dutch market of pharmaceuticals, a value of 0.55 was used.

[16] See, for example, Rizzo (1999); Dubé et al. (2005); Windmeijer et al. (2005) and Doganoglu and Klapper (2006).

2.8.4 *Lead Effects*

It is possible to extend marketing dynamics to include anticipations as well as carry-over effects. Leads occur when customers and/or competitors anticipate a marketing action and adjust their behavior before the action takes place. Consumers may expect prices to fall or a new product to be introduced. In these cases they may hold back purchases until the anticipated event occurs.

Consumers may also decelerate their purchases in anticipation of a promotion. This produces a pre-promotion dip. This theory has some support in the literature (Krishna 1992) and has been confirmed empirically in studies by Van Heerde et al. (2000) and Nies et al. (2014). Leads may also occur because sales persons delay sales if they anticipate or know that selling commissions are to rise (Doyle and Saunders 1985). Lead effects can be modeled in a similar manner as lagged effects, at least in principle. As an example we specify

$$q_t = \alpha + \beta_1 p_t + \beta_2 p_{t+1} + \beta_3 p_{t+2} + \varepsilon_t \tag{2.50}$$

where

$$\begin{aligned} p_t &= \text{price per unit in period } t, \text{and} \\ p_{t+1}, p_{t+2} &= \text{the announced prices or expected prices one- and two} \\ &\quad \text{periods ahead, respectively.} \end{aligned}$$

If consumers expect price increases in $t+1$ and $t+2$, $\beta_2, \beta_3 > 0$, i.e.: *positive* leads. Anticipation of price reductions result in *negative* leads, where the parameters are expected to have the same signs, i.e.: $\beta_2, \beta_3 > 0$. Negative leads were found by Doyle and Saunders (1985) and Van Heerde et al. (2000).

Lead and lag effects can also be combined: the two preceding subsections offer several ways to extend Eq. (2.50) so that it also accommodates lagged effects.

Van Heerde et al. (2000) applied three different distributed lead and lag structures to nine brands in the product categories: tuna fish and toilet tissue. Within each of three models they varied lead and lag lengths as well as the parameter describing the lag structure. They find that:

- Significant dynamic promotion effects exist and that these effects are substantial: between 4 and 25 %. In another study, Van Heerde et al. (2004) found that one third of the peak sales that result from a promotion are due to anticipation (leads) and stockpiling (lags) effects.
- Given the complexity of dynamic sales promotion effects, it is advisable to use a flexible specification such as the unrestricted model or the Almon model.

The selection of a dynamic model, in principle, can be based on theoretical criteria (size and development of lead/lagged effects over time) and statistical tests. We return to these tests in Chap. 5.

We do not specify a dynamic model for the Verhouten case here, but refer to Volume II of this book, where we will discuss several examples of dynamic model specifications.

References

Albers, S.: A framework for analysis of sources of profit contribution variance between actual and plan. Int. J. Res. Mark. **15**, 109–122 (1998)

Almon, S.: The distributed lag between capital appropriations and expenditures. Econometrica **33**, 178–196 (1965)

Assmus, G., Farley, J.U., Lehmann, D.R.: How advertising affects sales: Meta analysis of econometric results. J. Mark. Res. **21**, 65–74 (1984)

Baron, R.M., Kenny, D.A.: The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. **51**(6), 1173–1182 (1986)

Bass, F.M., Clarke, D.G.: Testing distributed lag models of advertising effect. J. Mark. Res. **9**, 298–308 (1972)

Bruce, N.I.: Pooling and dynamic forgetting effects in multitheme advertising: Tracking the advertising sales relationship with particle filters. Mark. Sci. **27**, 659–673 (2008)

Bucklin, R.E., Lattin, J.M.: A two-state model of purchase incidence and brand choice. Mark. Sci. **10**, 24–39 (1991)

Chintagunta, P.K.: Investigating purchase incidence, brand choice and purchase quantity decisions of households. Mark. Sci. **12**, 184–208 (1993)

Chintagunta, P.K.: Variety seeking, purchase timing and the 'lighting bolt' brand choice model. Manag. Sci. **45**, 486–498 (1999)

Clarke, D.G.: Econometric measurement of the duration of advertising effect on sales. J. Mark. Res. **13**, 345–357 (1976)

Doganoglu, T., Klapper, D.: Goodwill and dynamic advertising strategies. Quant. Mark. Econ. **4**, 5–29 (2006)

Doyle, P., Saunders, J.: The lead effect of marketing decisions. J. Mark. Res. **22**, 54–65 (1985)

Dubé, J.P., Hitsch, G.J., Manchanda, P.: An empirical model of marketing dynamics. Quant. Mark. Econ. **3**, 107–114 (2005)

Farris, P.W., Parry, M.E., Ailawadi, K.L.: Structural analysis of models with composite dependent variables. Mark. Sci. **11**, 76–94 (1992)

Gupta, S.: Impact of sales promotions on when, what and how much to buy. J. Mark. Res. **25**, 342–355 (1988)

Hanssens, D.M., Leeflang, P.S.H., Wittink, D.R.: Market response models and marketing practice. Appl. Stoch. Models Bus. Ind. **21**, 423–434 (2005)

Hanssens, D.M., Parsons, L.J., Schultz, R.L.: Market Response Models: Econometric and Time Series Analysis. Kluwer Academic, Boston (2001)

Hanssens, D.M., Pauwels, K., Srinivasan, S., Vanhuele, M., Yildirim, G.: Consumer attitude metrics for guiding marketing mix decisions. Mark. Sci. **33**, 534–550 (2014)

Horváth, C., Fok, D.: Moderating factors of immediate, gross and net cross-brand effects of price promotions. Mark. Sci. **32**, 127–152 (2013)

Irwin, J.R., McClelland, G.H.: Misleading heuristics for moderated multiple regression models. J. Mark. Res. **38**, 100–109 (2001)

Judge, G.G., W. E. Griffiths R. C. Hill, H.L., Lee, T.C.: The Theory and Practice of Econometrics, 2nd edn. Wiley, New York (1985)

Kadiyali, V., Vilcassim, N.J., Chintagunta, P.K.: Product line extensions and competive market interactions: An empirical analysis. J. Econom. **89**, 339–363 (1999)

Kalyanam, K., Shively, T.S.: Estimating irregular pricing effects: A stochastic spline regression approach. J. Mark. Res. **35**, 16–29 (1998)
Kanetkar, V., Weinberg, C.B., Weiss, D.L.: Estimating parameters of the autocorrelated current effects model from temporally aggregated data. J. Mark. Res. **23**, 379–386 (1986)
Kaul, A., Wittink, D.R.: Empirical generalizations about the impact of advertising on price sensitivity and price. Mark. Sci. **14**, G151–G160 (1995)
Koschate-Fischer, N., Cramer, J., Hoyer, W.D.: Moderating effects of the relationship between private label share and store loyalty. J. Mark. **78**(2), 69–82 (2014)
Koyck, L.M.: Distributed Lags and Investment Analysis. North-Holland, Amsterdam (1954)
Krishna, A.: The normative impact of consumer price expectations for multiple brands on consumer purchase behavior. Mark. Sci. **11**, 266–286 (1992)
Krishnamurthi, L., Raj, S.P.: A model of brand choice and purchase quantity price sensitivities. Mark. Sci. **7**, 1–20 (1988)
Lam, S.Y., Vandenbosch, M., Hulland, J., Pearce, M.: Evaluating promotions in shopping environments: Decomposing sales response into attraction, conversion, and spending effects. Mark. Sci. **20**, 194–215 (2001)
Larréché, J.C.: Marketing managers and models: A search for a better match. Unpublished Ph.D. thesis, Stanford University (1974)
Leeflang, P.S.H.: Modeling competitive reaction effects. Schmalenbach Bus. Rev. **60**, 322–358 (2008)
Leeflang, P.S.H., Bijmolt, T.H.A., Van Doorn, J., Hanssens, D.M., Van Heerde, H.J., Verhoef, P.C., Wieringa, J.E.: Creating lift versus building the base: Current trends in marketing dynamics. Int. J. Res. Mark. **26**, 13–20 (2009)
Leeflang, P.S.H., Mijatovic, G.M., Saunders, J.: Identification and estimation of complex multivariate lag structures: A nesting approach. Appl. Econ. **24**, 273–283 (1992)
Leeflang, P.S.H., Reuyl, J.C.: On the application of generalized least squares methods to logical consistent market share models. Onderzoeksmemorandum nr. 56, Institute of Economic Research, Faculty of Economics, University of Groningen, Groningen (1979)
Leeflang, P.S.H., Reuyl, J.C.: Advertising and industry sales: An empirical study of the West German cigarette market. J. Mark. **49**(4), 92–98 (1985)
Leeflang, P.S.H., Wedel, M.: Information based decision making in pricing. In: Proceedings ESOMAR/EMAC/AFM Symposium on Information Based Decision Making in Marketing, Paris (1993)
Leeflang, P.S.H., Wieringa, J.E.: Modeling the effects of pharmaceutical marketing. Mark. Lett. **21**, 121–133 (2010)
Leone, R.P.: Generalizing what is known about temporal aggregation and advertising carryover. Mark. Sci. **14**, G141–G150 (1995)
Lilien, G.L.: Bridging the academic - practitioner divide in marketing decision models. J. Mark. **75**(4), 196–210 (2011)
Lilien, G.L., Kotler, P.: Marketing Decision Making: A Model-Building Approach. Harper and Row, New York (1983)
Lilien, G.L., Rangaswamy, A.: Marketing Engineering, 2nd rev. edn. Trafford Publishing, Bloomington (2004)
Lilien, G.L., Rangaswamy, A., De Bruyn, A.: Principles of Marketing Engineering, 2nd edn. DecisionPro, Inc. (2013)
Little, J.D.C.: Models and managers: The concept of a decision calculus. Manag. Sci. **16**, B466–B485 (1970)
Little, J.D.C.: BRANDAID: A marketing-mix model, part 1: Structure. Oper. Res. **23**, 628–655 (1975a)
Little, J.D.C.: BRANDAID: A marketing-mix model, part 2: Implementation, calibration, and case study. Oper. Res. **23**, 656–673 (1975b)
Lodish, L.M., Abraham, M.M., Livelsberger, J., Lubetkin, B., Richardson, B., Stevens, M.E.: A summary of fifty-five in-market experimental estimates of the long-term effect of T.V. advertising. Mark. Sci. **14**, G133–G140. (1995)

Mahajan, V., Green, P.E., Goldberg, S.M.: A conjoint model for measuring self- and cross-price demand relationships. J. Mark. Res. **19**, 334–342 (1982)
Montgomery, D.B., Silk, A.J.: Estimating dynamic effects of market communications expenditures. Manag. Sci. **18**, B485–B501 (1972)
Naert, P.A., Bultez, A.V.: Logically consistent market share models. J. Mark. Res. **10**, 334–340 (1973)
Nies, S., Leeflang, P.S.H., Bijmolt, T.H.A., Natter, M.: Multi-unit price promotions and their impact on sales. Working paper. Faculty of Economics and Business, Department of Marketing, University of Groningen, Groningen (2014)
Pauwels, K., Currim, I., Dekimpe, M.G., Ghysels, E., Hanssens, D.M., Mizik, N., Naik, P.: Modeling marketing dynamics by time series econometrics. Mark. Lett. **15**, 167–183 (2004)
Pauwels, K., Erguncu, S., Yildirim, G.: Winning hearts, minds and sales: How marketing communication enters the purchase process in emerging and mature markets. Int. J. Res. Mark. **30**, 57–68 (2013)
Pauwels, K., Leeflang, P.S.H., Teerling, M.L., Huizingh, K.R.E.: Does online information drive offline revenues? Only for specific products and consumer segments! J. Retail. **87**, 1–17 (2011)
Pauwels, K., Weiss, A.: Moving from free to fee: How online firms market to change their business model successfully. J. Mark. **72**(3), 14–31 (2008)
Rizzo, J.A.: Advertising and competition in the ethical pharmaceutical industry: The case of antihypertensive drugs. J. Law Econ. **42**, 89–116 (1999)
Sethuraman, R., Tellis, G.J., Briesch, R.: How well does advertising work? Generalizations from a meta-analysis of brand advertising elasticity. J. Mark. Res. **48**, 457–471 (2011)
Srinivasan, S., Vanhuele, M., Pauwels, K.: Mind-set metrics in market response models: An integrative approach. J. Mark. Res. **27**, 672–684 (2010)
Theil, H.: Principles of Econometrics. Wiley, New York (1971)
Urban, G.L., Karash, R.: Evolutionary model building. J. Mark. Res. **8**, 62–66 (1971)
Vakratsas, D., Feinberg, F.M., Bass, F.M., Kalyanaram, G.: The shape of advertising response function revisited: A model of dynamic probabilistic thresholds. Mark. Sci. **23**, 109–119 (2004)
Van Heerde, H.J., Bijmolt, T.H.A.: Decomposing the promotional revenue bump for loyalty program members versus nonmembers. J. Mark. Res. **42**, 443–457 (2005)
Van Heerde, H.J., Gupta, S., Wittink, D.R.: Is 75% of the sales promotion bump due to brand switching? No, only 33% is. J. Mark. Res. **40**, 481–491 (2003)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: The estimation of pre- and postpromotion dips with store-level scanner data. J. Mark. Res. **37**, 383–395 (2000)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: How promotions work: SCAN*PRO-based evolutionary model building. Schmalenbach Bus. Rev. **54**, 198–220 (2002)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Decomposing the sales promotion bump with store data. Mark. Sci. **23**, 317–334 (2004)
Van Nierop, J.E.M., Leeflang, P.S.H., Teerling, M.L., Huizingh, K.R.E.: The impact of the introduction and use of an informational website on offline customer buying behavior. Int. J. Res. Mark. **28**, 155–165 (2011)
Vriens, M.: The Insights Advantage: Knowing How to Win. iUniverse, Bloomington (2012)
Wedel, M., Leeflang, P.S.H.: A model for the effects of psychological pricing in Gabor-Granger price studies. J. Econ. Psychol. **19**, 237–260 (1998)
Weiss, P.L., Windal, P.M.: Testing cumulative advertising effects: A comment on methodology. J. Mark. Res. **17**, 371–378 (1980)
Wierenga, B., Van Bruggen, G.H.: Marketing Management Support Systems: Principles, Tools, and Implementation. Kluwer Academic, Boston (2000)
Wierenga, B., Van Bruggen, G.H., Staelin, R.: The success of marketing management support systems. Mark. Sci. **18**, 196–207 (1999)
Wiesel, T., Pauwels, K., Arts, J.: Marketing's profit impact: Quantifying online and off-line funnel progression. Mark. Sci. **30**, 604–611 (2011)
Windmeijer, F., de Laat, E., Douven, R., Mot, E.: Pharmaceutical promotion and GP prescription behaviour. Health Econ. **15**, 5–18 (2005)

Chapter 3
Data

3.1 Introduction

Decision making in marketing must be based on appropriate, high quality data. Revolutionary developments in data collection during the last few decades offer many opportunities for advanced model building and the application of advanced research methods. For example, with the scanning revolution, the Internet invasion (Little 2004) and the "Big Data" era (see Pauwels 2014), we observed exponential increases in the availability of data.

Ideally, model development should precede data collection (Sect. 1.4). However, the *availability* of "good" data is a prerequisite for meaningful, and hence implementable, model building. For this reason data considerations affects model development in many practical cases. In this chapter we focus on various data issues before we turn to specification.

In most cases marketing managers have access to an abundance of data from different sources such as scanner-based data, Internet data and individual customer data. Not all sources are equally structured and (directly) suitable for modeling purposes. In Sect. 3.2 we discuss the most important data structures, and in Sect. 3.3 we explain what we mean by "good" data.

The model building step that follows specification is estimation. An important estimation issue is the selection of an appropriate estimation technique. In Sect. 3.4 we discuss how the choice for a certain estimation methodology depends on data characteristics.

In Sect. 3.5, we discuss well-known data sources, including scanner data, for the measurement of performance indicators and variables that influence performance. We also spend some attention to "Big Data" and subjective data.

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_3

3.2 Data Structures

Business data sets come in a variety of types. The most important data structures are:

- cross-sectional data;
- time series data;
- pooled cross-sectional data, and
- panel data.

A *cross-sectional* data set consists of a sample of customers, firms, brands, regions or other units taken at a single point in time. A common assumption in analyzing cross-sectional data is that they were obtained by random sampling from the underlying population. As a consequence, cross-sectional data typically do not have a natural ordering. In many practical cases however, this assumption is violated and researchers are then confronted with a sample selection problem.

Time series data sets consist of observations on a variable or several variables over time that are measured for a single unit. One feature that distinguishes time series data from cross-sectional data is that observations of a time series have a natural, temporal, ordering. In many cases time series observations are related to earlier observations and are not independent across time i.e., they exhibit serial correlation.

Data sets may have both cross-sectional and time series features. For example, a researcher has access on cross-sectional data that refer to January 2013 and January 2014. A *pooled cross section* is formed by combining the observations of these months, which increases the sample size.

A *panel* data (or longitudinal data) set consists of a time series of *each* cross-sectional member in the data set (Wooldridge 2012, p. 10). The key feature of panel data that distinguishes them from pooled cross-sectional data is that the *same* cross-sectional units are followed over a given time period.

Depending on the data structure, the marketing model builder faces different modeling challenges, such as the sample selection problem in cross-sectional models, the presence of serial correlation in time series analysis (Sect. 5.2.3) and the treatment of heterogeneity in units in (pooled) cross-sectional and panel data (Sect. 2.7, Sect. 4.5 and Chap. 8).

3.3 "Good Data"

Even for a very experienced and highly skilled model builder, the quality of the insights that are generated by an empirical modeling exercise depends strongly on the data that is used to calibrate the model. Data sets should contain good data. "Good" data encompasses *availability, quality, variability* and *quantity* (see also Vriens 2012, pp. 28–29).

3.3.1 Availability

The first requirement for "good" data is their availability. Most companies will have an internal accounting department that should be able to deliver data on a firm's own actions and performance. Thus, in principle, it should be relatively easy to obtain data on variables such as unit sales, revenues, prices of the firm's products, advertising expenditures, wholesale and retail margins, promotions and personal selling expenditures, and market coverage. The availability of such data does not imply that the data are directly usable. We return to this issue when we discuss data quality.

Other data might be more difficult to construct or to extract from company data bases. Moreover, many marketing models need to incorporate information from other firms, such as marketing actions for competitive products. For these reasons it is common for managers to purchase data, gathered regularly or incidentally, from market research firms. IRI, GfK and ACNielsen provide market feedback reports on a large number of product categories gathered through scanner equipment in supermarkets, drugstores, etc. Even then the available data may be incomplete. Incompleteness in the form of missing predictors is a problem because the estimated effects of the available variables will be biased (if there is covariation between included and excluded predictor variables).

Depending on the precise purpose of the model-building effort it may be useful to gather additional data. This leads to the question of how to properly combine data from different data sources. For example, a researcher might want to combine household-level scanner-based sales data with household-level media coverage data. These data are sometimes available from a single source, but quite often different data types are collected by different market research agencies. Initially, ad hoc procedures were used to match households from different data bases on the basis of similarity in (often demographic) characteristics. A similar problem arises when a researcher wants to compare the effect of classical (mass) media and social media.[1] Nowadays, data-fusion techniques are available that, based on multiple imputations of the missing data, yield fused data sets with known statistical properties (Kamakura and Wedel 1997; McDonnell Feit et al. 2013; Qian and Xie 2014).

Generally speaking, data on industrial markets are harder to obtain than consumer data. Systematic records are rarely kept by individual firms. Although systematic data gathering by commercial market research agencies is growing, this is typically limited to larger markets (such as electronic components). Customization in industrial markets complicates the problem: products, prices and services are often customer specific and usually not publicly available, and the actions of competitors are often unknown (Brand and Leeflang 1994).

[1] See, for example, Stephen and Galak (2012); Pauwels et al. (2013); De Vries et al. (2014).

More recently, we observe that researchers study dyadic relationships between firms (in industrial markets). Examples are studies by Wuyts et al. (2004), Wuyts (2007), and Van den Bulte and Wuyts (2007).

3.3.2 Quality

The quality of a data set is determined by the *validity* and the *reliability* of its data measures. A measure is valid if it measures what it is supposed to measure. Even if a measure is *valid* it will in many cases not be possible to measure it without error. The degree to which a measure is subject to random error is assessed by its *reliability*. Errors in the measurement of predictor variables can cause their estimated effects to be biased. It is in general quite difficult to define reliable and valid measures for variables such as "product quality" and "the value of a brand" (brand equity).[2] Much effort has been put into the development of appropriate measurement scales. Handbooks of validated scales are available that inform market research practice (Bearden et al. 2011). The validity of data used to measure the effectiveness of such variables on sales or profit is, generally speaking, not high. But the validity of directly observable data available from, say, the accounting department, may also be low. Furthermore, data obtained from panels and surveys are subject to biases and sampling error.

3.3.3 Variability

If a variable shows no variation we cannot measure its impact on the criterion variable. Generally speaking, the effect of a predictor variable will be estimated more precisely when the amount of sample variation is larger, where "precision" is inversely related to the "estimated standard error" or (statistical) unreliability of the effect estimate.

For models with multiple predictor variables, the "goodness" of the data also depends on the amount of *covariation* between the predictors. The precision of the estimated effect of a predictor variable usually decreases with the amount of covariation with other predictors.

The variability of marketing mix variables may be less than what is needed for the estimation of their effects. For example, due to price competition the price range observed in a market may be low, or the range of variation in a product attribute such as package size is limited. In such cases the revealed preference data (choices) in the market place may be supplemented with other data. Experiments may be used in which desired variation in marketing mix elements is induced. A powerful technique

[2]See for an overview of brand equity measurement Farris et al. (2006).

is conjoint choice experimentation in which product profiles are experimentally designed, and choices among them are made by respondents. Models are available to integrate such stated preference data with revealed preference data.

Sometimes there are multiple *sources of variation* that can be used to infer relations. For example, one may have access to time series data covering 52 weeks on 60 stores. In such a case it is suggested to use the time series and cross-sectional variation simultaneously and use parameterization methods that are appropriate for panel data.

3.3.4 Quantity

The final requirement for "good" data is quantity. If a probability sample of observations is used to estimate a population mean, the sample size influences the precision of the estimate. However, if the interest focuses on the relation between variables, it is the amount of variation in a predictor (as well as covariation between predictors) that influences the precision of estimated effects.

The quantity of observations is, however, a critical factor for the joint estimation of all model parameters. At the very bare minimum, the number of observations should exceed the number of model parameters. This is a necessary condition that needs to be met, otherwise it is impossible to estimate all parameters. Many researchers suggest that the quantity (number of observations) should be larger than five times the number of parameters. However, this is only a general rule of thumb, and the required number of observations to obtain a given precision depends, as we argued above, on the amount of variation in and the amount of covariation between the predictor variables.

Another concern regarding the number of observations is that it is useful to reserve some data for validation purposes (see Sect. 5.7). Model building is an iterative process, and if one data set is used to estimate several alternative specifications before one specification is selected, then the usual statistical criteria are no longer valid. This type of estimation is often referred to as pretest estimation (Leamer 1978). If the purpose of the model building is to test theories, the final specification should be validated on one or more new data sets. These other data sets could represent different product categories, different regions, different time periods, etc. If the model results are intended to be the basis for marketing decisions, an appropriate test is whether the model predictions, conditional upon specific marketing actions, outperform predictions from a manager (testing the model against managerial judgment). For more on validation, see Chap. 5.

If the initial quantity of data is inadequate for model estimation, the model builder can consider constrained estimation methods. For example, suppose one wants to estimate the demand for one brand as a function of six marketing variables. If there are nine other brands that belong to the same product category, and it is desirable to have cross-brand effects for all brands and all variables, then the number of marketing variable parameters is 60. Obviously 1 year of weekly data is insufficient

if the data are limited to time-series observations for one cross section. One solution to the problem is to obtain time-series data for other cross sections. In that case the issue of pooling cross-sectional and time-series variation has to be considered (see Sects. 4.5 and 5.4). An alternative solution is to employ constrained estimation. For example, one could force some of the other brands to have zero cross effects by eliminating those brands' variables altogether. Or one could combine some brands and force these parameters to be the same for brands belonging to common subsets.

3.4 Data Characteristics and Model Choice

In this section we focus on the characteristics of the variable that the model aims to explain: the dependent variable. In many cases, this is a performance measure such as sales, purchase intention or the brands that were bought by a certain set of customers. The scale of the dependent variable determines which models and estimation techniques are appropriate for analyzing these variables. Table 3.1, which is taken from Franses and Paap (2001), presents an overview.

An example of a typical *continuous* dependent variable is sales, measured in dollars or in Euros. In many cases sales are measured in units and the corresponding variable is then strictly speaking not a continuous variable, but a *count variable*. The regression model is typically also used in such situations because the error of applying a model for continuous data to count data is fairly small, provided the numbers are reasonably large. In cases where sales counts are small, as can be the case in B2B marketing situations, the researcher should use other models (such as Poisson models; see Sect. 8.3).

An example of a *binomial* dependent variable is the situation where we are modeling consumers' decisions to purchase a product or not. If we consider the choice between more than two brands, such as A, B and C, we encounter an *unordered multinomial variable*. If we only consider two brands, A and B, the dependent variable takes only two values and is again a *binomial variable*. If we measure the quality of a brand on a so-called Likert scale which runs from 1 to 7, the corresponding (dependent) variable is an *ordered multinomial variable*.

Censoring and truncation occur when not all values of a variable are recorded. An example of a *censored variable* is the situation when we are measuring relationship duration for a cohort of customers that signed a contract with a company 10 years ago. For the customers that terminated their contract a relationship duration is recorded. For those customers where the relation is still ongoing, we do not know the precise relationship duration, but we know is at least 10 years: this variable is (right) censored at 10. An example of a *truncated variable* is the yearly amount of miles traveled by customers of an airline company that have attained the platinum status in their frequent flyer program, which requires 80,000 miles. This variable is said to be (left) truncated at a value of 80,000. The difference between censoring and truncation is that with censoring we know how many cases are outside the threshold(s), whereas this is not the case with truncated data.

An example of a *duration variable* is the time that elapses between two purchases: the interpurchase time.

We discuss models for continuous dependent variables in Chaps. 4–7 and in Chap. 9. Chapter 8 discusses models where the dependent variable is binomial or multinomial. Models for truncated or censored data and models for duration data are also briefly discussed in Chap. 8

Table 3.1 Types of dependent variables and appropriate models/estimation techniques

Dependent variable	Name of model
Continuous	Standard Linear Regression model
Binomial	Binomial Logit/Probit model
Unordered multinomial	Multinomial Logit/Probit model
	Conditional Logit/Probit model
	Nested Logit model
Ordered multinomial	Ordered Logit/Probit model
Truncated/censored	Truncated Regression model
	Censored Regression (Tobit) model
Duration	Proportional Hazard model
	Accelerated Lifetime model

Source: Franses and Paap (2001, p. 27)

3.5 Data Sources

3.5.1 Introduction

Decision making in marketing must be based on profound data. Revolutionary developments in data collection (see Table 3.2) offer many opportunities for advanced model building and the application of advanced research methods. For example, the scanning revolution and Internet invasion (Little 2004) prompted exponential increases in the availability of data. Advances in data collection and store technologies have given rise to the customer data intermediary (CDI), a firm that collects customer data to offer customer-specific marketing services to marketers; see Pancras and Sudhir (2007). Access to and use of Internet data, social media, and data from customer relationship management (CRM) systems has multiplied this increase exponentially.

Before we discuss the most important data sources in more detail,[3] we spent some attention to the metrics/data that are used in marketing practice. Bendle et al. (2010) demonstrated that financial metrics are widely regarded as the most useful; of the

[3]For a more extensive discussion see e.g. Malhotra (2010, Chapter 4) and the surveys by the American Marketing Association which are known as the Honomichl Global Top 25.

Table 3.2 Data availability

1950:

- Store-level data (bimonthly ACNielsen data)
- (Relatively small and non-representative) samples of consumer data
- More representative and larger samples (Attwood statistics, GfK) (Leeflang and Olivier 1985)
- Ad hoc surveys (cross-sectional and time-series data)

1985:

- The scanning revolution (Bucklin and Gupta 1999)
  - Consumer panel data
  - Store-level data
  - Cross-sectional and time-series data (panel data)
  - Daily data

1995:

- Internet revolution
  - Internet data (special issue *Marketing Science*, vol. 19, no. 1)
  - Online publications and offline purchases, combined with website behavior (Pauwels et al. 2011)
  - Search engines (Telang et al. 2004)
  - Recommendation systems (Ansari et al. 2000)
  - Auctions (Yao and Mela 2008)
  - Web-based marketing research (Bucklin and Sismeiro 2009)

2000:

- Databases constructed by individual firms (CRM systems) (Blattberg et al. 2008)

2008:

- Data from social media (e.g. Facebook, LinkedIn, Twitter, Weblogs; Van Laer and De Ruyter 2010)

Source: Leeflang (2011, p. 81)

metrics that are usually considered marketing metrics, only customer satisfaction (71 %) and loyalty (69 %) make the top ten list, according to senior managers. In addition, Verhoef et al. (2009) study the data and metrics stored in the databases of 183 Dutch firms and find that many firms collect data systematically over time, a finding that appears clear in comparison with the metrics collected in a previous survey (Verhoef et al. 2002), as Table 3.3 illustrates. Yet, data collection does not necessitate that the variables are related formally in marketing decision models. Verhoef et al. (2009) conclude that only about 20 % of all firms perform statistical analyses using the data they collect. However, the trends in the types of analyses in Table 3.4 imply that advanced techniques have gained in importance over time.

3.5.2 Classification

A well-known classification of data distinguishes primary and secondary data. *Primary data* are collected by a researcher for the sole purpose of addressing a specific problem at hand. Secondary data are data that were gathered for purposes other than the problem at hand (Malhotra 2010, p. 132). The models that we discuss in this monograph usually utilize *secondary data*.

Table 3.3 Data availability: data stored in customer databases (percentage of firms)

	2003	2008
Type of product purchased	68	81
Demographics	34	56
Lifestyle data	17	40
Number of offers (outbound actions)	62	72
Share of wallet	7	34
Interaction information	42	76
Customer satisfaction data	12	60

Sources: Based on Verhoef et al. (2002, 2009)

Table 3.4 Use of statistical techniques for segmentation and forecasting (percentage of firms)

	2003	2008
Genetic algorithms	3	35
Neural networks	5	44
Factor analyses	19	56
Cluster analyses	32	67
Discriminant analyses	13	43
Logit/probit analyses	6	44
Linear regression analyses	33	60
CHAIN/CART	17	54
Cross tabulations	54	65
RFM analyses	42	52

Sources: Based on Verhoef et al. (2002, 2009)

Secondary data typically are easily accessible, relatively inexpensive, and quickly obtained. Secondary data may be further classified into external and internal data. *Internal data* are generated within the organization for which a model is developed; for example, transactional data that are stored in a company's CRM system. *External data* are generated by sources outside the organization. Examples are data provided by a country's national bureau of statistics and syndicated sources of data. The latter are provided by (market/marketing) research companies that collect and sell common pools of data. These companies primarily collect revealed preference data. Revealed preference data reflect choices and are measured in

terms of sales and market shares. Sales and market shares and a (sub)set of causal variables can be measured at, at least, three levels in the marketing channel, viz:

- at the household level, for example, through a household panel whose purchases are electronically recorded;
- at the retail level by means of a store audit or electronic (scanner-based) registration of activities;
- at the manufacturer level.

We discuss internal data in Sect. 3.5.3. Then we turn to external data, where we concentrate on data from syndicated services (Sect. 3.5.4). We discuss the advantages and disadvantages of household panel data and store-level data in Sect. 3.5.5. Finally, we spend attention to "Big Data" (Sect. 3.5.6) and subjective data (Sect. 3.5.7)

3.5.3 Internal Data

Companies store an abundance of data that are recorded for various purposes. For example, a company's call center may record the number complaints and qualitative information about the nature of the complaints as well as the follow-up action. The financial department may store data about prices, margins and paying behavior of the customers. The marketing department may store outcomes of market research studies and details of multi-media promotional campaigns. The sales department may store the purchasing history of individual customers as well as the amount of discount each customer received.

In the ideal situation, all these data are readily available to a model builder and are easily and real-time accessible via a data warehouse.

Some companies indeed have organized their data management process very well and are able to extract up to several hundreds of variables about the business at any desired aggregation level very easily. However, in many other situations, this is not the case, and it typically takes a lot of effort to (1) gain access to all different data sources, (2) link data from different sources together, and (3) clean the data and put it in such a format that it is usable for analysis.

Most companies gather (a subset of) the following types of internal variables:

1. Information on key performance indicators, such as:
   - sales (*M*oney/units);
   - *R*ecency (time since most recent visit/purchase);
   - *F*requency (number of visits/purchases)
   - loyalty/satisfaction measures (e.g. NPS scores);
   - conversion rates;
   - cross-selling rates;
   - Customer Lifetime Value (CLV).[4]

[4]See also Sect. 10.5.4 (Tables 10.4 and 10.5).

2. Information about possible explanatory variables such as:

   - number and/or quality of inbound and outbound contacts;
   - prices paid;
   - discounts given;
   - promotions received;
   - use of loyalty cards;
   - mailings sent, etc.

3. Information about customer characteristics:

   - demographics
   - interests (lifestyle data)
   - activities (memberships etc.)

These databases are in many cases complemented with external data (Kotler and Armstrong, 2012, p. 123). For example, many companies buy customer lifestyle data from syndicated data sources to further complete their customer view.

In the past few years many models have been developed in the area of database marketing that deal with customer acquisition, customer retention, and customer development. We discuss these models in Chap. 9. For a survey we refer to Bijmolt et al. (2010).

3.5.4 External Data

Traditionally, and in some countries this is still the default approach, information from households showing repeat purchase and brand switching behavior is obtained through diary panels. Such a panel consists of families who use a preprinted diary in which they record their weekly purchases in specified product categories. Typically, the item (brand name, type of product, weight or quantity, kind of package), number of units and price along with store information are reported for each purchase. The families are geographically dispersed, while the panel is demographically balanced so the data can be projected to the national level in a given country. Families are recruited so that the composition of the panel mirrors the population as much as possible on specified characteristics. Panel members are compensated for their participation, often with gifts. Families are dropped from the panel at their request, if they fail to return their diaries, or if their records prove to be unreliable. The diaries are returned weekly. Clients of diary panel services use the data to assess among other things: the size of the market for a product, the proportion of families buying over time, the amount purchased per household, brand share over time, the frequency of purchase and amount purchased per transaction, average price paid, etc. In some panels, the members are asked to record whether purchases are on promotion.

3.5.4.1 Store (Retail) Level

For many decades, the ACNielsen dominated the industry of systematic data gathering at the retail level for the purpose of tracking brand performance. The bimonthly audit data, still available in many countries all over the world, are based on national probability samples. Auditors visit the same stores every 2 months. During the visit they take a complete inventory of all items covered (for the product categories for which manufacturers are clients), and they record all invoices since the previous visit. Total purchases by the retailer plus the reduction in inventory provide the revenue (based on the shelf price prevailing at the time of the audit) and unit sales information for each item per store. Other information collected includes out-of-stock conditions (at the time of the audit), and certain promotional activities such as premiums, bonus packs, sampling and featuring. A probability sample of store data, weighted by certain store characteristics and previous period results, produces highly accurate estimates of brand performance at the national level. However, as the coverage of purchases by electronic means increases in a country, the bimonthly audit service disappears.

Another example of a retail panel is Info Scan Retail Tracking. Info Scan Retail Tracking, a service from IRI, provides manufacturers and retailers with access to detailed information on sales, market share, distribution, pricing, and provides access to a wide variety of retail channels and accounts. For other product categories not covered by ACNielsen, Audits and Surveys' National Total Market Audit provides data in some countries.

3.5.4.2 Manufacturer Level

The manufacturer's *internal accounting* system can provide "shipped sales" or "ex-factory sales". To measure household purchases, these data have to be corrected for changes in inventory at the wholesale and retail levels. Inventory changes at the retail level are obtained through store audits; inventory changes at the wholesale level are obtained from sales representatives who take on the role of intelligence gatherers or from wholesale audits. These corrections of the "ex-factory" sales are performed a couple of times per year. The ex-factory sales of a manufacturer's own brand can be combined with the corresponding figures of other brands in a given product category available from an independent institute. In that way, estimates of total industry sales, industry sales per segment and brands' market shares are obtained.

3.5.4.3 Evaluation

The precision of the data from each of these sources is not guaranteed. Yet, if household survey- or store audit data are used for marketing decisions, a certain

rigor must be imposed. Problems in this realm have been identified by, for example, Shoemaker and Pringle (1980); Leeflang and Olivier (1985) and Plat (1988).

Leeflang and Olivier observed substantial differences for measures of sales, market share and price between the different levels in the marketing channel that are distinguished in Sect. 3.5.2. Nonresponse bias is a major reason for problems with *sample survey data*. For example, households who buy relatively inexpensive brands have a higher response rate than households who buy the more expensive brands. The nonresponse bias in *store audit data* results from the refusal of some retailers such as discounters, to cooperate. Leeflang and Olivier (1985) demonstrated that the data bias leads to large differences in marketing decisions between consumer panel data and store audit data.

3.5.4.4 Scanner Data

The availability of data for parameterization of marketing models has increased during the past decade through adoption of scanners by retailers.[5] In the early 1970s laser technology in conjunction with small computers first enabled retailers in the US to record electronically or "scan" the purchases made in their stores. Since then, after a period of slow growth, the adoption of scanning has increased rapidly, and scanner data have become available for decision support in marketing at many organizations. Although scanning was originally used by retailers simply as a labor- and cost-saving device, the computerized accumulation of point-of-sale information puts a library of accurate and detailed purchase records at the disposal of many marketing researchers and marketing managers.

Several years after the first scanners were installed in the US, scanning was introduced in Europe. To coordinate the encoding of products and the exchange of data, several organizations in the European countries established clearing institutes to amalgamate the data of affiliated retailers. In addition, systems for automated transmission of transaction data were developed, and the possibilities for natural electronic funds transfer systems (EFTS) were studied.

Although the original intent for scanner data collected by market research firms was to aid brand managers, retailers increasingly recognize opportunities too. For example, scanner data analysis can be the basis for space allocation of items in a product category to maximize store profits. To be able to do this for as many stores equipped with scanners as possible, both ACNielsen and IRI have expanded the number of stores from which they obtain (purchase) data. And since the sales teams of the manufacturers want to maximize the space for the brands they have responsibility for, the manufacturers are also interested in having access to data on all scanner-based stores. Thus, the market research firms now favor having census of all scanner-equipped retailers. At the same time, manufacturers and retailers have an increasing need to obtain solutions that maximize their joint profits.

[5] For a more extensive discussion see Bucklin and Gupta (1999).

In *scanner panels*, each household may use an ID-card similar to a credit card. Panel members are asked to present their card at the checkout counter each time they shop. This allows each panel member's ID-number to be recorded next to the set of items purchased by the member. Such data from scanner panels are supplied by IRI to clients through its Infoscan Service. Behavior Scan (from IRI) is a household scanner panel designed to support controlled tests of new products or advertising campaigns. An alternative to the ID card is to ask households to scan at home or at the store and to identify the store in which the purchases were made. This is done by ACNielsen and GfK (Consumerscan and Microscan) both of which supply each participating household with a wand.

The data from scanning-based store samples are known as "volume tracking data". The data provide information on purchases by brand, size, flavor or formulation (Stock-Keeping Unit-level: SKU) and are based on sales data collected from the checkout scanner tapes. Volume tracking data are supplied through Infoscan (IRI), Scantrack (ACNielsen), Nabscan (The Newspaper Advertising Bureau) and TRIM (Tele-Research, Inc.). The following measures are reported:

- volumes (at the SKU-level);
- revenues;
- actual prices;
- ACV = the All Commodity Volume of the store or store revenue;
- ACV Selling = the ACV for an item or group of items reported only in stores with some movement of that item;
- baseline sales: an estimate of unit sales under non-promoted conditions.

Also, regular prices are estimated to make a distinction between those prices and promotional prices which reflect temporary discounts.

These descriptions indicate how data on relevant decision variables are captured. In addition the promotional environment in a store is measured through the separate collection of information on displays and features. Merchandising is a generic name for promotional activity conducted by a store to increase sales (Little 1998). The market research companies report four mutually exclusive types of "merchandising": (1) display only, (2) feature only, (3) display and feature together, and (4) unsupported price cuts. Several of these can be subdivided further, if desired, for example by type of display. Most of these non-price promotional variables are collected as zero-one measures. Measures of merchandising activity can also be defined analogously to those for distribution.

A more advanced system consists of the *combination* of scanner panel data with cable TV advertising exposure records. If all the information in a data set stems from one source, such a data set is referred to as single-source data. Single-source data provide integrated information on household purchases, media exposure, and other characteristics, along with information on marketing variables such as price, promotion and in-store marketing efforts.

3.5.4.5 HandScan Panels

A key issue with retail panel data is that it requires the explicit cooperation from the retailer, which is not always guaranteed. Thus, panel retail data misses a growing portion of the grocery retail market. Also, it misses purchases from non-grocery retailers such as drug stores and gasoline stations. To address these issues, market research companies such as GfK created consumer hand scan panels. Participating consumers scan all receipts with a hand scanner, allowing researchers to uncover purchases and prices at all visited retailers, drug stores and gasoline stations.

As an example of how such data are used for modeling, Van Heerde et al. (2008) study how the Dutch price war changed consumer shopping patterns and retailer profits. Their GfK Benelux panel consists of 4,400 households, which represents a stratified national sample. The panel also provides household perceptions of retailers, such as their price image and product quality.

3.5.4.6 Causal Data

Price-, promotion-, and distribution data are natural components of the data collection methods just discussed; however, data on manufacturer advertising and product quality are not. There are, however, agencies that specialize in the collection of *advertising* expenditures, such as Leading National Advertisers (LNA) in the US. In many models predictor variables such as gross rating points (GRP's) on TV-ratings are used. These data are collected, for example, by Nielsen (Nielsen Media Research[6]).

In most models discussed in this monograph, the fundamental unit of analysis is the brand. Given the wide range of assortments offered in many product categories, the brand is a product line comprising several Stock-Keeping Units (SKU's). Most SKU's can be described in terms of a few physical characteristics to distinguish the items. Examples are SKU-attributes such as brand name, package size, "formula", flavor, etc. Marketing research firms use several criteria[7] to determine what can be treated as an SKU-attribute: each attribute must be recognizable by consumers in an objective manner (i.e. physically distinguishable), the variation in each attribute across SKU's must be discrete, and each attribute must be applicable to every SKU.

3.5.4.7 Other Data Inputs

Other data inputs refer to environmental variables such as size and age distribution of the population, temperature, and macroeconomic variables such as gross national

[6] See, for applications, Bass et al. (2007); Chessa and Murre (2007).

[7] We closely follow Fader and Hardie (1996).

product, per capita income and employment.[8] One can also use stated preferences and other subjective consumer judgments.

Measures of sales, "merchandising", distribution, etc. can be defined at different levels of aggregation with respect to (groups of) stores, regions, time periods, and products. Little (1998) emphasizes that to be most useful, these measures should have parallel and consistent meanings across the different levels. He also suggests a class of integrated measures that start with information routinely provided by data suppliers. The information is first decomposed and then aggregated analytically to refer to store groups, product lines, and multi-week periods.

Other external data refer to the use of social media and the measurement of word-of-mouth (WOM).[9] As an example we mention a study by De Vries et al. (2012) who determine the drivers of liking and commenting. Brand fans can indicate they like brand posts of brand fan pages or comment on them. Their study is based on data that are obtained from social networking sites.

3.5.5 Household Data and/or Store Level Data?

There is an ongoing debate[10] about the choice to use either household level scanner panel data or store-level scanner data.[11] In this subsection we discuss advantages and disadvantages of both sources.

Panel data, which tracks purchases of a sample of households on an ongoing basis, allows managers to explore differences in purchase behaviors and preferences that lead to segmentation and targeting, to determine how these segments differ in terms of demographic characteristics, to examine brand switching and loyalty patterns, to track new product trial and repeat rates, to understand the impact of marketing variables on purchase timing and stockpiling, and to test theories of consumer behavior (Gupta et al. 1996). Hierarchical-level scanner panel data also have some disadvantages. It has been shown that inferences from panel data are usually not statistically representative (Bucklin and Gupta 1999; Gupta et al. 1996). This is particularly the case for brands with lower market shares and, more general, for low-incidence product categories. Store-level data are widely available to marketing managers, are used as a key resource for managerial decision making, are less expensive for firms to acquire, and require fewer computational resources than household-level data (Chintagunta et al. 2002). While panel and store-level data have several complementary uses, analysts in academic and industry settings use either type of data to predict sales response to price reductions and promotions.

[8] For a survey of these secondary data services, see Malhotra (2010, Chapter 4).

[9] See Sect. 9.7 for models that measure the impact of WOM.

[10] See Bucklin and Gupta (1999); Bodapati and Gupta (2004).

[11] This text is based on Andrews et al. (2011).

However, it is unclear whether the predictions from demand models using panel versus store-level scanner data are more or less biased and under what conditions. Complicating the comparison of promotional response predictions from models applied to panel and store-level data is the fact that different forms of consumer heterogeneity can be captured using the two types of data. With panel data, the focus is on heterogeneity in preferences and responses to marketing activity across households. Since store-level data lack household identifiers (Bodapati and Gupta 2004), heterogeneity recovered by typical store-level applications is actually heterogeneity across store visits (Besanko et al. 2003) and is often referred to as within-store heterogeneity. Bodapati and Gupta (2004) demonstrated that parameters from a store level model explaining within-store heterogeneity can approximate those of panel data models explaining household heterogeneity, especially when sample sizes are very large.

Andrews et al. (2011) show via simulation that demand models with various heterogeneity specifications do *not* produce *more accurate* sales response predictions than a homogeneous demand model applied to store-level data. In another study it is confirmed that accounting for heterogeneity in store-level models does not lead to better predictions.[12]

3.5.6 Big Data

Big Data is clearly "in": a Google search reveals over 1.5 billion hits in early 2014, and the prestigious journal *Marketing Science* features a special issue on Big Data in Marketing. The Big Data Market, at $18.6 B in 2014, is expected to grow to $50 B by 2017. Big Data's origins go back to 2001, when analyst Doug Laney at META Group (now Gartner) defined data growth challenges and opportunities in three dimensions: Volume, Variety and Velocity. Each of these poses a challenge for managers and researchers (Taylor 2014):

- Increasing data Volume undermines our ability to process using on-hand database management tools or traditional data processing applications. The challenges include collection storage, search, sharing, transfer, analysis, and visualization.
- Increasing data Velocity data arriving more quickly is driving us to increasingly faster analysis and faster action. We have to analyze the data more quickly, give ourselves time to act on this analysis by pushing our analysis into the future, and we have to act on our analysis more quickly.
- Increasing data Variety means that data arrives in different formats, such as structured and unstructured. In marketing, structured data include numerical data on consumer purchasing, participation in social media, or exposure to online marketing. Unstructured data, such as text, audio, or even video content are freely

[12] Andrews et al. (2008).

provided by consumers. Examples of applications that yield unstructured data are eye tracking (e.g. via Google Glass), in-store video tracking (Hui et al. 2013), or users interacting with the apps on their smart phones. Unstructured data can arise from interactions between humans, (e.g. a sequence of Facebook messages), from interactions between humans and machines (e.g. a Payment via Paypal) or from interactions between machines (e.g. a GPS device that receives localization data from a satellite). The latter is called "the Internet of Things".

What exactly is considered "Big Data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain.

> "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration" (Magoulas and Lorica 2009).

According to McKinsey (2011), Big Data can create value in five ways. First, Big Data can make information transparent and usable at much higher frequency. Second, analyzing more transactional data in digital helps companies expose variability and conduct controlled experiments to improve decisions. Third, Big Data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services. Fourth, sophisticated analytics can substantially improve the decision-making process. Finally, Big Data can guide product and service development. For instance, manufacturers are using data obtained from sensors embedded in products to create innovative after-sales service offerings such as proactive maintenance (preventive measures that take place before a failure occurs or is even noticed).

What are the implications of Big Data for marketing modelers? According to the Marketing Science call for papers[13]:

> "High volume implies the need for models that are scalable; high velocity opens opportunities for real-time, or virtually real-time, marketing decision making that may or may not be automated; and high variety may require integration across disciplines with the corresponding sensitivity to various methods and philosophies of research."

An early example of Big Data application in marketing is Reimer et al. (2014). The authors take on a data set of over 500,000 customers (Volume) with daily activities (Velocity) at a digital music download provider, which captures marketing activity in two ways (Variety): aggregate-level data on "push" mass media (e.g., TV, radio, print, banner ads) and customer-level data on what was "pulled" by the customer to enable purchase (e.g., permission-based communication, coupons claimed, newsletter emails). This combination should allow companies to profile customers based on their responsiveness to push marketing as well as their pull behavior. Unfortunately, realizing this potential is complicated by the sheer size of the customer base and the lack of a modeling framework combining response-based

[13] Available at http://pubsonline.informs.org/page/mksc/calls-for-papers (accessed 7-April-2014).

segmentation with long-term effect estimation. Historically, modeling in marketing has made use of either information-rich household panel data for a small number of households or data aggregated over consumers at the store or market-level. On the one hand, household panel data are used to infer the effects of mostly tactical actions (price and promotions) on the consumer's decision process (e.g., utility, consideration, learning, etc.) based on relatively small samples of consumers (typically 2,000–10,000). On the other hand, time series models using aggregate data are used to understand the short- and long-run effectiveness of aggregate-level spending (e.g. advertising) on sales. Reimer et al. (2014) parsimoniously combine individual-level and aggregate modeling to deal with large data sets at mixed aggregation levels. Latent-class segmentation on customer-level classifies thousands of consumers based on short-term pull marketing response and profiles the segments based on other customer information.[14] Next, Vector-AutoRegressive (VAR) models for each segment reveal substantial differences in long-term sales response to push marketing and guides a better marketing budget allocation across marketing actions and segments.

3.5.7 Subjective Data

3.5.7.1 Justification

In the previous sections of this chapter we dealt with objective data, data that represent observed or observable quantities. The question we address now is how we can model markets in the absence of objective data.

In the absence of models, decision makers (DM's) make judgments based on their own experiences, the experience of colleagues or the habits and beliefs that are part of an organizational culture. The judgments reflect implicit assumptions about response parameters. Rarely, however, do the implicit parameter values remain constant across conditions. It is especially for this reason that a "model of man" can outperform "man". That is, a *model* of repeated *judgments* made by one person can better predict the actual outcomes of those very *judgments*.

A benefit that we propose relevant to the use of models from subjective data is that it formalizes the process of predictions, and allows the decision maker to diagnose their accuracy. In addition, it forces decision makers to be explicit about how they believe variables under their control affect certain performance measures. And, when multiple experts provide judgments, the subjective estimation separately for each expert shows the nature and the extent of differences. If such differences get resolved before a decision is made, then the prevailing perspective gets disseminated. When experts cannot agree, future outcomes can serve as the basis for a determination of relative accuracy of alternative approaches.

[14] Latent class segmentation techniques and VAR models are discussed in Volume II.

The experts who supply subjective data are called assessors. The ultimate decision makers, internal and external consultants, and sales representatives are all potential assessors. Members of the sales force can be especially helpful when clues about future sales levels are gathered. The sales force members are in contact with customers, and this should allow them to provide relevant expertise. In addition, if sales forecasts are used for the determination of sales quotas it is helpful to have sales force members involved in the process. For example, their participation will increase their confidence in the quotas being fair, and this will increase their motivation to achieve the quotas. Of course, there is also the possibility that they will try to "game" the system.[15]

Sometimes expertise is gathered from various stakeholders whose differences of opinion can be vast. A broad set of stakeholders may include company executives, dealers, distributors, suppliers, consultants, forecast experts, etc. Some or all of these stakeholders may be asked to constitute a jury of executive opinion. When representatives of various groups of stakeholders get together, the purpose of the meeting is for the group to come as close as possible to a single judgment. A variation on this is the Delphi method in which experts write down their judgments in a first round. Each expert receives summary information about the independent judgments made, and this information can influence the experts judgments in subsequent rounds.

We now briefly discuss methods of obtaining subjective point estimates, response functions and probability assessments.

3.5.7.2 Obtaining Subjective Estimates

We consider three categories of subjective data obtainable from a single assessor:

1. point estimation;
2. response functions;
3. probability assessments.

Subjective data consist of opinions (judgments) and intentions. See, for example, Leeflang and Peluso (2012). Intentions are indications individuals provide about their planned behavior or about the decisions they plan to make or the outcomes of those decisions. Intention surveys are frequently used to forecast the demand for a new product (Jamieson and Bass 1989). Another application involves estimation of the impact of a possible entrant on a market (Alsem and Leeflang 1994).

Point Estimation

Point estimation provides partial information about the distribution of an unknown quantity. Suppose we ask an expert: "What is the probability that sales will be no

[15] Compare Lilien and Rangaswamy (2004, pp. 130–131).

more than one thousand units next month?" and it is 70 % in the expert's judgment. The expert gives us information about the cumulative distribution function of next month's sales, but it is only partial in the sense that we do not know the probability that sales will be no more than, for example, five hundred units. The expert's answer gives us one point (*A* in Fig. 3.1) of a cumulative distribution function. Figure 3.1

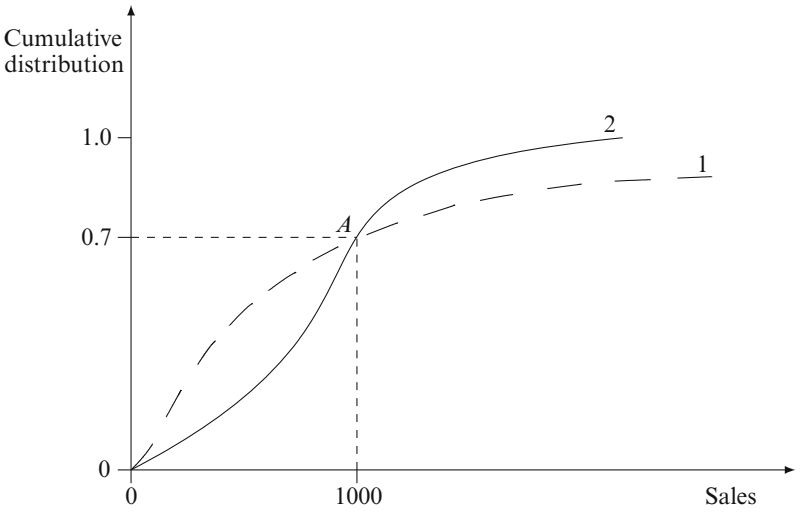

Fig. 3.1 Point estimate (*A*) and two cumulative distribution functions

shows two of the many possible functions that may pass through *A*. We note that the manner in which the question is formulated is important. It has, for example, no meaning for someone who is unfamiliar with the notion of probability. We can provide some training, or we can change the descriptions, for example by using "chances" or "odds" which may be more familiar terminology. If we only ask for a point estimate, we are often looking for a measure of central tendency. The answer may provide an estimate of the mode, the median, or mean, as illustrated below:

1. "What is your estimate of the most likely level of sales?" gives an estimate of the mode.
2. "What level of sales do you estimate you have an even chance of reaching?" provides an estimate of the median.
3. "What level of sales do you expect?" results in an estimate of the mean.

This type of questioning is appropriate if we desire to obtain estimates of such quantities as market share or sales. However, the assessor will by necessity give an estimate that is conditional upon an assumed level of various marketing instruments (for example, price, size of the sales force, etc.).

Response Functions

A logical extension is to obtain a set of point estimates, one for each of a number of values of the marketing instruments, which generates a point estimate of a response function. At the same time, the construction of a subjective response curve enables us to estimate quantities that cannot normally be assessed directly, such as elasticities or response parameters. We consider an example from the ADBUDG model[16]:

$$m_t = \alpha + (\beta - \alpha)\frac{a_t^\delta}{\gamma + a_t^\delta} \tag{3.1}$$

where

m_t = market share of a brand in period t,

a_t = advertising expenditures of the brand in period t, and

$\alpha, \beta, \gamma, \delta$ = the parameters.

Suppose we want to estimate the parameters of (3.1) subjectively. Specifically, assume that the brand in question is a detergent with a sizable market share. The obvious person to assist the model builder with parameter estimates is the brand manager. The brand manager has at least some knowledge of how the market operates, which competitors matter, and so on. Still, we cannot obtain the desired information if we simply ask: "what do you think α, β, γ and δ are?" Instead, we may ask: "what do you think market share will be in a few years, if all advertising is stopped from now on?" The answer is an estimate of α since for advertising equal to zero, market share is equal to α according to (3.1). Similarly we can ask the brand manager what will happen to market share if an unlimited amount is spent on advertising. The answer to this question is an estimate of β, since if advertising is very large, $a_t^\delta/(\gamma + a_t^\delta)$ approaches one, and thus m approaches $\alpha + (\beta - \alpha) = \beta$.

The next question is what market share the brand manager expects to obtain if advertising is left at its current level, say $a_t = c$ dollars. Let the answer be m_c. And we may ask what market share will occur if the advertising budget is increased by 50 %. Let the corresponding share be m_{c+}. We then have the following two points of the market share function:

$$m_c = \alpha + (\beta - \alpha)\frac{c^\delta}{\gamma + c^\delta} \tag{3.2}$$

$$m_{c+} = \alpha + (\beta - \alpha)\frac{(1.5c)^\delta}{\gamma + (1.5c)^\delta}\,. \tag{3.3}$$

[16] See Little (1970).

Since α and β are already "known", (3.2)–(3.3) is a system of two equations in two unknowns γ and δ. Solving this system yields estimates of the final two parameters.

To complete the process, suppose the brand manager estimates that without advertising, market share will drop to 10 % ($\alpha = 0.10$) and with saturation advertising it will reach 70 % ($\beta = 0.70$). Suppose further that with the current budget ($810,000), market share is expected to be 0.40, and with 50 % higher advertising market share should become 0.415. Estimates for γ and δ are then found by solving:

$$0.40 = 0.10 + 0.60\frac{(810,000)^{\delta}}{\gamma + (810,000)^{\delta}}$$

$$0.415 = 0.10 + 0.60\frac{(1,215,000)^{\delta}}{\gamma + (1,215,000)^{\delta}}.$$

The estimated values are approximately: $\gamma = 30$ and $\delta = 0.25$. Figure 3.2 shows the brand manager's implicit market share function. One may object that the four parameters in (3.2)–(3.3) are estimated from four observations. Hence the model fits the data perfectly. But there is no guarantee that additional subjective judgments fit the market share function shown in Fig. 3.2. Thus, we prefer to collect additional observations. For example, we may elicit market share estimates for advertising expenditures equal to the current budget plus 20 %, plus 40 %, ... minus 20 %, minus 40 %, and so on, thus providing a scatter of points through which a market

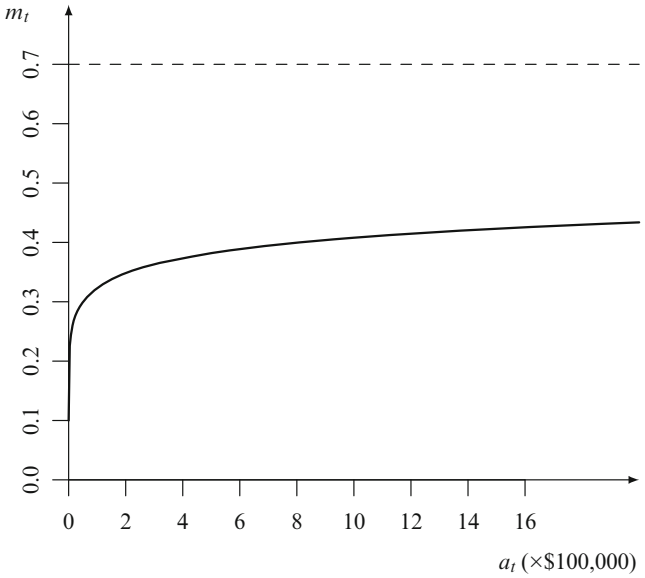

Fig. 3.2 A brand manager's implicit market share function

share function can be fitted using (non-linear) estimation methods. In that case, deviations of the subjective estimates from the fitted curve allow us to check the consistency of a manager's estimates. In case of systematic deviations, we can consider an alternative functional form, in close cooperation with the manager.

The disadvantage of asking additional judgments is that it requires more time and effort from the manager. These judgments only capture the manager's expectations about market share, given values for one marketing instrument. However, extensions to multiple predictors are straightforward. Importantly, the judgments about market share given advertising are conditional on, for example, the brand's price and the marketing instruments for other brands.

Probability Assessment

We now consider the notion of a probability assessment, that is the assessment of a probability distribution. We start with a simple illustration to suggest the usefulness of eliciting an entire distribution. Suppose that a firm considers a price increase. The desirability of the increase depends on the subjective estimate of how likely its sales manager considers the possibility of a price reaction by a major competitor. Assume that a modal value is used for an initial judgment. If the sales manager has much confidence in this estimate, implying a tight distribution such as f_A in Fig. 3.3, then the modal value gives very precise information about the competitive price reaction and this modal value is a good basis for (partly) determining the optimal price increase, if any.

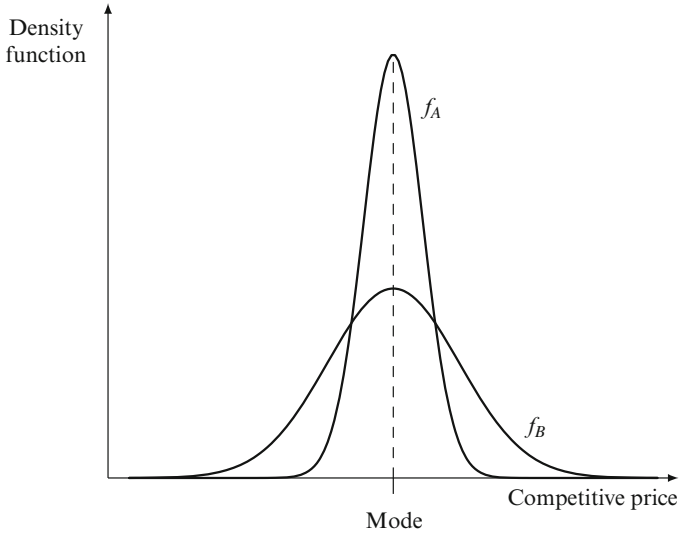

Fig. 3.3 Subjective estimates of a competitive reaction

On the other hand, if f_B in Fig. 3.3 is the proper density function, then the modal value is a very uncertain estimate. If we know the prevailing density function we can conduct sensitivity analyses, i.e. determine how the optimal price increase depends on alternative competitive price reactions. And, in case the sensitivity is very high, it may be useful to determine whether the uncertainty about the price reaction can be reduced.

Examples

Suppose that a DM is interested in a unit sales forecast for a new brand in period t: $\hat{S}_t$. To this end the DM asks two sales representatives (assessors 1 and 2) the following questions:

- what is your lowest estimate of S_t: S_t^L;
- what is your most likely estimate of S_t: S_t^M;
- what is your highest estimate of S_t: S_t^H.

And, assuming that the three values are the only possible outcomes, each assessor also provides an indication of the probabilities of occurrence. We show the representatives' answers in Table 3.5. Assessor 1 is more optimistic about the new

Table 3.5 Sales estimates and probabilities for two assessors

	Assessor 1		Assessor 2	
	Estimates	Probabilities	Estimates	Probabilities
S_t^L	6 million	0.1	4 million	0.3
S_t^M	8 million	0.8	7 million	0.5
S_t^H	10 million	0.1	10 million	0.2
$\mathrm{E}(S_t)$	8 million		6.7 million	
$\mathrm{Var}(S_t)$	0.8 million^2		4.41 million^2	

brand's sales than assessor 2, except for the highest estimate which is the same for the two assessors. Assessor 1 also has a much higher subjective probability for the most likely value of sales than assessor 2 does. As a result, the expected value for sales is higher while the variance is lower for assessor 1.

In the absence of knowledge about the shape of a subjective probability distribution, we can assess the uncertainty around the estimates by a variety of methods. One method focuses on the fractions of the Cumulative Distribution Function (CDF). Typically, five fractiles, 0.01, 0.25, 0.50, 0.75, and 0.99 are assessed. The first assessment is the 0.50 fractile which is the probability that sales in period t will be less than 0.50. $Q_{0.50}$ is obtained from the question: "Considering all possible levels of sales, what is the amount for which it is equally likely that sales will be more and that sales will be less than this value?" For assessor 2 in Table 3.5 the response would have been 7 Million, hence $Q_{0.50}$ = 7 Million. We can obtain the 0.25 and

0.75 fractiles by asking: "If sales are, in fact, less than 7 Million, what amount of sales would divide the interval from 0 to 7 Million units into equally likely parts?" The resulting value is the 0.25 fractile, denoted as $Q_{0.25}$. The 0.75 fractile can be obtained from the question: "How would you divide the interval of sales over 7 Million units into equally likely parts?" Finally, the values for $Q_{0.01}$ and $Q_{0.99}$ can be obtained by asking: "What value of sales would you use such that the changes of sales being greater (less) than this value is only one in 100?" We can then construct a CDF-curve of sales by plotting the fractiles and corresponding sales values. We provide an example in Fig. 3.4.

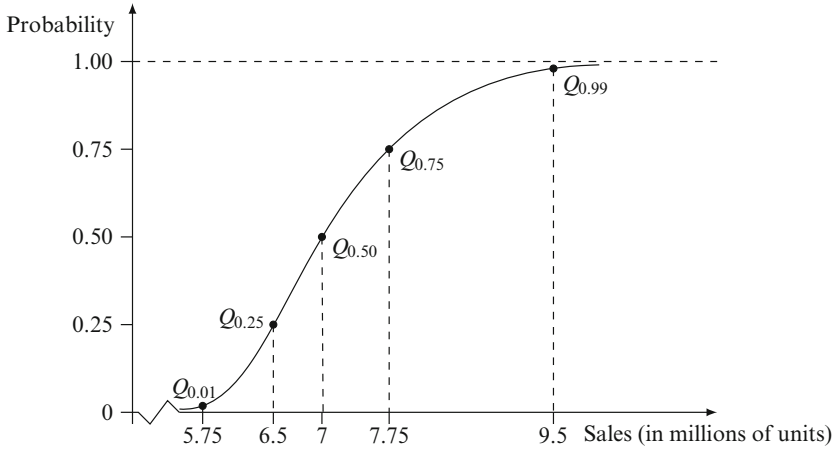

Fig. 3.4 A cumulative subjective probability function for sales (assessor 2)

In some cases it is possible to specify the distribution *a priori*. It then suffices to obtain estimates of the parameters. We consider a few examples. Suppose the quantity we want to assess is, in the mind of the assessor, a random variable that is *normally* distributed. This distribution is then characterized by its two parameters, the mean μ, and the standard deviation σ. Since most people cannot provide direct estimates of the standard deviation, we may use two questions to estimate both μ and σ, for example:

- "What is your estimate of sales such that there is a 2.5 % chance that it will be higher?" (S_t^H).
- "What is your estimate of sales such that there is a 2.5 % chance that it will be below that level?" (S_t^L).

The mean is then estimated by taking the average of S_t^L and S_t^H:

$$\hat{\mu} = \frac{S_t^L + S_t^H}{2}. \tag{3.4}$$

Since 95 % of the possible values of a normally distributed random variable lie within ±1.96 standard deviations of the mean, the standard deviation is estimated by:

$$\hat{\sigma} = \frac{S_t^H - S_t^L}{3.92}. \tag{3.5}$$

For asymmetric distributions such as the beta distribution, low (S_t^L) high (S_t^H), and modal (S_t^M) estimates are necessary. The estimates of the mean and standard deviation are then:v

$$\hat{\mu} = \frac{S_t^L + 4S_t^M + S_t^H}{6} \tag{3.6}$$

$$\hat{\sigma} = \frac{S_t^H - S_t^L}{6}. \tag{3.7}$$

So far we have considered distributions of the values a random variable may take. This idea can also be used for distributions of individual points on a response function. For example, we could construct an optimistic, a pessimistic and a modal response function. Figure 3.5 shows an example of a market share-advertising response function for which the degree of uncertainty stays approximately constant over the whole range of values of advertising. This is not the case in Fig. 3.6, where the uncertainty is modest for advertising expenditures within the interval *RA*, but large outside that interval. This could result if the firm's advertising expenditure levels have normally fallen within *RA*. Managers will be less confident about points on the response curve that are outside their experience.

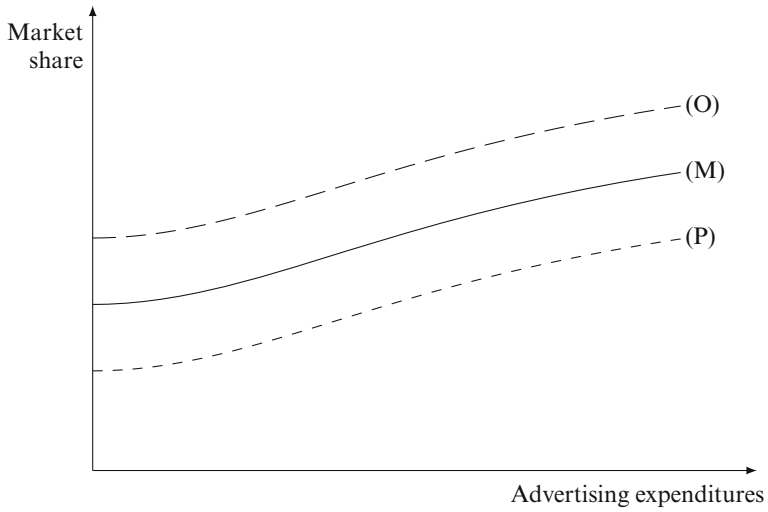

Fig. 3.5 Optimistic (O), modal (M) and pessimistic (P) market share-advertising response functions (constant uncertainty)

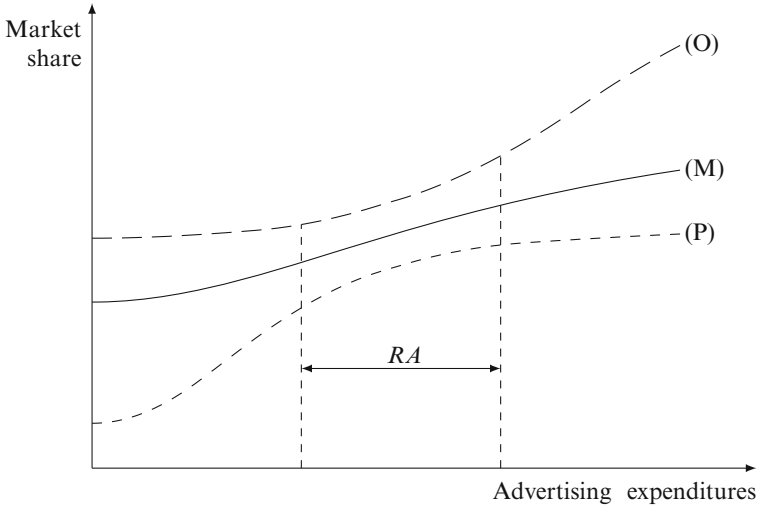

Fig. 3.6 Optimistic (O), modal (M) and pessimistic (P) market share-advertising response functions (varying uncertainty)

References

Alsem, K.J., Leeflang, P.S.H.: Predicting advertising expenditures using intention surveys. Int. J. Forecast. **10**, 327–337 (1994)

Andrews, R.L., Currim, I.S., Leeflang, P.S.H.: A comparison of sales response predictions from demand models applied to store-level vs. panel data. J. Bus. Econ. Stat. **29**, 319–326 (2011)

Andrews, R.L., Currim, I.S., Leeflang, P.S.H., Lim, J.: Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? Int. J. Res. Mark. **25**, 22–33 (2008)

Ansari, A., Essegaier, S., Kohli, R.: Internet recommendation systems. J. Mark. Res. **37**, 363–375 (2000)

Bass, F.M., Bruce, N., Majumdar, S., Murthi, B.P.S.: Wearout effects of different advertising themes: A dynamic Bayesian model of the advertising-sales relationship. Mark. Sci. **26**, 179–195 (2007)

Bearden, W.O., Netemeyer, R.G., Haws, K.: Handbook of Marketing Scales, 3rd edn. Sage, Los Angeles (2011)

Bendle, N., Farris, P., Pfeifer, P., Reibstein, D.: Metrics that matter to marketing managers. Mark. J. Res. Manag. **6**(1), 18–23 (2010)

Besanko, D., Dubé, J.P., Gupta, S.: Competitive price discrimination strategies in a vertical channel using aggregate retail data. Manag. Sci. **49**, 1121–1138 (2003)

Bijmolt, T.H.A., Leeflang, P.S.H., Block, F., Eisenbeiss, M., Hardie, B.G.S., Lemmens, A., Saffert, P.: Analytics for customer engagement. J. Serv. Res. **13**, 341–356 (2010)

Blattberg, R.C., Kim, B.D., Neslin, S.A.: Database Marketing: Analyzing and Managing Customers. Springer Science, New York (2008)

Bodapati, A.V., Gupta, S.: The recoverability of segmentation structure from store-level aggregate data. J. Mark. Res. **41**, 351–364 (2004)

Brand, M.J., Leeflang, P.S.H.: Research on modeling industrial markets. In: G. Laurent, G.L. Lilien, B. Pras (eds.) Research Traditions in Marketing. Kluwer, Boston (1994)

Bucklin, R.E., Gupta, S.: Commercial use of UPC scanner data: Industry and academic perspectives. Mark. Sci. **18**, 247–274 (1999)

Bucklin, R.E., Sismeiro, C.: Click here for interact insight: Advances in clickstream data analysis in marketing. J. Interact. Mark. **23**, 35–48 (2009)

Chessa, A.G., Murre, J.M.J.: A neurocognitive model of advertisement content and brand name recall. Mark. Sci. **26**, 130–141 (2007)

Chintagunta, P.K., Dubé, J.P., Singh, V.: Market structure across stores: An application of a random coefficients logit model with store level data. In: P.H. Franses, A. Montgomery (eds.) Advances in Econometrics, vol. 16. Emerald Group Publishing Limited, Binley (2002)

De Vries, L., Gensler, S., Leeflang, P.S.H.: Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. J. Interact. Mark. **26**, 83–91 (2012)

De Vries, L., Gensler, S., Leeflang, P.S.H.: Effects of owned and earned social media and paid traditional media on consumer mind-set metrics and acquisition. Working paper, University of Groningen, Groningen (2014)

Fader, P.S., Hardie, B.G.: Modeling consumer choice among SKUs. J. Mark. Res. **33**, 442–452 (1996)

Farris, P.W., Bendle, N.T., Pfeifer, P.E., Reibstein, D.J.: Marketing Metrics: 50+ Metrics Every Executive Should Master. Wharton School Publishing, Upper Saddle River (2006)

Franses, P.H., Paap, R.: Quantitative Models in Marketing Research. Cambridge University Press, Cambridge (2001)

Gupta, S., Chintagunta, P.K., Kaul, A., Wittink, D.R.: Do household scanner data provide representative inferences from brand choices: A comparison with store data. J. Mark. Res. **33**, 383–398 (1996)

Hui, S.K., Huang, Y., Suher, J., Inman, J.J.: Deconstructing the "first moment of truth": Understanding unplanned consideration and purchase conversion using in-store video tracking. J. Mark. Res. **50**, 445–462 (2013)

Jamieson, L., Bass, F.M.: Adjusted stated intention measures to predict trial purchase of new products: A comparison of models and methods. J. Mark. Res. **26**, 336–345 (1989)

Kamakura, W.A., Wedel, M.: Statistical data fusion for cross-tabulation. J. Mark. Res. **34**, 485–498 (1997)

Kotler, P., Armstrong, G.: Principles of Marketing, 14th edn. Pearson, Boston (2012)

Leamer, E.E.: Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York (1978)

Leeflang, P.S.H.: Paving the way for distinguished marketing. Int. J. Res. Mark. **28**, 76–88 (2011)

Leeflang, P.S.H., Olivier, A.J.: Bias in consumer panel and store audit data. Int. J. Res. Mark. **2**, 27–41 (1985)

Leeflang, P.S.H., Peluso, A.M.: Knowledge generation in marketing. In: A. Diamantopoulos, W. Fritz, L. Hildebrandt (eds.) Quantitative Marketing and Marketing Management: Marketing Models and Methods in Theory and Practice. Springer Gabler, Wiesbaden (2012)

Lilien, G.L., Rangaswamy, A.: Marketing Engineering, 2nd rev. edn. Trafford Publishing, Bloomington (2004)

Little, J.D.C.: Models and managers: The concept of a decision calculus. Manag. Sci. **16**, B466–B485 (1970)

Little, J.D.C.: Integrated measures of sales, merchandising and distributions. Int. J. Res. Mark. **15**, 475–485 (1998)

Little, J.D.C.: Models and managers: The concept of a decision calculus. Manag. Sci. **50**(12), 1841–1853 (2004)

Magoulas, R., Lorica, B.: Introduction to Big Data, release 2.0 edn. O'Reilly Media, Sebastopol (2009)

Malhotra, N.K.: Marketing Research, 6th edn. Prentice-Hall International Editions, Upper Saddle River (2010)

McDonnell Feit, E., Wang, P., Bradlow, E., Fader, P.: Fusing aggregate and disaggregate data with an application to multiplatform media consumption. J. Mark. Res. **50**, 348–364 (2013)

McKinsey Global Institute: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey & Company (2011)

Pancras, J., Sudhir, K.: Optimal marketing strategies for a customer data intermediary. J. Mark. Res. **44**, 560–578 (2007)
Pauwels, K.: It's Not the Size of the Data – It's How You Use It: Smarter Marketing with Analytics and Dashboards. AMACOM, New York (2014)
Pauwels, K., Leeflang, P.S.H., Teerling, M.L., Huizingh, K.R.E.: Does online information drive offline revenues? Only for specific products and consumer segments! J. Retail. **87**, 1–17 (2011)
Pauwels, K., Stacey, C., Lackmann, A.: Beyond likes and tweets: Marketing, social media content, and store performance. Report MSI 13-125, Marketing Science Institute (2013)
Plat, F.W.: Modelling for markets: applications of advanced models and methods for data analysis. Unpublished Ph.D. thesis, University of Groningen, Groningen (1988)
Qian, Y., Xie, H.: Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches. Mark. Sci. **33**, 437–448 (2014)
Reimer, K., Rutz, O.J., Pauwels, K.: How online consumer segments differ in long-term marketing effectiveness. J. Interact. Mark. **28**, forthcoming (2014)
Shoemaker, R.W., Pringle, L.G.: Possible biases in parameter estimation with store audit data. J. Mark. Res. **17**, 91–96 (1980)
Stephen, A.T., Galak, J.: The effects of traditional and social earned media on sales: A study of a microlending marketplace. J. Mark. Res. **49**, 624–639 (2012)
Taylor, J.: To succeed with big data, begin with the decision in mind. Information Management Online (2014) [Published online on February 10; accessed 7 Apr 2014]
Telang, R., Boatwright, P., Mukhopadhyay, T.: A mixture model for internet search engine visits. J. Mark. Res. **41**, 206–214 (2004)
Van den Bulte, C., Wuyts, S.: Social Networks and Marketing. MSI Relevant Knowledge Series. Marketing Science Institute, Cambridge (2007)
Van Heerde, H.J., Gijsbrechts, E., Pauwels, K.: Winners and losers in a major price war. J. Mark. Res. **45**, 499–518 (2008)
Van Laer, T., De Ruyter, K.: In stories we trust: How narrative apologies provide cover for competitive vulnerability after integrity-violating blog posts. Int. J. Res. Mark. **27**, 164–174 (2010)
Verhoef, P.C., Hoekstra, J.C., Van der Scheer, J., De Vries, L.: Competing on analytics: Status quo van customer intelligence in Nederland. Onderzoeksrapport No. 2. Customer Insights Center, Groningen (2009)
Verhoef, P.C., Spring, P.N., Hoekstra, J.C., Leeflang, P.S.H.: The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. Decis. Support Syst. **34**, 471–487 (2002)
Vriens, M.: The Insights Advantage: Knowing How To Win. iUniverse, Bloomington (2012)
Wooldridge, J.M.: Introductory Econometrics: A Modern Approach, 5th edn. Cengage Learning, Mason (2012)
Wuyts, S.: Extra-role behavior in buyer-supplier relationships. Int. J. Res. Mark. **24**, 301–311 (2007)
Wuyts, S., Stremersch, S., Van den Bulte, C., Franses, P.H.: Vertical marketing systems for complex products: A triadic perspective. J. Mark. Res. **41**, 479–487 (2004)
Yao, S., Mela, C.F.: Online auction demand. Mark. Sci. **27**, 867–885 (2008)

Chapter 4
Estimation and Testing

4.1 Introduction

In this chapter we turn to estimation, the step in the marketing model building process that follows model specification, where we consider methods and procedures for obtaining numerical values for the model parameters in the model. Throughout this chapter we will mainly consider the case where the independent variables are linearly related to the dependent variable or where one or more variables can be transformed in such a way that the relation between the variables becomes linear. In such cases, it is appropriate to estimate a linear model. Linear models do not only provide reasonable specifications for many practical applications, they are also attractive for a careful treatment of model assumptions, and for a conceptual explanation of the basis for the assumptions. Most of the principles that apply to the linear model remain relevant as long as nonlinear effects for the original variables can be accommodated by transforming variables (so that the transformed variables are linearly related).

We discuss methods and procedures for estimation of linear models in Sect. 4.2, as well as the assumptions that they require. In Sect. 4.3 we discuss measures of the goodness of fit and test for the significance of the model and its individual parameters. In Sect. 4.4 we apply the general linear model to data of the Verhouten case, and estimate Eqs. (2.26) and (2.27). Estimation methods for pooled models are discussed in Sect. 4.5.

We note that our discussion is limited to standard procedures. More detailed treatments can be found in econometrics textbooks. Goldberger (1998) and Wooldridge (2012) provide very clear and lucid discussions. Intriligator et al. (1996), Pindyck and Rubinfeld (1997), Johnston and Dinardo (1997), Baltagi (2011), and Gujarati and Porter (2009) are standard texts that contain relatively easy-to-follow treatments. Judge et al. (1985), Heij et al. (2004), Verbeek (1972) and Greene (2012)

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_4

are general state-of-the-art textbooks. Amemiya (1985), Hamilton (1994), Davidson and MacKinnon (2004) and Cameron and Trivedi (2009) provide advanced descriptions.

4.2 The General Linear Model

In Sect. 2.4 we discussed that an important decision during the specification step is the selection of a functional form that provides a reasonable mathematical representation of the relations between the variables in the model. In this section we explain how linear relations between one criterion variable and one (Sect. 4.2.1) or more (Sect. 4.2.2) predictor variables are estimated. In Sect. 4.2.3 we list the assumptions that the estimation procedures require.

4.2.1 One Explanatory Variable

Let us first assume the following simple linear additive relation for T time-series observations:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \quad t = 1, \ldots, T \tag{4.1}$$

where, at time t:

y_t = the value of the criterion variable,

x_t = the value of the predictor variable,

ε_t = the (unobserved) value of the disturbance term,

α = the unknown intercept,

β = the unknown slope parameter.

In Eq. (4.1), β is the systematic change in y_t when x_t increases by one unit, and α represents the value y_t when x_t equals zero. When data are available for both variables in the model for $t = 1, \ldots, T$, the most common method for estimating the unknown parameters is Ordinary Least Squares (OLS).

A basic objective for model estimation is that for each $t = 1, \ldots, T$, the estimated value of the criterion variable (indicated by $\hat{y}_t$) is close to the observed value y_t. This is measured by the *residual*, which is defined as:

$$e_t = y_t - \hat{y}_t. \tag{4.2}$$

For each t, given that we know the value for the independent variable, we can obtain the estimated value for the criterion variable by replacing the unknown parameters by the estimated values, and assuming that the best prediction for the disturbance term is zero:

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t. \tag{4.3}$$

When the model parameters are estimated with OLS, the requirement that the estimated values of the criterion variable should be close to the observed ones across all observations, is translated into the objective that the sum of squared residuals should be minimized. That is: the parameters estimates $\hat{\beta}$ and $\hat{\alpha}$ are determined in such a way that:

$$\sum_{t=1}^{T} (e_t)^2 \tag{4.4}$$

is minimal. Substituting (4.2) and (4.3) in Eq. (4.4), the objective for OLS estimation can be summarized as:

$$\min_{\hat{\alpha},\hat{\beta}} \left[\sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t \right)^2 \right]. \tag{4.5}$$

The minimum for the part between square brackets in Eq. (4.5) can be found by taking partial derivatives with respect to $\hat{\alpha}$ and $\hat{\beta}$, and setting those equal to zero:

$$\frac{\partial \left[\sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t \right)^2 \right]}{\partial \hat{\alpha}} = 0 \tag{4.6}$$

$$\frac{\partial \left[\sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t \right)^2 \right]}{\partial \hat{\beta}} = 0. \tag{4.7}$$

Let us rewrite the part between square brackets:

$$\sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t \right)^2 = \sum_{t=1}^{T} \left(y_t^2 + \hat{\alpha}^2 + \left(\hat{\beta} x_t \right)^2 - 2 y_t \hat{\alpha} - 2 \hat{\beta} y_t x_t + 2 \hat{\alpha} \hat{\beta} x_t \right). \tag{4.8}$$

Substituting (4.8) in Eqs. (4.6) and (4.7) and taking derivatives we obtain:

$$\sum_{t=1}^{T} \left(2\hat{\alpha} - 2 y_t + 2 \hat{\beta} x_t \right) = 0 \tag{4.9}$$

$$\sum_{t=1}^{T}\left(2\hat{\beta}x_t^2 - 2y_t x_t + 2\hat{\alpha}x_t\right) = 0. \tag{4.10}$$

Solving Eq. (4.9) for $\hat{\alpha}$ results in the following expression:

$$\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T} y_t - \hat{\beta}\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) = \bar{y} - \hat{\beta}\bar{x}, \tag{4.11}$$

and substituting Eq. (4.11) in (4.10) we can derive for $\hat{\beta}$:

$$\begin{aligned}
&\hat{\beta}\sum_{t=1}^{T} x_t^2 - \sum_{t=1}^{T} x_t y_t + \sum_{t=1}^{T}\left(\bar{y} - \hat{\beta}\bar{x}\right)x_t = 0 \\
\Rightarrow\ &\hat{\beta}\left(\sum_{t=1}^{T} x_t^2 - \bar{x}\sum_{t=1}^{T} x_t\right) = \sum_{t=1}^{T} x_t y_t - \bar{y}\sum_{t=1}^{T} x_t \\
\Rightarrow\ &\hat{\beta}\left(\sum_{t=1}^{T}\left(x_t^2 - \bar{x}^2\right)\right) = \sum_{t=1}^{T}(x_t y_t - \bar{y}\bar{x}) \\
\Rightarrow\ &\hat{\beta}\left(\sum_{t=1}^{T}\left(x_t^2 - 2\bar{x}^2 + \bar{x}^2\right)\right) = \sum_{t=1}^{T}(x_t y_t - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}) \\
\Rightarrow\ &\hat{\beta}\left(\sum_{t=1}^{T}\left(x_t^2 - 2x_t\bar{x} + \bar{x}^2\right)\right) = \sum_{t=1}^{T}(x_t y_t - x_t\bar{y} - \bar{x}y_t + \bar{x}\bar{y}) \\
\Rightarrow\ &\hat{\beta}\left(\sum_{t=1}^{T}(x_t - \bar{x})^2\right) = \sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y}),
\end{aligned}$$

so that the OLS estimator for β equals:

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}. \tag{4.12}$$

4.2.2 The K-Variable Case

The basic model (4.1) can be extended to include K predictor variables:

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t \tag{4.13}$$

where, in addition to the variables for Eq. (4.1), we define for $t = 1,\ldots,T$:

$$x_{kt} = \text{the value of the } k\text{-th predictor variable, } k = 1,\ldots,K.$$

We can rewrite the relations in Eq. (4.13) as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{K1} \\ 1 & x_{12} & x_{22} & \ldots & x_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1T} & x_{2T} & \ldots & x_{KT} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix} \tag{4.14}$$

which in matrix notation becomes:

$$y = X\beta + \varepsilon \tag{4.15}$$

where

y = a column vector of size T with values of the criterion variable,

X = a matrix of dimensions $(T \times K + 1)$ with a column of ones and values of the K predictor variables,

β = a column vector of $K + 1$ unknown parameters, and

ε = a column vector of T disturbance terms.

In Appendix A we discuss the most important aspects and issues of matrix and vector calculation that are used throughout the text of this monograph.

Parameter estimates for Eq. (4.13) are obtained analogously to the process shown in (4.4)–(4.12). Thus, the OLS estimates of the parameters $\alpha,\beta_1,\ldots,\beta_K$ in (4.13) are the values $\hat{\alpha},\hat{\beta}_1,\ldots,\hat{\beta}_K$ which minimize the sum or the squared values of the residuals $e_1,\ldots,e_T$, also known as the Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{t=1}^{T} (e_t)^2 = \sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \sum_{k=1}^{K} \hat{\beta}_k x_{kt} \right)^2 . \tag{4.16}$$

In matrix notation (4.16) becomes:

$$\begin{aligned} \text{RSS} = e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}. \end{aligned} \tag{4.17}$$

Differentiating RSS with respect to $\hat{\beta}$ and setting the derivatives equal to zero, we obtain:

$$-2X'y + 2X'X\hat{\beta} = 0 \tag{4.18}$$

or

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{4.19}$$

This expression for the OLS estimator of β is similar to the corresponding expression of $\hat{\beta}$ in the two-variable case, see Eq. (4.12).[1]

For statistical interference about the parameters of the linear model we also need a specification of the probability distribution of the disturbance terms. We discuss such a specification and associated model assumptions in Sect. 4.2.3.

4.2.3 Model Assumptions

When OLS is employed to obtain estimates of the parameter in a model, several assumptions about the model elements in (4.1) or (4.13) need to be satisfied. Four of these concern the disturbance term:

1. $\mathrm{E}(\varepsilon_t) = 0$ for all t;
2. $\mathrm{Var}(\varepsilon_t) = \sigma^2$ for all t;
3. $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t'}) = 0$ for $t \neq t'$;
4. ε_t is normally distributed.

Two other assumptions are:

5. There is no relation between the predictors and ε_t, i.e. $\mathrm{Cov}(x_t, \varepsilon_t) = 0$ (one-variable case). In other words the x_t are nonstochastic, exogenous or "fixed". For the K-variable case this implies $\mathrm{Cov}(X', \varepsilon) = 0$, in which case we have $\mathrm{E}(\varepsilon \mid X) = \mathrm{E}(\varepsilon) = 0$. If the covariance between the disturbance term and an independent variable is not zero, we encounter the problem of endogeneity.
6. For the K-variable case, the matrix of observations X has full rank, which is the case if the columns in X are linearly independent (see footnote 1). This means that none of the independent variables is constant, and that there are no exact linear relationships among the independent variables.

The variance-covariance matrix of the disturbance vector summarizes the relationships in ε, the T-vector of disturbances, and is denoted by $\mathrm{Var}(\varepsilon)$. Its dimensions are $(T \times T)$. The elements on the diagonal of $\mathrm{Var}(\varepsilon)$ are the variances of the elements

[1] We assume that X has rank $K+1$ (or $X'X$ is nonsingular), and therefore its inverse $(X'X)^{-1}$ exists. See also Appendix A.8.

in ε: the element in row 1, column 1 is $\text{Var}(\varepsilon_1)$, the element in row 2, column 2 is $\text{Var}(\varepsilon_2)$, and so on until the element in row T, column T, which equals $\text{Var}(\varepsilon_T)$. The off-diagonal elements in $\text{Var}(\varepsilon)$ represent the covariances between the disturbances: the element in row t_1 and column t_2 equals $\text{Cov}(\varepsilon_{t_1}, \varepsilon_{t_2})$. By definition, $\text{Var}(\varepsilon)$ is a symmetrical matrix.

If conditions 2 and 3 are both satisfied, $\text{Var}(\varepsilon)$ has the following structure:

$$\text{Var}(\varepsilon) = \sigma^2 I \tag{4.20}$$

where

$$I = \text{a } T \times T \text{ identity matrix.}$$

In Eq. (4.19) we presented the vector of OLS estimates for the parameters in the model:

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{4.21}$$

If assumptions 1 and 5 hold, the OLS estimate for β is unbiased, which means that its expected value equals β:

$$\begin{aligned}
\text{E}(\hat{\beta}) &= \text{E}\left((X'X)^{-1}X'y\right) \\
&= \text{E}\left((X'X)^{-1}X'(X\beta+\varepsilon)\right) \\
&= \beta + \text{E}\left((X'X)^{-1}X'\varepsilon\right) \\
&= \beta + (X'X)^{-1}X'\text{E}(\varepsilon) \\
&= \beta.
\end{aligned}$$

Let us denote the variance-covariance matrix of $\hat{\beta}$ in Eq. (4.19) as $\text{Var}(\hat{\beta})$. We derive the following expression for $\text{Var}(\hat{\beta})$:

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \text{E}\left((\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right) \\
&= \text{E}\left((X'X)^{-1}X'(X\beta+\varepsilon)-\beta\right)\left((X'X)^{-1}X'(X\beta+\varepsilon)-\beta\right)' \\
&= \text{E}\left((X'X)^{-1}X'\varepsilon\right)\left((X'X)^{-1}X'\varepsilon\right)' \\
&= \text{E}\left((X'X)^{-1}X'\varepsilon\right)\left(\varepsilon'X(X'X)^{-1}\right)
\end{aligned}$$

or

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{4.22}$$

because of assumption 2. Since σ^2 is unknown it is usually replaced by an unbiased estimator:

$$\hat{\sigma}^2 = \frac{1}{T-K-1}\sum_{t=1}^{T} e_t^2. \tag{4.23}$$

$\hat{\sigma}$ is also known as the *standard error of the regression* or the *standard error of estimate*.

We have seen that the OLS estimate of β has desirable properties if the assumptions are satisfied. In fact, it can be shown that in this case, OLS is the optimal estimation technique. However, when the assumptions are violated, the statistical inferences in this section and in the next sections are invalid. In Chap. 5 we discuss possible reasons for violations of each of the assumptions, the consequence for parameter estimation, how each violation can be detected, and the available remedies.

4.3 Statistical Inference

If the assumptions for the error-term and other model assumptions are satisfied, we can evaluate the quality of the model, and identify substantive implications. In this section we discuss criteria pertaining to the model's goodness of fit and the significance of the coefficients.

4.3.1 Goodness of Fit

An important measure for assessing the quality of a model is the extent to which fluctuations in the criterion variable are explained by the model. This is known as the "fit" of the model. A criterion for fit is the coefficient of determination or R^2. This measure is also known as the squared multiple correlation coefficient. It measures the proportion of total variance in the criterion variable "explained" by the model. With (4.13) as the model, let

$$\hat{y}_t = \hat{\alpha} + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \cdots + \hat{\beta}_K x_{Kt} \tag{4.24}$$

and

$$e_t = y_t - \hat{y}_t.$$

R^2, the coefficient of determination is defined as[2]:

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\text{explained variation in } y}{\text{total variation in } y}. \tag{4.25}$$

The explained variation in y is also known as the explained sum of squares or the regression sum of squares. The total sum of squares in y is referred to as the total variation in y. Expression (4.25) can also be written as:

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = 1 - \frac{\text{unexplained variation in } y}{\text{total variation in } y}. \tag{4.26}$$

The unexplained variation is also called residual variation, or residual sum of squares.

Expression (4.26) can be written in matrix notation as:

$$R^2 = 1 - \frac{e'e}{y^{*\prime}y^*} \tag{4.27}$$

where

$$y^* = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_T - \bar{y} \end{pmatrix}.$$

We emphasize that R^2 is a relative measure. Its value depends on [see (4.26) or (4.27)]:

1. how well the regression line fits the data as measured by variation in the residuals, and
2. the amount of dispersion in the values of the criterion variable.

It is tempting for researchers to regard estimated equations with high R^2 values favorably and low R^2 values unfavorably. Indeed, it is straightforward to agree with

[2]If the model does not contain a constant term it is not meaningful to express the observed values in deviations from the mean, in which case the denominator should read $\sum_{t=1}^{T} y_t^2$.

this notion if everything else remains the same. However, in practice, models that are structurally deficient (e.g. with implausible substantive implications) can have higher R^2 values than models with acceptable specifications.

In addition, for some types of problems and data, all models necessarily have low R^2 values. Thus, low R^2 values should not be interpreted as indicating unacceptable or useless results nor should high R^2 values be interpreted to mean that the results are useful.

The use of many predictor variables may result in artificially high R^2 values. Each predictor with a nonzero slope coefficient makes a positive contribution to R^2, with the actual contribution determined by the slope and the amount of variation for the predictor in the sample. Model comparisons based on R^2 values are not meaningful unless the comparisons are made on one set of data for the criterion variable (and the number of predictors is taken into account).[3]

The artificiality of R^2 is further illustrated by examples that indicate how researchers can manipulate the value of R^2 while the standard deviation of the residuals remains constant (see Wittink 1988, pp. 209–213). The *standard deviation of residuals* is an absolute measure of lack of fit. It is measured in the same units as the criterion variable, and therefore it has substantive relevance.

If the criterion variable differs between two equations, either in terms of its definition or in terms of the data, comparisons based on R^2 values are *meaningless*. Models of the same criterion variable on the same data, that differ in the number of predictor variables, can be compared on R_a^2, the adjusted coefficient of determination that punishes models with many independent variables. In this coefficient both the unexplained variation and the total variation are adjusted for degrees of freedom, $(T-K-1)$ and $(T-1)$ respectively. The adjusted coefficient of determination is defined as:

$$R_a^2 = 1 - \frac{e'e/(T-K-1)}{y^{*\prime}y^*/(T-1)}. \tag{4.28}$$

From (4.27) and (4.28) we can show that R^2 and R_a^2 are related as follows:

$$R_a^2 = R^2 - \left[\frac{K}{T-K-1}\right](1-R^2). \tag{4.29}$$

It follows that $R_a^2 < R^2$ (except for the irrelevant case of $K=0$).

To summarize, the coefficient of determination, R^2, is the percent of variation in the criterion variable, which exists in a *sample* of data, that is accounted for or explained by the estimated equation. It can be shown that this measure overestimates the true explanatory power of the model. A better estimate of the true explanatory power of this model is provided by the adjusted coefficient of determination R_a^2. In addition, for comparisons of alternative models, applied to the same criterion

[3] See also Wooldridge (2012, p. 192).

variable and the same data, R_a^2 is a more useful basis. Alternatively, statistical tests of the difference between alternative models are based on the difference in (unadjusted) R^2 values and the difference in degrees of freedom.

We also note that aggregation tends to produce "inflated" R^2 values. Thus, models of monthly scanner data typically have higher R^2 values than models of weekly or daily scanner data. Similarly, models of purchases aggregated across households will have higher R^2 values than models of individual household purchase data. If high R^2 values are desired, the logical conclusion is to use aggregated data. However, aggregation almost always causes parameter estimates to be biased. Thus, even though R^2 values have some role to play in the model-building process, we also need other criteria to determine the substantive usefulness of a model.

To illustrate the difficulty associated with the use of relative measures for substantive questions, we consider the following. As indicated, R^2 measures the percent of variation in the criterion variable explained by the model in the sample. A related measure determines the marginal contribution of each predictor variable to the model's explanatory power. In a simplified setting, imagine that the predictor variables in (4.13) are uncorrelated. In that case:

$$R^2 = \sum_{k=1}^{K} r_{y,x_k}^2 \tag{4.30}$$

where

$$r_{y,x_k} = \text{the simple correlation coefficient of } y \text{ and } x_k.$$

For uncorrelated predictor variables, we can also show that:

$$r_{y,x_k}^2 = \hat{\beta}_k^2 \frac{s_{x_k}^2}{s_y^2} \tag{4.31}$$

where

$$s_{x_k} = \text{the standard deviation in the sample for } x_k,$$

$$s_y = \text{the standard deviation in the sample for } y.$$

Thus, (4.30) provides a measure of the contribution for each predictor variable to the model's explanatory power. Some software packages automatically provide *standardized* regression coefficients, defined as:

$$\text{beta}_k = \hat{\beta}_k \frac{s_{x_k}}{s_y}. \tag{4.32}$$

Hence beta coefficients measure the effects not in terms of the original units but in standard deviation units. The interest in *beta coefficients* stems from the difficulty model users have in comparing slope coefficients. As indicated in Sect. 2.3, each slope coefficient, $\hat{\beta}_k$, is interpretable only with respect to the unit in which predictor

variable k is measured. With different units of measurement for the predictor variables, the slope coefficients are not comparable. The beta coefficients are unitless, as are the correlation coefficients.

A comparison of (4.32) with (4.30) and (4.31) makes it clear that the beta coefficients have similar properties as R^2. Thus, the shortcomings associated with R^2 with respect to substantive interpretations also apply to the beta coefficients. In fact, if the model user wants to discuss the "relative importance" of predictor variables, based on the beta coefficients, several conditions must be met. Since the amount of sample variation in x_k plays a role in (4.32), this variation must be obtained through *probabilistic processes*. This means that the "importances" of predictor variables manipulated by managers, such as price and advertising, cannot be meaningfully assessed in this manner.

To demonstrate the problem for predictor variables manipulated by managers, consider the demand for a brand as a function of price and advertising. For these variables, the sample variances are determined by management based on, say, market characteristics. Suppose that two brand managers operate with similar products in two different geographic areas. Over time, one manager varies price a lot but keeps advertising at a fairly constant level. The other manager varies advertising a lot, but holds price approximately constant. Assume that in both cases, each variable varies sufficiently for the marginal effects of both predictor variables to be reliably estimable. But since the units of measurement differ between the predictors, one might be tempted to use beta coefficients. Since the standard deviation for the price variable will be large, but for the advertising variable small, the first manager is likely to conclude that price is "most important". Conversely, since the standard deviation for the advertising variable will be large but for the price variable small, the second manager may conclude that advertising is "most important". This example makes it clear why economists do not rely on beta coefficients for policy questions, and neither should marketing modelers.

On the other hand, household income, demographics, and attitudinal variables have acceptable properties for the use of relative measures. If the sample data *have been generated via probability sampling* it is *meaningful to use beta coefficients* for substantive conclusions about the relative importances of such variables in explaining household behavior. Thus, only if all predictor variables have "natural" amounts of variation, and the data represent a probability sample, is it meaningful to use the beta coefficients to infer the relative importances of the predictor variables in a population.

4.3.2 Assessing Statistical Significance

1. Test of the Equation as a Whole

As the number of predictor variables included in a model increases, the probability that at least one of the slope coefficients is statistically significant increases, even if

none of the predictor variables *truly* relates to the criterion variable. For that reason, we should *first* determine whether the equation as a whole is better than what could be due to chance variation.

The test statistic for this uses the explanatory power of the estimated equation. The null hypothesis is that the model lacks explanatory power. More specifically, all slope parameters are zero under the null hypothesis, i.e.:

$$\text{Null hypothesis } (H_0): \beta_1 = \cdots = \beta_k = \cdots = \beta_K = 0$$

$$\text{Alternative hypothesis } (H_A): \text{at least one } \beta_k \text{ is different from zero } (k = 1, \ldots, K).$$

If the error term assumptions stated earlier hold, then under the null hypothesis the explained variance should be equal to the unexplained variance. The ratio of these two variances follows an F-distribution with K degrees of freedom for the numerator and $T - K - 1$ degrees of freedom for the denominator :

$$\frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2/(K)}{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2/(T-K-1)} = \frac{\text{Explained variance}}{\text{Unexplained variance}} \sim F_{(K,T-K-1)}. \tag{4.33}$$

Since R^2 is the proportion of total variation in y that is explained by the model and $(1 - R^2)$ is the proportion that is unexplained, it is easy to show that an alternative expression for this F-ratio is:

$$\frac{R^2/(K)}{(1-R^2)/(T-K-1)} \sim F_{(K,T-K-1)}. \tag{4.34}$$

In both expressions (4.33) and (4.34), the number of slope parameters in the model equals K which is also the number of degrees of freedom used to explain variation in the criterion variable. The number of degrees of freedom left for the unexplained variation is $(T - K - 1)$, since one additional degree of freedom is used by the intercept. If the calculated F-value exceeds the critical value for a given significance level, we reject H_0. Only if the equation as a whole explains more than what could be due to chance it makes sense to investigate the statistical significance of individual slope coefficients.

Strictly speaking the F-test is valid only if all error-term assumptions are met. In practice, it is very difficult to specify a model that is complete and has no identifiable shortcomings. Typically, the residuals that remain after the model is first estimated will be reviewed and tested. Based on this residual analysis the model will be adjusted, and a second model will be estimated. This process often proceeds until the model builder has a specification that is theoretically acceptable and is consistent with the data. Such an iterative model-building process, while necessary, reduces

the validity of statistical inferences. For this reason, and also for other reasons, it is important for the model builder to "test" the ultimate specification on a validation sample (see Chap. 5).

2. Tests for (Individual) Slope Parameters

If the null hypothesis that all slope parameters are zero is rejected, we can conduct statistical tests for individual response parameters in the model. Assuming, again, that the four assumptions about the error term are valid (Sect. 4.2.3), we can evaluate the significance of each of the estimates $\hat{\beta}_k$, $(k = 1, \ldots, K)$. Under these conditions, for a model linear in the parameters and with normally distributed disturbance terms, the OLS estimator $\hat{\beta}_k$ of β_k also follows a normal distribution. It then follows that $(\hat{\beta}_k - \beta_k)/\sigma_{\hat{\beta}_k}$ is standard normally distributed. The standard deviation $\sigma_{\hat{\beta}_k}$ is a function of σ, the standard deviation of the disturbance term. Since the latter has to be estimated from the residuals, we replace $\sigma_{\hat{\beta}_k}$ by its estimate $\hat{\sigma}_{\hat{\beta}_k}$ [Eqs. (4.22) and (4.23)], so that the test statistic:

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \tag{4.35}$$

is t-distributed with $(T - K - 1)$ degrees of freedom.[4] Usually the null hypothesis is formulated as follows:

$$H_0 : \beta_k = 0$$

and H_0 is rejected in favor of the *alternative hypothesis* if the calculated t-value for a given coefficient is more extreme than can be expected, given H_0. The alternative hypothesis (H_1) that corresponds to H_0 above is:

$$H_1 : \beta_k \neq 0$$

which is a two-sided alternative hypothesis.[5]

The decision whether or not to reject H_0 can be based on so-called "critical values" or on p-values. Both approaches lead to the same decision.

When the "critical value"-approach is used, a *significance level* should be determined a priori. A significance level is the maximum probability of rejecting H_0 when it is in fact true. A 5 % significance level is a well-accepted and popular choice. Thus, we are willing to accept a probability of up to 5 % that we mistakenly reject H_0 when it is true. The decision rule to reject H_0 is based on the test statistic

[4] For more detail see introductory statistics and econometrics textbooks, such as Wittink (1988); Greene (2012); Wooldridge (2012). Tables of the t-distribution are also reproduced in most of these textbooks.

[5] It is also possible to formulate one-sided tests for β_k, but here we limit ourselves to the most common situation of a two-sided test.

in (4.35): if its value is more extreme (i.e. either too low or too high) than can be expected under H_0, it is unlikely that H_0 is true. To determine cut-off values for the test statistic in (4.35), we divide the significance level of 5 % evenly over the two tails of the t-distribution, and we reject H_0 if:

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} < t_{.025}^{T-K-1} \quad \text{or if} \quad \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} > t_{.975}^{T-K-1}$$

where $t_{.025}^{T-K-1}$ and $t_{.975}^{T-K-1}$ are the 2.5th percentile and the 97.5 percentile, respectively, in a t-distribution with $T - K - 1$ degrees of freedom. Because the t-distribution is symmetrical, we have that $t_{.025}^{T-K-1} = -t_{.975}^{T-K-1}$, so that we can write the decision rule as:

$$\left| \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \right| > t_{.975}^{T-K-1}$$

and $t_{.975}^{T-K-1}$ is called the "critical value".

Most statistical software programs provide p-values, which can also be used for deciding between H_0 and H_1. A p-value is the probability of rejecting H_0 when in fact it is true. Alternatively, a p-value can also be interpreted as: 'the smallest significance level at which the null hypothesis is rejected'. Given that a common and quite well-accepted choice for the significance level is 5 %, the decision rule in the "p-value-approach" is to reject H_0 if:

$$p < 0.05. \tag{4.36}$$

In most cases, the p-value is reported for a two-sided test for β_k.

Whichever of the two approaches we use, when H_0 is not rejected this is expressed as: "we fail to reject H_0 at the 5 % level" rather than: "H_0 is accepted at the 5 % level".

As the amount of empirical evidence with regard to demand sensitivity to specific marketing instruments accumulates, it seems inappropriate to continue the use of hypotheses of no effect. Farley et al. (1995) argue that model builders should instead rely on average estimated effects from previous empirical analyses. Such average effects have been reported in various meta-analyses (see, for example, Hanssens 2009). New empirical results for problems studied earlier should then be tested against prevailing average effects. Suppose that the current average effect for predictor variable x_k is c. Then the null hypothesis is:

$$H_0 : \beta_k = c.$$

If we cannot reject this null hypothesis, the product/time/region for which data have been obtained is comparable to previous sources with regard to the average effect of variable x_k. If we obtain an effect that is significantly different from the

prevailing standard, we have an opportunity to consider the reason(s) for divergence. Farley, Lehmann, Sawyer argue that this manner of testing will prove to be more informative than the traditional procedure.

There are many possible reasons why the t-ratio for a given slope coefficient can be insignificant:

1. the predictor variable has an effect that is different from the functional form assumed (incorrect functional form);
2. the model excludes other relevant predictor variables (omitted variables);
3. the predictor variable is highly correlated with one or more other predictor variables included in the model (multicollinearity);
4. the number of data points in the sample is insufficient (lack of power);
5. the predictor variable has no relation with the criterion variable (irrelevance).

Insignificance due to either of the first two reasons should stimulate us to investigate the existence of superior functional forms and/or additional predictor variables. The third reason requires the exploration of additional data sources, model reformulation, or alternative estimation procedures (as might be the case for the fourth reason). Only the fifth reason is a proper justification for eliminating a predictor variable from the model. We must consider each of these possible reasons before we eliminate a predictor variable from the model.

We note that in a regression analysis with one predictor variable, the F-test in (4.33) and the t-test based on (4.35) for testing $\beta_k = 0$ provide the same conclusion.[6] Curiously, in a multiple regression analysis it is possible to reject the null hypothesis that all slope parameters are zero and at the same time find that none of the slope coefficients differ significantly from zero based on the t-test. This may occur if the predictor variables are highly intercorrelated (multicollinearity). In this case the estimates for separate influences of $x_k, k = 1, \ldots, K$ on y are unreliable (i.e., they have large estimated standard errors), while the model's explanatory power may be high. Marketing data, especially at high aggregation levels, often have a high degree of collinearity between the predictor variables. We pursue the issue of multicollinearity in more detail in Chap. 5.

In some cases it is of interest to test a hypothesis that involves more than one of the slope parameters $\beta_1, \cdots, \beta_K$. Assume for example that a company can invest in advertisements in print (x_1) and in commercials on television (x_2) and that management wants to know whether or not the two media are equally effective (i.e. are the two associated parameters, β_1 and β_2, similar or significantly different). To this end one may use a t-test statistic such as (4.37):

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_1 - \hat{\beta}_2}} \tag{4.37}$$

[6] This is true if both tests are conducted at the same level of significance and against the same alternative (i.e. a two-sided t-test).

where $\hat{\sigma}_{\hat{\beta}_1-\hat{\beta}_2}$ is the standard error of the difference between $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\hat{\sigma}_{\hat{\beta}_1-\hat{\beta}_2} = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1} + \hat{\sigma}^2_{\hat{\beta}_2} + \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2}} \tag{4.38}$$

where $\hat{\sigma}_{\hat{\beta}_1\hat{\beta}_2}$ denotes an estimate of $\text{cov}(\hat{\beta}_1,\hat{\beta}_2)$.

4.4 Numerically Specified Models for the Verhouten Case

In this section, we estimate the two models that we specified in Sect. 2.6 for the Verhouten case. First we estimate the linear additive model that we presented in Eq. (2.26), thereafter we estimate the multiplicative model of Eq. (2.27). For both models we present:

1. OLS estimates of the parameters;
2. standard errors of the estimated parameters;
3. t-values of the estimated parameters;
4. p-values of the estimated parameters;
5. the number of observations;
6. R-squared and adjusted R-squared;
7. the standard error of the estimate (the standard deviation of the residuals, $\hat{\sigma}$);
8. the Residual Sum of Squares (RSS);
9. the Explained Sum of Squares (ESS);
10. the F-value for testing the equation as a whole, see Eq. (4.33).

Table 4.1 displays the estimation outcomes for the *linear additive model* for Verhouten. Note that an estimate for β_{42}, the parameter of Feature and Display for Droste, is lacking. This is due to unavailability of this variable in the data set.

In order to be able to estimate the *multiplicative model* with OLS, we first need to linearize Eq. (2.27) using the logarithmic transformation:

$$\begin{aligned}
\ln(S_{1t}) &= \ln\left(\theta\left[\prod_{j=1}^{n}\left(P_{jt}^{\gamma_{1j}}\gamma_{2j}^{F_{jt}}\gamma_{3j}^{D_{jt}}\gamma_{4j}^{FD_{jt}}\right)\right]Temp_t^{*\gamma_5}\gamma_6^{\delta_{1t}}\gamma_7^{\delta_{2t}}\varepsilon_{1t}\right) \\
&= \ln(\theta) + \sum_{j=1}^{n}\left(\gamma_{1j}\ln(P_{jt}) + \ln(\gamma_{2j})F_{jt} + \ln(\gamma_{3j})D_{jt} + \ln(\gamma_{4j})FD_{jt}\right) \\
&\quad + \gamma_5\ln(Temp_t^*) + \ln(\gamma_6)\delta_{1t} + \ln(\gamma_7)\delta_{2t} + \ln(\varepsilon_{1t})
\end{aligned}$$

Table 4.1 Estimation results of the linear additive model for Verhouten

Equation (2.26)	Parameter estimates	Standard error	t-value	p-value
Intercept ($\hat{\alpha}$)	1331.83	139.78	9.53	0.00
Price Verhouten ($\hat{\beta}_{11}$)	−756.17	51.76	−14.61	0.00
Price Droste ($\hat{\beta}_{12}$)	−52.07	68.25	−0.76	0.45
Price Baronie ($\hat{\beta}_{13}$)	5.96	34.71	0.17	0.86
Price Delicata ($\hat{\beta}_{14}$)	−2.11	28.02	−0.08	0.94
Feature-only Verhouten ($\hat{\beta}_{21}$)	−50.50	35.04	−1.44	0.16
Feature-only Droste ($\hat{\beta}_{22}$)	98.64	67.30	1.47	0.15
Feature-only Baronie ($\hat{\beta}_{23}$)	17.43	22.43	0.78	0.44
Feature-only Delicata ($\hat{\beta}_{24}$)	48.39	38.29	1.26	0.21
Display-only Verhouten ($\hat{\beta}_{31}$)	22.22	93.76	0.24	0.81
Display-only Droste ($\hat{\beta}_{32}$)	−1.58	46.52	−0.03	0.97
Display-only Baronie ($\hat{\beta}_{33}$)	−148.40	89.81	−1.65	0.10
Display-only Delicata ($\hat{\beta}_{34}$)	86.93	96.25	0.90	0.37
Feature and Display Verhouten ($\hat{\beta}_{41}$)	100.66	167.56	0.60	0.55
Feature and Display Baronie ($\hat{\beta}_{43}$)	109.03	113.43	0.96	0.34
Feature and Display Delicata ($\hat{\beta}_{44}$)	−4.82	45.48	−0.11	0.92
Temperature ($\hat{\beta}_5$)	−2.28	0.77	−2.96	0.00
Dummy variable December ($\hat{\beta}_6$)	−51.76	25.10	−2.06	0.04
Dummy variable Easter ($\hat{\beta}_7$)	−48.86	48.98	−1.00	0.32
Number of observations = 68, $R^2 = 0.93$, $R_a^2 = 0.91$, $\hat{\sigma} = 30.53$				
RSS = 45685.06, ESS = 631560.45, F-value = 37.63				

in which we recognize a linear additive function if we redefine some of the variables and some of the parameters:

$$\begin{aligned} S_{1t}^* = \theta^* + \sum_{j=1}^{n} \left(\gamma_{1j} P_{jt}^* + \gamma_{2j}^* F_{jt} + \gamma_{3j}^* D_{jt} + \gamma_{4j}^* FD_{jt} \right) \\ + \gamma_5 Temp_t^{**} + \gamma_6^* \delta_{1t} + \gamma_7^* \delta_{2t} + \varepsilon_{1t}^* \end{aligned} \tag{4.39}$$

where, for $j = 1, \ldots, n$:

$$\begin{aligned} S_{1t}^* &= \ln(S_{1t}), \\ \theta^* &= \ln(\theta), \\ P_{jt}^* &= \ln(P_{jt}), \\ \gamma_{2j}^* &= \ln(\gamma_{2j}), \\ \gamma_{3j}^* &= \ln(\gamma_{3j}), \\ \gamma_{4j}^* &= \ln(\gamma_{4j}), \\ Temp_t^{**} &= \ln(Temp_t^*), \end{aligned} \tag{4.40}$$

$$\gamma_6^* = \ln(\gamma_6),$$
$$\gamma_7^* = \ln(\gamma_7),$$
$$\varepsilon_{1t}^* = \ln(\varepsilon_{1t}).$$

In Table 4.2 we present the estimation results for (4.39).

Table 4.2 Estimation results of the multiplicative model for Verhouten

Equation (4.39)	Parameter estimates	Standard error	t-value	p-value
Intercept ($\hat{\theta}^*$)	20.97	6.15	3.41	0.00
ln of Price Verhouten ($\hat{\gamma}_{11}$)	−3.87	0.32	−11.99	0.00
ln of Price Droste ($\hat{\gamma}_{12}$)	−1.24	0.51	−2.45	0.02
ln of Price Baronie ($\hat{\gamma}_{13}$)	0.42	0.21	1.98	0.05
ln of Price Delicata ($\hat{\gamma}_{14}$)	−0.02	0.24	−0.08	0.94
Feature-only Verhouten ($\hat{\gamma}_{21}^*$)	0.04	0.17	0.23	0.82
Feature-only Droste ($\hat{\gamma}_{22}^*$)	0.10	0.34	0.29	0.78
Feature-only Baronie ($\hat{\gamma}_{23}^*$)	0.23	0.11	2.10	0.04
Feature-only Delicata ($\hat{\gamma}_{24}^*$)	0.04	0.19	0.19	0.85
Display-only Verhouten ($\hat{\gamma}_{31}^*$)	0.35	0.47	0.74	0.46
Display-only Droste ($\hat{\gamma}_{32}^*$)	−0.24	0.23	−1.01	0.32
Display-only Baronie ($\hat{\gamma}_{33}^*$)	−0.25	0.45	−0.55	0.59
Display-only Delicata ($\hat{\gamma}_{34}^*$)	0.50	0.48	1.03	0.31
Feature and Display Verhouten ($\hat{\gamma}_{41}^*$)	0.52	0.84	0.62	0.54
Feature and Display Baronie ($\hat{\gamma}_{43}^*$)	0.21	0.57	0.36	0.72
Feature and Display Delicata ($\hat{\gamma}_{44}^*$)	−0.03	0.23	−0.12	0.91
ln of Temperature in Kelvin ($\hat{\gamma}_5$)	−2.57	1.09	−2.36	0.02
Dummy variable December ($\hat{\gamma}_6^*$)	−0.35	0.13	−2.82	0.01
Dummy variable Easter ($\hat{\gamma}_6^*$)	−0.15	0.25	−0.60	0.55
Number of observations = 68,	$R^2 = 0.93$,	$R_a^2 = 0.90$,	$\hat{\sigma} = 0.15$	
RSS = 1.15, ESS = 14.25,	F-value = 33.78			

In Tables 4.1 and 4.2 we report both standard errors and t-values. Some authors prefer to report t-statistics rather than standard errors. There is some preference for reporting standard errors, because this makes it easy to compute confidence intervals (Wooldridge 2012, p. 146).

From Tables 4.1 and 4.2 we observe that:

- The goodness of fit of both models indicates that about 93 % of the fluctuations of the sales of Verhouten can be explained by the models. Relatively high values for R^2 were expected in this application, given that we deal with time series data. The adjusted coefficient of determination, R_a^2, is 0.91 for the linear additive model, and 0.90 for the multiplicative model. The adjustments relative to R^2 are due to the large number of independent variables in Eqs. (2.26) and (2.27). The value of the F-statistic (4.33) is in both cases substantially larger than 1.819,

which is the critical value of an F distribution with $K = 18$ degrees of freedom for the numerator and $T - K - 1 = 68 - 18 - 1 = 49$ degrees of freedom for the denominator, assuming a significance level of 5 %. This indicates that both models are highly significant.

- The number of significant parameters is not the same in both models. In the linear additive model, the intercept (α) is significant, as well as the parameter for price of Verhouten (β_{11}), the parameter for temperature (β_5) and the parameter for the December dummy (β_6). These parameters are significant because the corresponding p-values are smaller than 0.05. The same parameters are also significant in the multiplicative model, but now the parameters for the ln of price of Droste (γ_{12}), the ln of price of Baronie (γ_{13}), and for feature-only for Baronie (γ_{23}) are also significant.
- Following Farley et al. (1995)'s suggestion we compare the significant price estimates to earlier generalizations in literature. Bijmolt et al. (2005) find that the average price-to-sales elasticity at the brand level equals −2.62. Our estimate for γ_{11} indicates a somewhat stronger price response (−3.87), indicating that the market for chocolate might be somewhat more price sensitive than average. From earlier studies we know that cross-price-to-sales elasticities are asymmetric (Blattberg and Wisniewski 1989): the cross-price elasticity of lower-priced brands on higher-priced brands' sales is not as large as the cross-price elasticity of higher-priced brands on lower-priced brands' sales. The Verhouten brand has the lowest price, followed by Droste, Baronie, and Delicata respectively. In their meta-analysis on cross-price elasticities, Sethuraman et al. (1999) estimate that the average cross-price elasticity of a brand on its closest lower-priced neighbor is 0.754. If we compare this to −1.24, the estimate for γ_{12}, we find that the sign is opposite of what we expect and also the magnitude is unexpectedly high. Apparently, Verhouten's sales benefit from Droste's price promotions. Sethuraman et al. (1999) estimate that the average cross-price elasticity of a brand on its second-closest lower-priced neighbor is 0.344. This compares favorably to our estimate of 0.42 for γ_{13}.
- RSS, ESS, and $\hat{\sigma}$ are not comparable in size across the models. This is due to the fact that several variables, including the dependent variable, were ln-transformed in order to be able to estimate the multiplicative model with OLS. As a consequence, one can also not compare the estimated values for the parameters across both models.

Note that the numbers in Table 4.2 are estimates for the parameters in Eq. (4.39), the linearized version of the multiplicative model. Some of the parameters in this model correspond directly to the parameters of Eq. (2.27), the original multiplicative model. In our application these are the four price-coefficients and the temperature coefficient. Other parameters required a ln-transformation to obtain linearity in parameters [see Eq. (4.40)]; these are superindexed with an asterisk in Table 4.2. If we are interested in estimates for the parameters in Eq. (2.27), we need to apply an 'anti-ln' transformation to all estimated values for the parameters that are superscripted with an asterisk in Table 4.2. It appears sensible to take the exponential of these values to obtain proper estimates for the parameters in the multiplicative

model. However, Goldberger (1968) showed that this approach produces biased results. Wittink et al. (2011) propose a correction to reduce this bias that relates an estimate for a parameter γ in the multiplicative model to an estimate for the ln-transformed parameter γ^* as follows:

$$\hat{\gamma} = \exp(\hat{\gamma}^*) \times \exp\left(-\frac{1}{2}\hat{\sigma}^2_{\hat{\gamma}^*}\right), \tag{4.41}$$

where $\hat{\sigma}_{\hat{\gamma}^*}$ is the estimated standard error for $\hat{\gamma}^*$. So, for example, an estimate for $\hat{\gamma}_{21}$, the coefficient of Feature-only of Verhouten, equals $\exp(0.04) \times \exp(-\frac{1}{2} \times (0.17)^2) = 1.03$.[7]

4.5 Estimating Pooled Models

4.5.1 Introduction

In Sect. 2.7, we discussed several approaches for treating heterogeneity across entities or cross-sections. In this section, we discuss how to estimate such models. Following our line of discussion in Sect. 2.7, we subsequently treat unit-by-unit models, pooled models, and partially pooled models. We do not consider aggregate models, as they do not require pooling.

We explain estimation of pooled models using the simple model for Verhouten that was discussed in Sect. 2.7, where we assumed that there are only two brands in the market and that a useful model for explaining brand sales included only own price. For estimation purposes, we assume that we have four $(T \times 1)$ vectors of data available:

$$S_1 = \begin{pmatrix} S_{11} \\ S_{12} \\ \vdots \\ S_{1T} \end{pmatrix}, \quad S_2 = \begin{pmatrix} S_{21} \\ S_{22} \\ \vdots \\ S_{2T} \end{pmatrix}, \quad P_1 = \begin{pmatrix} P_{11} \\ P_{12} \\ \vdots \\ P_{1T} \end{pmatrix}, \quad P_2 = \begin{pmatrix} P_{21} \\ P_{22} \\ \vdots \\ P_{2T} \end{pmatrix}, \tag{4.43}$$

using notation that was introduced in Sect. 2.6.

[7] A similar correction is needed when predicting values for the dependent variable, based on predicted values for the ln-transformed version of that variable. For example, predictions for S_{1t} are related to predictions for S^*_{1t} as follows:

$$\hat{S}_{1t} = \exp(\hat{S}^*_{1t}) \times \exp\left(\frac{1}{2}\hat{\sigma}^2\right), \tag{4.42}$$

where $\hat{\sigma}^2$ is an estimate for the variance of the residuals, see Eq. (4.23).

4.5.2 Estimating Unit-by-Unit Models

The unit-by-unit approach was introduced in Sect. 2.7, where we explained that this involves estimating a separate model for each entity or cross-section. The top panel of Fig. 2.5 shows the corresponding model specification:

$$S_{it} = \alpha_i + \beta_i P_{it} + \varepsilon_{it}\,, \quad t = 1,\ldots,T, \quad i = 1,2. \tag{4.44}$$

Estimation of Eq. (4.44) is quite straightforward: it involves running a separate regression for each of the two cross-sections (brands) in our example. In the first regression, we regress S_1 on P_1, and in the second regression we regress S_2 on P_2. This generates a separate set of parameter estimates for each of the two cross sections, and provides maximum flexibility for accommodating all potential differences between the brands. As illustrated in the top panel of Fig. 2.5, it allows for different intercepts, and different response parameters. However, in some cases this approach is not feasible, for example when T is too small to reliably estimate all the coefficients for each cross-section separately.

4.5.3 Estimating Fully Pooled Models

The middle panel of Fig. 2.5 depicts a situation where a fully pooled model is appropriate. In that case it is assumed that all parameters are the same across cross sections. The corresponding model specification for this approach is:

$$S_{it} = \alpha + \beta P_{it} + \varepsilon_{it}\,, \quad t = 1,\ldots,T, \quad i = 1,2. \tag{4.45}$$

Estimation of Eq. (4.45) requires that we first stack the sales vectors and the price vectors as follows:

$$S = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} = \begin{pmatrix} S_{11} \\ S_{12} \\ \vdots \\ S_{1T} \\ S_{21} \\ S_{22} \\ \vdots \\ S_{2T} \end{pmatrix}, \quad \text{and} \quad P = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \begin{pmatrix} P_{11} \\ P_{12} \\ \vdots \\ P_{1T} \\ P_{21} \\ P_{22} \\ \vdots \\ P_{2T} \end{pmatrix}. \tag{4.46}$$

Subsequently, S is regressed on P. This results in a combined estimate for α and a combined estimate for β. Equation (4.46) illustrates the most important benefit of pooling: by combining data from cross-sections, there are more data points

available so that the coefficients can be estimated with greater statistical efficiency (i.e. smaller variance of the estimates). In general, when data are pooled across n cross sections, and for each cross section T observations were recorded, the length of the data vectors becomes $n \times T$.

Baltagi (2008, pp. 6–8) mentions other benefits of pooled models:

1. pooled models offer opportunities to control for heterogeneity (e.g. by using partial pooling, see below);
2. pooled models give *more* data, more *variability*, less collinearity among "independent" variables, more degrees of freedom and more efficiency;
3. pooled models are better able to identify and measure effects that are simply not detectable in pure cross-section and pure time-series models.

If a pooled model is applied while (some of) the parameters are not similar, then the estimates based on the pooled model (4.45) lack meaning (at best the estimates represent weighted averages of the separate parameter values). Nevertheless, if the differences in parameters between the cross sections are small we may be willing to accept some bias in return for smaller variances of the estimated coefficients.[8] Consequently, even if there are differences between the sets of parameters across the cross sections, pooled models may be preferred over unit-by-unit estimation, because the statistical uncertainty due to separate regressions on small data sets may "inflate" the differences considerably. Thus, the model builder faces the question how to balance bias (unit-by-unit estimation minimizes bias) against variance (pooled estimation minimizes the variance). In Sect. 5.4 we discuss how to test whether pooling is appropriate for a given data set.

4.5.4 Estimating Partially Pooled Models

Partial pooling is appropriate when some coefficients are quite similar across cross-sections, while others are not. The bottom panel of Fig. 2.5 depicts a situation where the price response parameters are about the same, but the intercepts differ between the two brands. The corresponding model specification for this approach is:

$$S_{it} = \alpha_i + \beta P_{it} + \varepsilon_{it}, \quad t = 1, \ldots, T, \quad i = 1, 2. \tag{4.47}$$

Partial pooling allows a model builder to strike a balance between the benefits of pooling (increased statistical efficiency) and accommodating heterogeneity between the cross sections.

Estimation of Eq. (4.47) requires stacking of the dependent variable across cross-sections as in the fully pooled model. Similarly, the variables whose coefficients we are planning to pool across cross sections are also stacked. However, those variables

[8] To this end Wallace (1972) suggested weaker criteria. See also Brobst and Gates (1977).

that require cross-section specific estimates are not stacked. Since we want to obtain a different estimate for each cross section for each of these variables, we retain different variables for the different cross sections. However, in order to make sure that we can estimate (4.47), all data vectors must be of equal length. To that end, the remaining elements are filled up with zeros.

Let us illustrate this for the simplified Verhouten case that we consider in this section. Estimation of (4.47) uses the stacked vectors S and P, because we aim to obtain a pooled price coefficient. However, we do not stack the part that is required for estimating the constant. In the first column of the X matrix of Eq. (4.14), we see that this is a column of ones. Hence, in order to obtain separate estimates for α_1 and α_2, we define two variables δ_1 and δ_2 as follows:

$$\delta_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{and} \quad \delta_2 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix}}\right\} T \text{ observations for brand 1} \\ \left.\vphantom{\begin{matrix} 1 \\ 1 \\ \vdots \\ 1 \end{matrix}}\right\} T \text{ observations for brand 2} \end{matrix} \tag{4.48}$$

Note that δ_1 and δ_2 have the same length as S and P. Subsequently, we regress S on δ_1, δ_2 and on P, where we do not estimate an overall constant, since we replaced the column of ones in Eq. (4.14) by δ_1 and δ_2. The estimated coefficient for δ_1 is our estimate for α_1, the estimated coefficient for δ_2 is the estimate for α_2, and the estimated coefficient for P is the estimate for β.

The situation in our example is a special case of partial pooling, where we only estimate cross-section specific intercepts, while all other parameters are pooled across cross sections. In that case, the variables that are created to obtain cross-section specific estimates are dummies as we saw in Eq. (4.48). Therefore, this case of partial pooling is also referred to as *OLS with Dummy Variables*, or OLSDV. This is a type of pooling that is relevant in many marketing applications where the cross sections have different overall levels of the dependent variable, but it is reasonable to assume that the response to the marketing instruments is about the same across the cross sections. This occurs for example in sales models for a specific chain that are specified at the store level. Store-specific intercepts then capture differences in store sizes, while it is reasonable to work with pooled estimates of the response parameters, because, given that they belong to the same chain, the stores are likely to attract similar customers (see Horváth et al. 2005, for an example).

OLSDV provides a method for estimating models in which a different effect (i.e. intercept) is specified for each cross section. Such models are referred to as *fixed effects* models, to distinguish them from *random effects models*, that provide another way of accommodating cross-sectional differences in the intercept.

In cases where there is only a small number of cross sections, we might be interested in determining a (fixed) effect for each cross section separately, and OLSDV can provide these estimates. Random effects models are applicable in situations where we want to account for differences in the intercepts, but where we are not interested in obtaining estimates of these effects per se. Such cases may arise in applications with a relatively large number of cross sections (e.g. stores or individual customers). Each cross-sectional effect α_i is then thought of as a drawing from a statistical distribution (hence "random effect"), e.g. $\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2)$, and the parameters of this distribution (μ_α and σ_α^2) are used to characterize the heterogeneity in the intercepts. Even with a small to moderate number of cross-sectional units, estimating the parameters of this distribution (usually only one or two) can be much more efficient than estimating a separate effect for each cross section. However, the random effects approach cannot always be applied, because technically, the cross-sectional differences are included as an additional disturbance term, and the joint disturbance term needs to satisfy the conditions that we discussed earlier. A full treatment of the resulting statistical intricacies of fixed and random effects models is beyond the scope of this monograph, and we refer to Greene (2012, Sections 11.4 and 11.5) or Wooldridge (2012, Sections 14.1 and 14.2) for a detailed discussion.

References

Amemiya, T.: Advanced Econometrics. Harvard University Press, Cambridge (1985)

Baltagi, B.H.: Econometric Analysis of Panel Data, 4th edn. Wiley, Chichester (2008)

Baltagi, B.H.: Econometrics, 5th edn. Springer, Berlin (2011)

Bijmolt, T.H.A., Van Heerde, H.J., Pieters, F.G.M.: Determinants of price elasticity: New empirical generalizations. J. Mark. Res. **42**, 141–156 (2005)

Blattberg, R.C., Wisniewski, K.J.: Price-induced patterns of competition. Mark. Sci. **8**, 291–309 (1989)

Brobst, R., Gates, R.: Comments on pooling issues and methods in regression analysis. J. Mark. Res. **14**, 598–606 (1977)

Cameron, A.C., Trivedi, P.K.: Microeconometrics Using Stata. Stata Press (2009)

Davidson, R., MacKinnon, J.G.: Econometric Theory and Methods. Oxford University Press, New York (2004)

Farley, J.U., Lehmann, D.R., Sawyer, A.: Empirical marketing generalization using meta-analysis. Mark. Sci. **14**, G36–G46 (1995)

Goldberger, A.S.: The interpretation and estimation of Cobb-Douglas functions. Econometrica **9**(3/4), 464–472 (1968)

Goldberger, A.S.: Introductory Econometrics. Harvard University Press, Cambridge (1998)

Greene, W.H.: Econometric Analysis, 7th international edn. Prentice-Hall, Upper Saddle River (2012)

Gujarati, D.N., Porter, D.: Essentials of Econometrics, 4th edn. McGraw-Hill, New York (2009)

Hamilton, J.D.: Time Series Analysis. Princeton University Press, Princeton (1994)

Hanssens, D.M.: Empirical Generalizations About Marketing Impact. Marketing Science Institute, Relevant Knowledge Series (2009)

Heij, C., De Boer, P.M.C., Franses, P.H., Kloek, T., Van Dijk, H.K.: Econometric Methods with Applications in Business and Economics. Oxford University Press, Oxford (2004)

Horváth, C., Leeflang, P.S.H., Wieringa, J.E., Wittink, D.R.: Competitive reaction- and feedback effects based on VARX models of pooled store data. Int. J. Res. Mark. **22**, 415–426 (2005)
Intriligator, M.D., Bodkin, R.G., Hsiao, C.: Econometric Models, Techniques and Applications, 2nd edn. Prentice Hall, Upper Saddle River (1996)
Johnston, J., Dinardo, J.: Econometric Methods, 4th edn. McGraw-Hill, New York (1997)
Judge, G.G., W. E. Griffiths R. C. Hill, H.L., Lee, T.C.: The Theory and Practice of Econometrics, 2nd edn. Wiley, New York (1985)
Pindyck, R.S., Rubinfeld, D.L.: Econometric Models & Economic Forecasts, 4th edn. McGraw-Hill, New York (1997)
Sethuraman, R., Srinivasan, V., Kim, D.: Asymmetric and neighborhood cross-price effects: Some empririca generalizations. Mark. Sci. **18**, 23–42 (1999)
Verbeek, M.: A Guide to Modern Econometrics, 4th edn. Wiley, Chichester (2012)
Wallace, T.D.: Weaker criteria and tests for linear restrictions in regression. Econometrica **40**, 689–698 (1972)
Wittink, D.R.: The Application of Regression Analysis. Allyn and Bacon, Boston (1988)
Wittink, D.R., Addona, M.J., Hawkes, W.J., Porter, J.C.: SCAN*PRO: The estimation, validation and use of promotional effects based on scanner data. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in Honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)
Wooldridge, J.M.: Introductory Econometrics: A Modern Approach, 5th edn. Cengage Learning, Mason (2012)

Chapter 5
Validation and Testing

5.1 Introduction

Two critical steps in the model building process are model specification and model estimation. In this chapter we turn to the next stage in model building: validation (also verification or evaluation).

In its broadest sense validation is an assessment of the quality of the model outcomes. Validation criteria for model building can relate to:

- the model structure (specification);
- the data quality;
- the estimation method;
- the applicability of statistical tests (e.g. with regard to error term assumptions);
- the correspondence of model results to theoretical and common sense expectations;
- the model's (relative) performance against alternative models;
- the relevance of model results for intended use.

In previous chapters we already covered issues related to specification, data quality and estimation. In this chapter we turn to the four remaining validation criteria. In Sect. 5.2 we discuss tests that can be used to validate the six assumptions that required for application of OLS (see Sect. 4.2.3). In Sect. 5.3 we introduce mediation tests and in Sect. 5.4 we consider several tests that allow us to judge whether assumptions are jointly satisfied. In Sect. 5.5 we discuss *face validity* criteria which are used to determine whether model results are in accordance with theoretical and/or common-sense expectations. (Literally, face validity refers to the extent to which one's face becomes red if one is questioned about the meaningfulness of the empirical results.)

We introduce criteria for model selection in Sect. 5.6. The idea of model selection is that we often have alternative model specifications, and we use data to distinguish between the alternatives. The superiority of one model over another may depend

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_5

on the product category and on competitive conditions but also on the quality of data. Even though theoretical arguments should inform the model specification, in marketing we want the empirical results to be not only consistent with what sound arguments dictate but also with how the marketplace behaves. With new data, the question is whether extant models apply, and with new models the question is whether the proposed specification outperforms prevailing benchmarks.

If the model-building effort is intended to have descriptive validity, we could restrict the validation effort to such aspects as model tests and face validity. However, in marketing, the empirical research almost always includes a measure of predictive validity. This is perhaps a reflection of the philosophy that for any model to be useful (even for descriptive purposes), it must have predictive validity. We introduce a framework, along with criteria, for predictive validity in Sect. 5.7. Finally, we illustrate the validation of models in Sect. 5.8, using data from the Verhouten case.

5.2 Testing the Six Basic Assumptions of the General Linear Model

We have argued in Sect. 4.2.3 that several assumptions concerning different model elements need to be satisfied when the parameters are estimated with OLS. We repeat here the disturbance term assumptions:

1. $\mathrm{E}(\varepsilon_t) = 0$ for all t;
2. $\mathrm{Var}(\varepsilon_t) = \sigma^2$ for all t;
3. $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t'}) = 0$ for $t \neq t'$;
4. ε_t is normally distributed.

The two other assumptions are:

5. There is no relation between the predictors and ε_t, i.e. $\mathrm{Cov}(x_t, \varepsilon_t) = 0$ (one-variable case). In other words the x_t are nonstochastic or "exogenous". For the K-variable case this implies $\mathrm{Cov}(X', \varepsilon) = 0$, in which case we have $\mathrm{E}(\varepsilon \mid X) = \mathrm{E}(\varepsilon) = 0$.
6. For the K-variable case, the matrix of observations X has full rank, that is the vectors in X are linearly independent.

In Table 5.1 we show possible reasons for violations of each of the assumptions, the consequence of this violation for parameter estimates, how each violation can be detected, and the available remedies.

Table 5.1 Violations of the assumptions about the disturbance term: reasons, consequences, tests and remedies

Violated assumption	Possible reasons	Consequence	Detection	Remedy
1. $E(\varepsilon_t) \neq 0$	- Incorrect functional form(s) - Omitted variable(s) - Varying parameter(s)	- Biased parameter estimate[a]	- Plot residuals against each predictor variable - RESET-test - White test	- Modify the model specification in terms of functional form - Add relevant predictors - Allow parameters to vary
2. $\text{Var}(\varepsilon_t) \neq \sigma^2$	- Error proportional to values of a predictor	- Inefficient parameter estimate	- Plot residuals against each predictor variable - Goldfeld/Quandt test, Breusch–Pagan test, White test	- Modify the specification (e.g. per capita) - Use heteroscedasticity-consistent estimation (e.g. WLS, GLS[b])
3. $\text{Cov}(\varepsilon_j, \varepsilon_{t'}) \neq 0$, $t \neq t'$	- See 1.	- See 2.	- Plot residuals against time - Durbin Watson test - Durbin's h-test	- See 1.
4. Nonnormal errors	- See 1.	- p-values cannot be trusted	- Plot distribution of residuals - χ^2-test - Kolmogorov–Smirnov - Shapiro Wilk - Bera–Jarque	- See 1. - Robust regression - Box–Cox-transformation
5. Stochastic predictor correlated with the disturbance term: $E(\varepsilon \mid x_i) \neq 0$	- Measurement errors - Endogeneity of predictor variable(s)	- See 1.	- Diagnose specification	- Simultaneous equations - Instrumental Variable estimation

(continued)

Table 5.1 (continued)

Violated assumption	Possible reasons	Consequence	Detection	Remedy
6. Multicollinearity	- Relations between predictor variables	- Unreliable parameter estimates	- Inspect the correlation matrix of the predictor variables - Regress x_k on the other predictor variables - some VIF values ≥ 5	- Reformulate model - Create new predictors (e.g. factor scores) - Obtain more data - Apply other estimation methods - Eliminate predictor variable(s)

[a]The bias due to an omitted relevant predictor variable depends on the correlation between included and excluded variables. If this correlation is zero, there is no bias (but the estimates are less precise)
[b]*WLS* weighted least squares, *GLS* generalized least squares, VIF variance inflation factor

5.2.1 Nonzero Expectation

Violation of the first assumption, i.e. $E(\varepsilon) \neq 0$, is the most serious one. One of the principal desiderata of parameter estimates is unbiasedness (or consistency). In regression analysis unbiasedness can only be obtained if the model is correctly specified. All relevant predictor variables should be included, the proper functional form of the partial relation with respect to each predictor must be accommodated, etc. Misspecification of the model causes the parameter estimates to be biased. For example, an omitted predictor variable causes the parameter estimates to be biased, unless the omitted variable is uncorrelated with the included predictor variables. The amount of the bias increases with the degree of (positive or negative) correlation. Because of this, it is always necessary to specify a model that is complete on relevant issues, even in cases when a researcher is interested in the effect of one specific variable (see also Sect. 2.2.2). Including irrelevant variables can have undesirable effects on the variance of the OLS estimates.

Violations of the first assumption are rarely detectable from a plot of the residuals, $e_t = (y_t - \hat{y}_t)$, $t = 1, \ldots, T$, against each predictor variable. However, if only the assumed functional form is incorrect, such a plot should show a systematic pattern in the residual values. On the other hand, this plot will not suggest that a relevant predictor has been omitted (unless one has information about the values of the omitted variable).

To test for the possibility of omitted variables, the model can be extended with additional terms. The explanatory power of these extra terms is then used as a basis for detecting the omission of relevant predictor variables. A formal statistical test

for misspecification in this context is the RESET-test.[1] The null hypothesis of the test is Eq. (4.13) and the alternative hypothesis is $E(\varepsilon_t) = \xi \neq 0$. The test is based on the estimation of an extended model:

$$\begin{aligned} y_t &= \alpha + \beta_1 x_{1t} + \cdots + \beta_K x_{Kt} + \\ &\quad + \gamma_1 z_{1t} + \cdots + \gamma_m z_{mt} + \varepsilon_t^*, \quad t = 1, \ldots, T \end{aligned} \tag{5.1}$$

where

$$\begin{aligned} z_{1t}, \ldots, z_{mt} &= \text{the additional variables chosen in such a way} \\ &\quad \text{as to explain the elements } \xi_t,\ t = 1, \ldots T \text{ of } \xi \text{ best.} \end{aligned}$$

If $E(\varepsilon) = 0$ is true for (4.13) then $\gamma_1 = \cdots = \gamma_m = 0$ in (5.1). The alternative hypothesis is accepted if $\gamma_\ell \neq 0$ for at least one ℓ, $\ell = 1, \ldots, m$. For the additional variables $z_{1t}, \ldots, z_{mt}$, Ramsey (1969) recommends adding the powers of $\hat{y}_t$ (i.e. $\hat{y}_t^2$, $\hat{y}_t^3$, $\hat{y}_t^4, \ldots$) obtained from the (OLS-)estimation of the original regression (4.13). The justification for this is that the powers of $\hat{y}_t$ are functions of the powers and cross products of the original regressors. A limitation is that each power of $\hat{y}_t$ is a linear combination of the original predictor variables, the squared predictors and cross products along with the coefficients of the original regression.

The RESET-test is an F-test based on incremental R^2. We provide details about this test in Sect. 5.4.1, but mention a few caveats here. First, the test only considers powers (and cross products, based on Ramsey's suggestion) of the "original" predictor variables $x_{1t}, \ldots, x_{Kt}$. If the null hypothesis is rejected, the model builder will lack information about the identity of the "real" missing variables. Interestingly, and not surprisingly given the use of powers, the test turns out to be useful for detecting nonlinearities. This weakens its attractiveness for detecting other specification problems, since rejection of a model may be due to either an omitted explanatory variable or a wrong functional form. The identification of missing variables is especially arduous. Second, we mentioned earlier that the bias in parameter estimates is a positive function of the degree of correlation (absolute value) between included and excluded predictors: the stronger the correlation, the greater the bias. However, if this correlation is strong, then the power of the test for omitted variables based on squared and higher-order terms of included predictors decreases. Third, the RESET-test is quite sensitive to autocorrelation (which we argue is likely to occur in the error term in case of model misspecification). Fourth, the RESET-test is a test of functional form misspecification only for nested models. The concept of nested models is discussed in Sect. 5.6.2.

In this discussion we emphasize the importance of proper specification of the predictors. One needs a substantial amount of knowledge about consumers, competitors, and markets in general (e.g. market structure) and the nature of possible

[1] Ramsey (1969, 1974).

effects of marketing instruments to justify a specific model formulation. This knowledge should be used to the fullest extent possible in the specification of a theoretically appropriate model. The model can then be tested against alternative specifications. A possibility to mitigate the omitted variable bias in an equation is to obtain *proxy variables* for the omitted variable. A proxy variable is a variable that is related to the unobserved variable (Wooldridge 2012, Section 9.2). This is called the *plug-in solution for the omitted variable problem*.

5.2.2 Heteroscedasticity

The second assumption, that the error term is *homoscedastic* (i.e. it has the same variance in all cases either cross-sectionally and/or over time), is not as critical as the first. Its violation "merely" reduces the *efficiency* of (OLS) parameter estimates. Thus if only the homoscedasticity assumption is violated, the least-squares estimator is (usually) unbiased but does not have minimum variance. In addition, the covariance matrix of the parameter estimates provides incorrect values because the estimates of the variances of the parameters are biased. In many cases the critical remedy is to use an appropriately adjusted formula for the variances and covariances of the parameter estimates.

The benefit of using an estimator that incorporates the heteroscedasticity is not always obvious. There are two relevant aspects to this. One is that the true source of heteroscedasticity is usually unknown. Thus, a deviation from homoscedasticity is usually determined based on data which introduces uncertainty. The other is that in this case the theoretically superior estimator is only asymptotically more efficient than the least-squares estimator. The benefit in practice, therefore, depends also on the sample size.

Experience in model building reveals that heteroscedasticity occurs especially if cross-sectional data are used for estimation. Thus, heteroscedastic disturbances

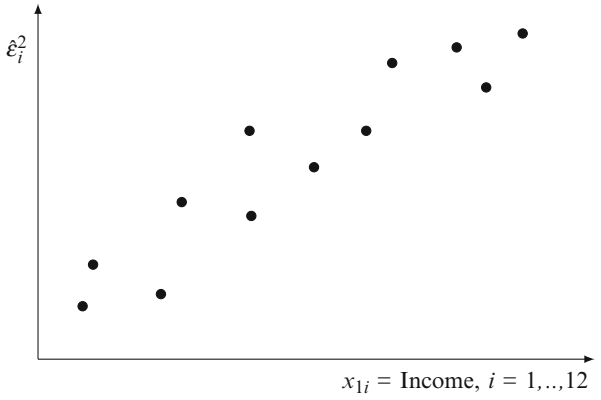

Fig. 5.1 A relation between the squared disturbance and income

have traditionally been accommodated in analyses of cross-sectional data. In the literature, fairly complex tests for homoscedastic disturbances exist. We restrict the discussion here to the basic phenomenon.

For heteroscedastic disturbance terms, consider a relation between the (squared) disturbance and a predictor variable such as income, as shown in Fig. 5.1. Income of family x_{1i}, $i = 1,\ldots,12$ is measured along the horizontal axis and the squared residual values occur along the vertical axis. The apparent dependence of the squared disturbance on income can be tested by estimating the following model:

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 x_{1i} + \nu_i \tag{5.2}$$

where ν_i = a disturbance term. If $H_0 : \gamma_1 = 0$ is rejected, then the assumption of homoscedasticity for ε_i is rejected.

Perhaps the best-known test of heteroscedasticity is the Goldfeld–Quandt test. This test is applicable if the model builder has a preconceived notion about the nature of heteroscedasticity (i.e. the residuals are not used to infer the nature of a possible violation of the assumption). Goldfeld and Quandt considered the problem that the residuals obtained from estimating *one* set of parameters from a sample of data are not independent. Thus, if the 12 observations in the application involving income as a predictor can be categorized prior to data analysis according to the expected magnitude of the squared disturbance (i.e. let $\mathrm{E}(\varepsilon_1^2) \geq \mathrm{E}(\varepsilon_2^2) \geq \mathrm{E}(\varepsilon_3^2)\ldots$), the ratio $\sum_{i=1}^{6}\hat{\varepsilon}_i^2 / \sum_{i=7}^{12}\hat{\varepsilon}_i^2$ does *not* follow the F-distribution. This is because the numerator and denominator in the ratio can be shown to be dependent.

Goldfeld and Quandt (1965) proposed a ratio whose numerator and denominator are independent under the null hypothesis of homoscedasticity by partitioning the original data as follows:

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} X_A & O \\ O & X_B \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} + \begin{bmatrix} \varepsilon_A \\ \varepsilon_B \end{bmatrix} \tag{5.3}$$

where the vectors and matrices with subscript A refer to the first $\frac{1}{2}n$ observations and those with B subscripts to the last $\frac{1}{2}n$ (where the observations are still ranked according to the expected magnitude of the squared disturbance under the alternative hypothesis of heteroscedasticity). Importantly, the residuals $\hat{\varepsilon}_i$, $i = 1,\ldots,n$ are obtained from fitting separate regressions to the first $\frac{1}{2}n$ and to the last $\frac{1}{2}n$ observations. The ratio of the residual sums of squares from these regressions:

$$\frac{\hat{\varepsilon}'_A \hat{\varepsilon}_A}{\hat{\varepsilon}'_B \hat{\varepsilon}_B} \tag{5.4}$$

is F-distributed with $\frac{1}{2}n - K - 1$ and $\frac{1}{2}n - K - 1$ degrees of freedom under the null hypothesis.

A *joint test* for homoscedasticity and correct model specification is given by White (1980).[2] Another well-known test to detect heteroscedascticity is the Breusch–Pagan test for heteroscedascticity (Breusch and Pagan 1979). This test is also based on the idea whether $\hat{\varepsilon}_t^2$ is related to one or more independent variables:

$$\hat{\varepsilon}_t^2 = \delta_0 + \delta_1 x_{1t} + \delta_2 x_{2t} + \cdots + \delta_K x_{Kt} + v_t \tag{5.5}$$

where v_t is a disturbance term with mean zero given the x_{kt}, $k = 1, \ldots, K$. The null hypothesis of homoscedasticity is:

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_K = 0.$$

The Breusch–Pagan test is based on running the regression (5.5). The R-squared from the regression is denoted by $R^2_{\hat{\varepsilon}^2}$, and is used to compute the F-statistic

$$F = \frac{R^2_{\hat{\varepsilon}^2}/K}{(1 - R^2_{\hat{\varepsilon}^2})/(T - K - 1)}, \tag{5.6}$$

which has (approximately) an $F_{K,T-K-1}$ distribution under the null hypothesis of homoscedasticity, and where T is the number of observations.

We now examine the heteroscedasticity issue more closely by assuming it to exist in a prespecified form. To simplify the exposition we consider the basic model (4.1) and suppose that:

$$\begin{aligned} &1.\ \mathrm{E}(\varepsilon_t) \quad = 0 \\ &2.\ \mathrm{Var}(\varepsilon_t) = \sigma^2 x_t^2. \end{aligned} \tag{5.7}$$

This second assumption indicates that the error variance increases with the squared value of the predictor variable (somewhat analogous to the relation in Fig. 5.1). For this specific form of heteroscedasticity we can use a transformation that results in the estimation of a different equation, where the new error term meets the homoscedasticity assumption and allows OLS estimation of the parameters of interest. If we multiply both sides of (4.1) by $1/x_t$ we obtain:

$$\frac{y_t}{x_t} = \frac{\alpha}{x_t} + \beta + \frac{\varepsilon_t}{x_t}. \tag{5.8}$$

In this form the error term ε_t/x_t still has the expected value equal to zero, since $\mathrm{E}(\varepsilon_t) = 0$ for a given value of x_t. The variance of the error term in (5.8) is:

$$\mathrm{Var}\left(\frac{\varepsilon_t}{x_t}\right) = \frac{1}{x_t^2}\mathrm{Var}(\varepsilon_t) = \frac{1}{x_t^2}\sigma^2 x_t^2 = \sigma^2. \tag{5.9}$$

[2] See also Judge et al. (1985, p. 453); Wooldridge (2012, p. 269).

In this case OLS is best, linear unbiased if it is applied to the new criterion variable y_t/x_t and the new predictor variable $1/x_t$. This transformation involves *weighting* each observation by $1/x_t$ so that as the value for x_t increases, the weight declines. This estimation method is also known as *weighted* least squares (WLS), and it is a special case of the generalized least squares (GLS-) estimation methods. We discuss these methods in Chap. 6.

5.2.3 Correlated Disturbances

If the residuals exhibit some systematic pattern over time, some or all of the covariances between residuals at different points in time will be nonzero, and the third assumption is violated. An important example is when the residuals exhibit first-order autocorrelation. Let us illustrate this by a modification of the residual term of the basic two-variable regression equation (4.1):

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, \ldots, T \tag{5.10}$$

where we use a different notation for the disturbance term because we assume that successive disturbances are related as follows:

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1. \tag{5.11}$$

We also assume:

$$\begin{aligned} \mathrm{E}(\varepsilon_t) &= 0 \\ \mathrm{Cov}(\varepsilon_t, \varepsilon_{t'}) &= 0, \quad t \neq t'. \end{aligned} \tag{5.12}$$

In (5.10) the error terms $u_1, u_2, \ldots, u_T$ are not independent but follow a so-called first-order AutoRegressive (AR) process with parameter ρ. This feature is called *autoregression*, autocorrelation or serial correlation. The parameter ρ is known as the autocorrelation parameter or autocorrelation coefficient. Under these assumptions, the efficiency of OLS-estimation is reduced, similarly to a violation of the second assumption. It is possible to show mathematically that the parameter estimates α and β are *unbiased*, but in practical applications, the parameter estimates may exhibit seizable deviations from their true values due to the reduced estimation efficiency. A violation of the third assumption can result from model misspecification (e.g. incorrect functional form, omitted variable(s), varying parameters).

To detect a violation of the assumption that the disturbances for different observations have zero covariance, one can *plot* the residuals ($\hat{u}_t$) against time. Figures 5.2 and 5.3 show cases of positive and negative autocorrelation, respectively. In Fig. 5.2, a positive residual tends to be followed by another positive one, and a negative residual tends to be followed by a negative one. Positive autocorrelation

means that the residual in t tends to have the same sign as the residual in $t-1$. On the other hand, in Fig. 5.3 we see that the observations tend to have positive values followed by negative ones, and vice versa, which result in a typical alternating pattern of negative autocorrelation.

The best-known test statistic to detect (first-error) autocorrelation is the one developed by Durbin and Watson (1950, 1951). The Durbin–Watson test statistic is based on the variance of the difference between two successive disturbances:

$$\mathrm{E}(u_t - u_{t-1})^2 = \mathrm{E}(u_t^2) + \mathrm{E}(u_{t-1}^2) - 2\mathrm{E}(u_t, u_{t-1}). \tag{5.13}$$

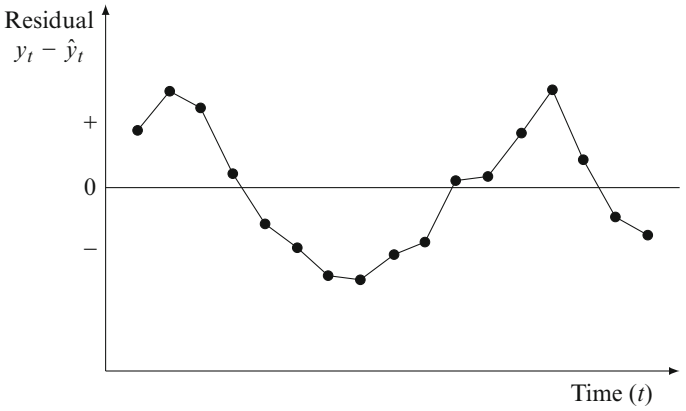

Fig. 5.2 Positive autocorrelation

If successive disturbances are positively correlated (positive autocorrelation), the expected value of (5.13) is small, because of the negative sign of $-2\mathrm{E}(u_t, u_{t-1})$. Similarly, negative autocorrelation causes this last part of (5.13) to contribute to a high expected value for $\mathrm{E}(u_t - u_{t-1})^2$. To the extent that the residuals $\hat{u}_t$ obtained by the ordinary least squares method are satisfactory approximations of the corresponding random disturbance terms u_t, we have a similar result for $(\hat{u}_t - \hat{u}_{t-1})^2$. These considerations lead to the Durbin–Watson test statistic:

$$\mathrm{DW} = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2}. \tag{5.14}$$

The DW statistic varies between zero and four. Small values indicate positive autocorrelation, large values negative autocorrelation.

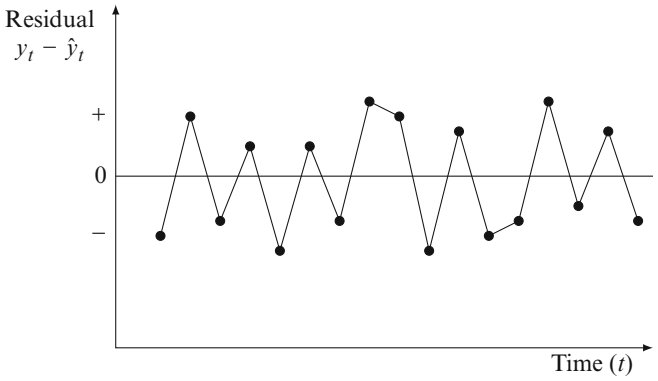

Fig. 5.3 Negative autocorrelation

Durbin and Watson (1950, 1951) formulated lower and upper bounds (d_L, d_U) for various significance levels,[3] and for specific sample sizes and numbers of parameters. The test statistic is used as follows:

1. Tests for positive autocorrelation:

   (a) If DW $< d_L$, there is positive autocorrelation;
   (b) If $d_L <$ DW $< d_U$ the result is inconclusive;
   (c) If DW $> d_U$, there is no positive autocorrelation.

2. Tests for negative autocorrelation:

   (a) If DW $> 4 - d_L$, there is negative autocorrelation.
   (b) If $4 - d_U <$ DW $< 4 - d_L$ the result is inconclusive.
   (c) If DW $< 4 - d_U$, there is no negative autocorrelation.

The Durbin–Watson test is not very powerful in the sense that the inconclusive range can be quite large. Judge et al. (1985, p. 330) recommend that the upper critical bound (d_U) is used instead of the lower bound. Thus they essentially include the inclusive region as evidence of autocorrelation. We note that the Durbin–Watson statistic is a test for first-order autocorrelation, and does not consider higher-order autoregressive schemes [see: (5.16)].

In the presence of a lagged criterion variable among the predictor variables, the DW statistic is biased towards finding no autocorrelation. For such models Durbin (1970) proposed a statistic (Durbin's h) defined as:

$$h = \hat{\rho} \sqrt{\frac{T}{1 - T\hat{\sigma}^2_{\hat{\lambda}}}} \tag{5.15}$$

[3]Tabulated values of d_L, d_U for different significance levels can be found in most textbooks on econometrics.

where $\hat{\rho}$ is an estimate for ρ in (5.11), T is the number of observations, and $\hat{\sigma}^2_{\hat{1}}$ is the estimated variance of the slope coefficient for the lagged endogenous variable. Durbin (1970) shows that, asymptotically, h is a standard normal deviate. However for *small samples* the Durbin–Watson test is more powerful than the Durbin's h statistic in detecting $\rho \neq 0$ as demonstrated by Kenkel (1974).

Many software programs automatically compute and report the value of the Durbin–Watson statistic. This is very useful for longitudinal data, but for cross-sectional observations the test is meaningless, because the test assumes that the observations are ordered, which is not the case in a cross-sectional setting.

We note that econometric textbooks suggest a remedy that essentially incorporates the systematic pattern in the residuals in the estimation method. We discuss such an approach in Sect. 6.2. We believe that this remedy should only be a last-resort option. That is, the model builder should first do everything possible to obtain an acceptable model specification, such that the residuals are uncorrelated. If the residuals of an early version of a model turn out to be correlated, the first option should be to investigate whether the model properly accommodates the dynamic effects of marketing efforts. In Sect. 2.8 we extensively discussed several options for incorporating marketing dynamics in the model. If none of these options provide the desired uncorrelatedness of the residuals, the procedure in Sect. 6.2 can be used.

Traditionally, the first-order autoregressive process (AR(1)) represented by (5.11) has been the primary autocorrelation process considered in econometrics (Judge et al. 1985, pp. 226–227). For annual data, the AR(1)-process is often sufficient for models with autocorrelated disturbances (with the caveat that it may mask other shortcomings in model specification). More recently, for models of frequently occurring data such as daily online expenditures and with the virtual elimination of computational constraints in estimation, other specifications have been considered. These include autoregressive processes of finite order greater than one.[4] An autoregressive process of order p, AR(p), has the following form (for the disturbance term):

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \tag{5.16}$$

where ε_t satisfies (5.12).

To accommodate seasonal processes, for quarterly data, p may be 4. Thomas and Wallis (1971) suggest that with quarterly data only the disturbances in corresponding quarters of successive years should be correlated. This leads to the specification of a *restricted* AR(4)-process:

$$u_t = \rho_4 u_{t-4} + \varepsilon_t. \tag{5.17}$$

For monthly data, $p = 12$ may be appropriate.

[4]For other stochastic processes such as moving average processes and combined autoregressive moving average processes, see, e.g., Judge et al. (1985, Chapter 8).

Lagrange Multiplier (LM) tests can be used to determine the choice of an autocorrelation process. For restricted AR(p)-processes, the LM-test statistic $T\hat{\rho}_p^2$ asymptotically has a $\chi^2_{(1)}$ distribution where T = the total number of observations, and

$$\hat{\rho}_p = \frac{\sum_{t=1}^{T-p} \hat{u}_t \hat{u}_{t+p}}{\sum_{t=1}^{T} \hat{u}_t^2} \tag{5.18}$$

where $\hat{u}_t$, $\hat{u}_{t+p}$ are the OLS-residuals.

For the more general alternative (5.16), to test the null hypothesis $\rho_1 = \rho_2 = \cdots = \rho_p = 0$ the LM test statistic

$$T \sum_{r=1}^{p} \hat{\rho}_r^2 \tag{5.19}$$

asymptotically follows a $\chi^2_{(1)}$ distribution.

In Sect. 6.2 we consider how to estimate the parameters in a model with AR(1) disturbances. For the estimation of models with disturbances that follow higher-order AR processes we refer to Judge et al. (1985, pp. 293–298).

5.2.4 Nonnormal Errors

The fourth assumption, that the disturbances are normally distributed, may also be violated due to model misspecification. It makes sense, therefore, that this assumption will not be examined until the model specification is reconsidered, if necessary. For the same reason it is efficient to examine the plausibility of the second and third assumptions before one checks the fourth. The disturbances need to be normally distributed for the standard test statistics for hypothesis testing and confidence intervals to be applicable. We can examine the validity of the fourth assumption indirectly through the residuals. If each error term separately satisfies all four assumptions, then the estimated error values (residuals) as a group will be normally distributed. Thus, if the residuals appear to be normally distributed, we cannot reject the hypothesis that each unobservable error term follows a normal distribution.

To examine whether the residuals are approximately normally distributed, we can categorize the residuals and construct a histogram. Such a histogram shows the

relative frequency with which the (standardized)[5] residuals fall into defined classes. An inspection of the histogram may suggest deviations from normality in the form of skewness or kurtosis.

The literature on testing for normality is vast. Some commonly used tests are the Kolmogorov–Smirnov test, the likelihood ratio test and the Kuiper test. Other tests of the normality assumption are the Shapiro–Wilk test[6] and the Jarque–Bera test.[7] The Jarque–Bera test statistic for non-normality is chi-square distributed (with two degrees of freedom):

$$\chi^2_{(2)} = \frac{(T-L)}{\hat{\sigma}^2}(\widehat{sk}^2 + \tfrac{1}{4}\widehat{ek}^2) \tag{5.20}$$

where

T = number of observations,

L = number of parameters,

$\hat{\sigma}$ = standard deviation of the residuals,

$\widehat{sk}$ = skewness of the distribution of the residuals (3rd moment),

$\widehat{ek}$ = excess kurtosis of the distribution of the residuals (4th moment).

Thus, this test determines whether the third and fourth moments of the residuals are consistent with the null hypothesis of normality.

For some models that have nonnormal disturbances a transformation may produce normally distributed errors. Suppose that, if the effects of predictor variables are accounted for, y_t is log-normally distributed. Then by taking the logs of y_t the new criterion variable $\ln y_t$ is normally distributed (and hence we have normally distributed errors). This is a special case of a class of transformations considered by Box and Cox (1964). They assume that there exists a value λ such that

$$\frac{y_t^\lambda - 1}{\lambda} = \beta_1 x_{1t} + \cdots + \beta_K x_{Kt} + \varepsilon_t, \quad \lambda > 0 \tag{5.21}$$

where the disturbance term ε_t is normally distributed (and homoscedastic). It can be shown that:

$$\lim_{\lambda \to 0} \frac{y_t^\lambda - 1}{\lambda} = \ln y_t. \tag{5.22}$$

[5]The residuals are standardized by dividing the observed values by the standard deviation of the residuals (the average of the residuals equals zero under usual conditions).

[6]See Shapiro and Wilk (1965).

[7]See Bera and Jarque (1981, 1982); Stewart (1991, p. 162).

Apart from a difference in the intercept, $\lambda = 1$ yields the basic model for the K-variate case (4.13). In general, however, the Box–Cox transformation is applied primarily to let the data determine the most appropriate functional form. See Kristensen (1984) for a marketing application.

We note that with increases in sample sizes, the null hypothesis of normally distributed errors will be rejected more often. Consequently, even minor deviations from normality can signify a violation of the assumption required for traditional statistical inference. If the model specification seems appropriate, it will not be appealing to follow strict rules with regard to such violations. To accommodate cases in which the normality assumption does not hold but the model specification is acceptable, researchers have developed robust regression methods. The interested reader is referred to Huber (1973), Hinich and Talwar (1975) and Judge et al. (1985, pp. 828–839). For marketing applications, see Mahajan et al. (1984).

One of the reasons why non-normality tests may indicate deviations from the normal distribution is the presence of outliers in the residuals. Especially in small data sets the OLS estimates are sensitive to outliers, because large residuals receive a lot of weight in the least squares minimization problem. Consequently, when deviations from normality are caused by outliers, OLS results may no longer be unbiased or efficient. Figure 5.4 shows different patterns of outliers.

In panel (a) of Fig. 5.4, the outlier is an extreme observation relative to the other observations of both x_t and y_t, but strengthens the relationship between x and y that we see in these observations. This outlier does not result in biased estimates, and actually increases R^2. That is not the case in panels (b)–(e). In those situations, the outliers are not in line with the relationship between x and y that is suggested by the other observations, and the outliers reduce R^2. In panel (b) the outlier is extreme relative to the other observations of both x_t and y_t. The outlier in panel (c) is extreme relative to the other observations of y_t but not with respect to the other observations of x_t. In panel (d) the outlier is not extreme relative to the other observations of x_t, but it is relative to other observations of y_t. The outlier in panel (e) is neither extreme in x_t, nor in y_t.

The take-away from Fig. 5.4 is that we should not judge normality or the presence of outliers in terms of x or y alone: a proper assessment also takes the relation between x and y into account. Therefore it is advisable to start with testing the residuals for outliers and normality. If outliers are detected, x and y variables should be inspected for anomalies.

In practical settings, it is advisable to estimate the model twice, with and without outliers, to assess the influence of the unusual observations on the estimation results. However, inspecting observations in trying to determine which are outliers, and which ones have substantial influence on the (OLS) estimator is a difficult endeavor (Wooldridge 2012, p. 316). Formal tests have been developed and are discussed by Belsley et al. (1980).

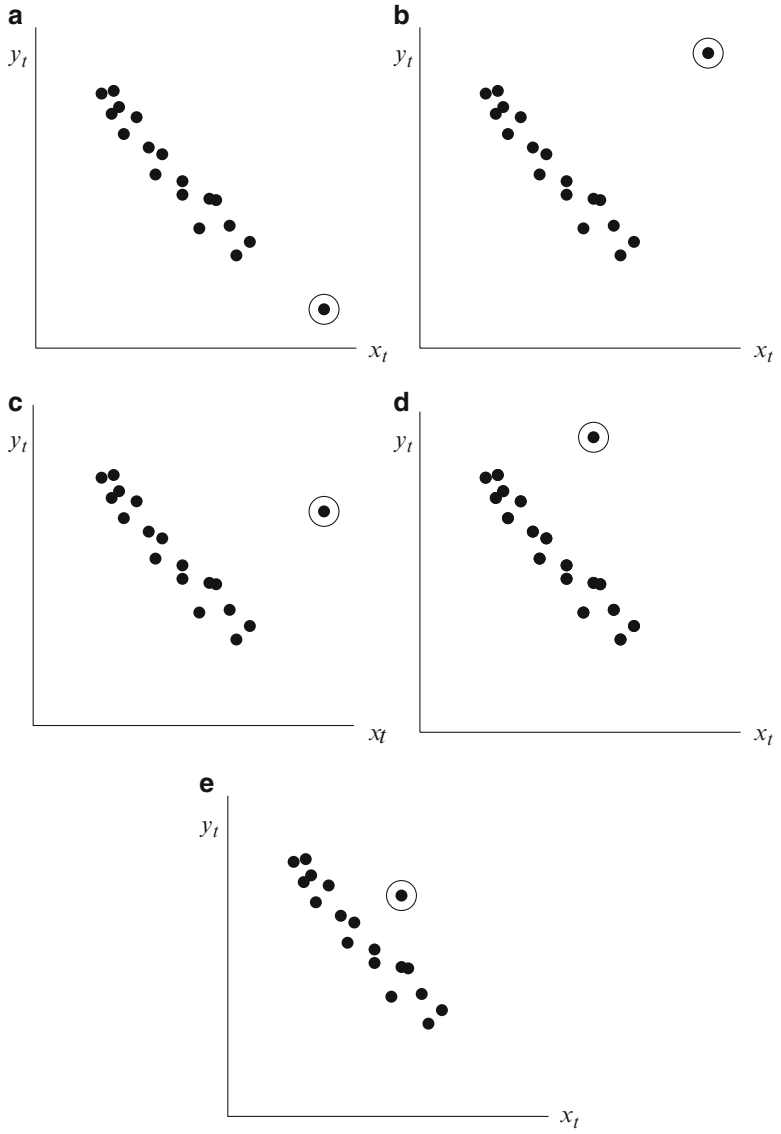

Fig. 5.4 Patterns of outliers

5.2.5 Endogenous Predictor Variables

It is often convenient to consider the error-term assumptions under the condition that the predictor variables x_k, $k = 1,\ldots,K$ are nonstochastic or "exogenous". Essentially this means that we would take the observed values for the predictor

variable(s) as given. The assumption is violated (i.e. $\mathrm{E}(\varepsilon_t \mid x_{kt}) \neq 0$) amongst others if x_{kt} in fact depends on y_t. Given that y_t depends on ε_t there is a non-zero correlation between the independent variable and the disturbance term. This is called *endogeneity*. The consequence of a violation of this assumption is that the ordinary least squares parameter estimates are biased. Endogeneity makes the least-squares estimator biased and inconsistent. We note that endogeneity is not detectable from an inspection of the residuals in the original equation. Thus, the model builder must possess the substantive knowledge that is critical for meaningful model building. Endogeneity can arise from a number of different sources (Ebbes 2004, pp. 8–15):

1. relevant omitted variables;
2. measurement error in the regressors;
3. serially correlated error in the presence of lagged dependent variables in the set of regressors;
4. simultaneity.

Proofs of how the first three sources lead to violation of assumption 5 ("endogeneity") are given by, for example, Ebbes (2004) and Wooldridge (2012, p. 491). Here we discuss how simultaneity leads to endogeneity. If, for example, price and advertising are used as predictor variables in a demand equation, the model builder must know how price and advertising decisions are made and whether this decision-making process should be taken into account. For example, advertising expenditures of brand j may be based on the sales of brand j, prices may be based on past prices *and* sales of brand j, and competitive sales (Horváth et al. 2005). But if price depends on sales, it is also related to the error term in the sales equation, resulting in endogeneity. We elaborate on this example in Sect. 6.5.

Before we pinpoint tests for endogeneity, we first briefly touch on possible remedies to account for or to reduce endogeneity.

One remedy for a violation of the assumption is to find *instrumental variables* Z such that $\mathrm{E}(\varepsilon \mid Z) = 0$, while the variables in the matrix Z are highly correlated with the endogenous variables in X, and the number of variables in Z is at least K. It is easy to demonstrate that the instrumental variable estimator is consistent since $\mathrm{E}(\varepsilon \mid Z) = 0$.

Specifically, suppose that only *one* predictor variable in the X matrix is correlated with the error term ε, say x_{k^*}. We then only need to find instrumental variables to obtain $\hat{x}_{k^*}$ such that $\hat{x}_{k^*}$ is uncorrelated with ε but (highly) correlated with x_{k^*}. For the other predictors we use $\hat{X}_k = X_k$ $(k \neq k^*)$, since the predictors themselves are uncorrelated with ε and there are no other variables more highly correlated with X_k for $k \neq k^*$.

We want to employ multiple instrumental variables, if possible, to create the highest possible correlation between $\hat{x}_{k^*}$ and x_{k^*}, subject to $\mathrm{E}(\hat{x}_{k^*} \mid \varepsilon) = 0$. This can often be accomplished by using lagged values of the criterion and predictor variables. Thus, let

$$x_{k^*} = Z\gamma + v \tag{5.23}$$

where

x_{k^*} = the predictor variable that is correlated with the error term ε,

Z = a matrix of instrumental variables, possibly consisting of all lagged y and X variables,

v = vector of disturbances.

Then $\hat{x}_{k^*} = Z(Z'Z)^{-1}Z'x_{k^*}$. We replace the values for x_{k^*} in the X-matrix with the values for $\hat{x}_{k^*}$, and apply least squares (this is the Two-Stage Least Squares or 2SLS estimator). For further detail, see: Sect. 6.6 and for example Greene (2012, p. 357); Wooldridge (2012, Chapter 15).

Tests for endogeneity compare the estimates of the so-called instrumental variables (IV) method with the OLS estimates (Hausman 1978). This Hausman test and other endogeneity tests are discussed after a more in-depth introduction of the IV-estimation method in Sect. 6.6.

Other remedies to deal with endogeneity are:

- modeling a simultaneous system of equations (see Sect. 6.5);
- the use of spatial models (Bronnenberg and Mahajan 2001; Van Dijk et al. 2004);
- the use of so-called control variables (Card 1999, 2001);
- latent instrumental variables (Ebbes 2004, Ebbes et al. 2005, 2009);
- a number of specific approaches which are model-specific (Kuskov and Villas-Boas 2008; Park and Gupta 2009);
- the control function approach (Petrin and Train 2010; Sridhar and Srinivasan 2012);
- copulas (Park and Gupta 2012).

5.2.6 *Multicollinearity*

Real-world data often show high degrees of correlation between predictors. Take, for example, the specification of a model to explain the demand for a product category at the household level. Variables such as household income and age of the head of the household may be relevant predictor variables. However, these variables tend to be correlated. In particular, income is partly a function of age.

In Model (4.14) we assume that the matrix of observations X has full rank (see Sect. A.4 in the Appendix on Matrix Algebra), or that the vectors in X are linearly independent. If this assumption is violated we encounter perfect collinearity. If X has a rank smaller than K, $X'X$ has a determinant with value zero. In that case $(X'X)^{-1}$ does not exist and the parameter estimates cannot be uniquely determined. Perfect collinearity is, however, unusual. If it happens, it is often due to a duplicate variable. The notion of "almost" linearly dependent vectors is, however, meaningful. This means that the determinant of $X'X$ is not equal to zero, but its value will be near zero. Hence, $(X'X)^{-1}$ exists, but its elements will be large. As a consequence, the parameter estimates will have large variances and covariances [see (4.22)]. Thus,

multicollinearity (almost linearly dependent vectors in the X-matrix) makes the parameter estimates unreliable.

Collinearity between predictor variables can also result from decisions made by managers. Consider again the example that we discussed before, that unit brand sales is a function of advertising and price. It seems inappropriate to assume that either variable is "fixed". Instead, a decision maker manipulates both variables, based on strategic considerations, marketplace conditions, cost considerations, etc. Imagine that consumers' price sensitivity depends on the amount of advertising.[8] A manager who manipulates both variables will then refrain from changing these variables independently. In that case, a real-world correlation between these marketing decision variables reflects, at least partly, the decision maker's belief about interactions between these variables. If her belief is very strong (i.e. price changes are strongly related to advertising changes), then it can become difficult or impossible to obtain reliable estimates of the separate main effects from marketplace data. And, verification of an interaction effect will then be especially difficult.

Some of the procedures for detecting the presence of multicollinearity are:

1. The traditional approach to check for collinearity is to examine the correlation matrix R of the predictor variables. Highly positive or negative correlations are indications of potential difficulties in the estimation of reliable effects. Often empirical researchers examine the matrix of all bivariate correlation coefficients between the predictor variables. Based on some cut-off value a decision might be made about which pairs of predictors should not be included together. One problem with this procedure is that all bivariate correlations may be low and yet one predictor may be highly related to a linear combination of the remaining predictors (see approach 3 below). Another problem is that the severity of a high bivariate correlation between two predictors depends on the sample size. For that reason statistical tests of multicollinearity[9] are meaningless. In this regard, Mason and Perreault (1991); Cameron and Trivedi (2009, p. 350) demonstrated that the harmful effects of collinear predictions are often exaggerated and that collinearity cannot be viewed in isolation. They argue that the effects of a given level of collinearity must be evaluated in conjunction with the sample size, the R^2 value of the estimated equation, and the magnitudes of the slope coefficients. For example, bivariate correlations as high as 0.95 have little effect on the ability to recover the true parameters if the sample size is 250 and the R^2 is at least 0.75. By contrast, a bivariate correlation of 0.95 in conjunction with a sample size of 30 and an R^2 of 0.25 results in failure to detect the significance of individual predictors.[10]

[8]For a discussion of theoretical arguments and empirical generalizations, see Kaul and Wittink (1995).

[9]See Farrar and Glauber (1967); Kumar (1975) or Friedman (1982).

[10]Currently, the spectral decomposition of $X'X$ is advocated for the quantification of multicollinearity, based on one or more characteristic roots. In this respect researchers quite often use the *condition numbers* of $X'X$, that is the square root of the ratio of the largest to smallest eigenvalue of $X'X$; see Cameron and Trivedi (2009, pp. 350–351).

2. Do a factor analysis of the predictor variables to examine the structure of overlapping variation between the predictors. This is a more comprehensive examination of multicollinearity than the traditional approach which considers only the simple correlations.
3. Use multiple *regression* with x_{kt} as the *criterion variable* and the x_{rt}, $r = 1,\ldots,K, r \neq k$ as the predictor variables, for all $k = 1,\cdots,K$. Each regression indicates how much one predictor is related to a linear combination of the other predictors. The R_k^2 values that result from these regressions, one at a time, can be used to quantify the overlap among any number of predictor variables. A related measure is the Variance Inflation Factor (VIF) computed as $1/(1-R_k^2)$. A VIF greater than 5 is often taken to signal that collinearity is a problem (De Vaus 2013). An equivalent measure is the *"tolerance"*, which is 1/VIF.
4. Estimate the complete model with y_t as a function of all K predictor variables. Next, *eliminate* one of the predictor variables x_{kt}, and *re-estimate* the model. The parameter estimates that change the most are associated with predictor variables that are collinear with the deleted variable x_{kt}. This can be done for each predictor variable in turn.
5. *Comparing results for F-test and t-tests*. Multicollinearity may be regarded as acute if the F-statistic shows significance and none of the t-statistics for the slope coefficients is significant.

5.2.6.1 Solutions to Multicollinearity

If any two predictors are perfectly correlated, the parameters of a regression equation cannot be estimated. Thus no solution to a multiple regression problem (other than combining the two predictors or deleting one) can be obtained if there is *extreme* multicollinearity. The approaches available for the resolution of non-extreme but severe multicollinearity include:

1. obtain more data relevant to the problem;
2. reformulate the model with the specific objective to decrease multicollinearity;
3. create new predictors;
4. apply estimation methods specifically developed for cases with severe multicollinearity;
5. eliminate a predictor variable with a t-ratio close to zero.

The last approach is simply that a predictor variable with a statistically insignificant t-ratio be eliminated. In general, an insignificant t-ratio indicates that the predictor variable is irrelevant or that it has an effect but we cannot obtain a reliable parameter estimate. The elimination of such a predictor variable from the model should be a last-resort option.

A better solution, of course, is to obtain more data. With more data, especially data with a reduced degree of multicollinearity, it is more likely that a significant effect can be obtained for any predictor variable. However, the opportunity to add data to the sample is often limited.

Another possible solution is to reformulate the model. In some sense, the elimination of one predictor variable amounts to a model reformulation. But we may also combine two predictor variables and in that manner resolve the problem. For example, this is appropriate if the two variables are substitute measures of the same underlying construct such as when the observations represent characteristics of individuals, and two of the predictors measure age and work experience. These two variables tend to be correlated, because both may measure individuals' learned skills for a certain job as well as their maturity. It may be sufficient to use one of these variables or to define a new variable that combines these two predictor variables.

If the source of multicollinearity stems from two or more predictor variables that capture different phenomena, such combinations are not appropriate. However, it may be possible to redefine the variables. Consider, for example, a demand model in which product category sales in t (Q_t) is explained by predictors such as total disposable income in t (Inc_t) and population size in t (N_t). These predictors are collinear because Inc_t is a function of the size of the population. However, by defining the variables on a per capita basis, such as Q_t/N_t and Inc_t/N_t, we create a simpler model and eliminate the collinearity between Inc_t and N_t.

Other functional specifications of the model can also reduce multicollinearity. For example, consider relation (4.13). The predictor variables $x_{1t},\ldots,x_{Kt}$ vary over time,[11] and each of these variables may have a component that follows a common trend. In that case, bivariate correlation between the predictors will be high due to the common factor. The multicollinearity can be considerably reduced in the following manner. First specify (4.13) for $t-1$:

$$y_{t-1} = \alpha + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \cdots + \beta_K x_{K,t-1} + \varepsilon_{t-1}. \tag{5.24}$$

By subtracting (5.24) from (4.13), we express the model in terms of changes over time:

$$\begin{aligned}(y_t - y_{t-1}) = {} & \beta_1(x_{1t} - x_{1,t-1}) + \beta_2(x_{2t} - x_{2,t-1}) + \cdots + \beta_K(x_{Kt} - x_{K,t-1}) \\ & + (\varepsilon_t - \varepsilon_{t-1}).\end{aligned} \tag{5.25}$$

Note that the reformulated model has no intercept. Apart from that, the model contains the same parameters. Importantly, this reformulation will not have the same degree of multicollinearity as the original model. Such opportunities for model reformulation are useful to consider when severe multicollinearity is encountered.

[11] See also Rao et al. (1988).

Some researchers advocate the creation of a matrix of orthogonal variables, constructed as linear combinations of the original predictor variables, based on principal components analysis or factor analysis.[12] If all components or factors are included in a subsequent multiple regression analysis, we can claim to have "solved" the collinearity problem in the sense that the components (factors) are uncorrelated. However, we are unlikely to have a useful substantive interpretation for the estimated effects. And if we transform these estimated effects based on the relations between the factors and the original predictors, we will derive exactly the same slope coefficients as if we had not first constructed the matrix of orthogonal variables. Only by deleting at least one orthogonal component (e.g. the one with the smallest explanatory power of y) do we have a chance of improved results. In this case we can also transform the effects back in terms of the original variables, in which case we have (slightly) biased effects with reduced statistical uncertainty.

Finally, we briefly consider the consequences of multicollinearity on understanding versus forecasting. If the purpose of the study is to develop a forecasting model, multicollinearity may not be an issue. For example, suppose x_1 is eliminated from the model because it is highly correlated with a linear combination of the other predictor variables. If x_1 is truly a relevant predictor, then the parameter estimates for the remaining predictors will be biased. Nevertheless, the forecasts produced by this otherwise deficient model may still be accurate, since most of the explanatory power of x_1 is contained in one or more other predictor variables. As long as this correlation between x_1 and the other predictors *continues to exist*, the forecasts from this deficient model will *not* be systematically affected. On the other hand, if this correlation changes, the accuracy of the forecasts will be affected by the model's deficiency. This happens especially when the predictor variables are controlled by a manager or a policy maker. The reason is that the elimination of a relevant predictor variable biases the coefficients of the remaining predictor variables which affects the accuracy of conditional forecasts, i.e. the forecasts that depend on the predictor variables that are controlled by a manager or policy maker.

The nature and magnitude of this bias depends, among other things, on the degree of multicollinearity. If the degree of multicollinearity changes from the estimation sample to new data, then the nature of the bias changes, thereby affecting the forecasting accuracy. Thus, for *conditional* forecasting multicollinearity is a serious problem, and the "solution" has to be similar to the situation when the primary objective of model building is to understand relationships.

If the objective is to understand or describe the relationship between a criterion variable and several predictor variables, it is critical that all relevant variables are included in the model. This is also true for *normative decisions*. The presence of multicollinearity may make it difficult or impossible to obtain the desired degree of reliability (i.e., low standard errors) for the coefficients. As a result the computed

[12] See, for example, Massy (1965) and Sharma (1996, Chapter 4).

t-ratio for the slope coefficient of one or more predictor variables may be below the critical value. Yet for understanding it is not advisable to eliminate a (relevant) predictor.

5.3 Mediation Tests

Using the notation that was introduced in Fig. 2.4 in Sect. 2.5, testing for mediation involves the steps outlined in Baron and Kenny (1986):

1. Regress the outcome variable y_t on the causal variable x_{1t} to confirm x_{1t} significantly predicts y_t (path c).
2. Regress the mediator variable x_{2t} on the causal variable x_{1t} to confirm x_{1t} significantly predicts x_{2t} (path a).
3. Regress the outcome variable y_t on both the mediator (x_{2t}) and the causal (x_{1t}) variable to confirm that the mediator significantly predicts y_t (path b), while controlling for x_{1t} (path c').

The steps for testing mediation are the same for different types of models and do not depend on the estimation method of the model, which could include ordinary least squares regression, logistic regression and multilevel (e.g. hierarchical linear) models (Kenny 2013).

For mediation, the path b coefficient (x_{2t} on y_t) needs to be significant and the path c' coefficient should be smaller in absolute value than the total effect (path c in step 1). Full mediation occurs when the path c' coefficient does not significantly differ from 0.

Unfortunately, this step-by-step testing approach, popular in psychology, has low statistical power (MacKinnon et al. 2002). From sociology comes the approach to test for the product of coefficients ($a \times b$): mediation is more likely the higher the $x_{1t} \rightarrow x_{2t}$ effect (path a) and the $x_{2t} \rightarrow y_t$ effect (path b) are, and the lower their standard errors. Therefore, this test divides the estimate of the mediating variable effect $a \times b$ by its standard error and compares this value to a standard normal distribution. The most commonly used standard error is the approximate formula derived by Sobel (1982) using the multivariate delta method based on a first-order Taylor series approximation, which is then compared to a standard normal distribution to test for significance ($H_0 : a \times b = 0$). The derivation of the unbiased variance of the product of two normally distributed variables is based on Goodman (1960), hence the name 'Sobel–Goodman' test. Comparing these and other methods in a Monte Carlo study, MacKinnon et al. (2002) report an excellent balance of Type I error and statistical power for this test of the joint significance of the two effects ($a \times b$) comprising the mediating variable effect. Practically, the Sobel–Goodman test is incorporated in software packages such as STATA. Alternatives include the bootstrapping test by Hayes and Scharkow (2013) in SPSS and SAS macros, and the Monte Carlo Simulation by MacKinnon et al. (2004) in the software package R.[13]

[13] Available from Selig and Preacher at http://www.quantpsy.org/medmc/medmc.htm.

Hanssens et al. (2014) use the Sobel–Goodman mediation test (Sobel 1982) on their model of partial mediation by "brand health indicators" of the marketing-sales relationship. Examples of these indicators are awareness, consideration and liking. The results for the Juice category are shown in Table 5.2. The indirect effect of advertising on sales via brand health indicator "awareness" is 0.0008. The total indirect effect of advertising via all brand health indicators is 0.0038. The proportion of the total effect that is mediated due to the brand health indicators is 48.60 % of the total effect of advertising on brand sales. As expected, the more tactical action of "promotion" affects sales less through brand health indicators (8.86 %) and thus more through the direct "transactions" route. Across all studied brands and product categories, the Sobel–Goodman tests reveal evidence of partial mediation of the marketing-sales effect by the three brand health indicators. This leads Hanssens et al. (2014) to include the brand health indicators in their marketing-sales model.

Table 5.2 How brand health indicators partially mediate the effect of advertising on juice brand sales

	Awareness	Consideration	Liking	Total indirect	Direct effect	Mediated effect
Price	−0.0080	0.0083	−0.0728	−0.0725	−0.2020	26.42 %
Promotion	0.0012	0.0076	0.0007	0.0095	0.0980	8.86 %
Advertising	0.0008	0.0003	0.0027	0.0038	0.0040	48.60 %

Source: Based on Hanssens et al. (2014)

As a final note, Kenny (2013) emphasizes that a mediation model is a causal model: if the researcher is incorrect on the causation, the results of a mediation analysis are "of little value". Using time series data, Srinivasan et al. (2010) show that marketing actions and brand health indicators Granger cause sales more often than the other way around (see Sect. 5.4.3 for Granger Causality Tests).

5.4 Joint Tests, Pooling Tests and Causality Tests

In this section we consider joint tests (Sect. 5.4.1), pooling tests (Sect. 5.4.2) and introduce causality testing (Sect. 5.4.3).

5.4.1 Joint Tests

Most of the diagnostic tests that we have reviewed in this chapter were developed to validate a single specific assumption of the classical linear model. However, the tests may require the acceptance of other assumptions. For example, the tests for autocorrelation assume that the disturbances are homoscedastic. In empirical

applications, various assumptions may be jointly violated. This implies that the diagnostic value of individual test results is in doubt. An alternative approach is to apply joint tests for multiple assumptions. For example, the RESET test can be considered as a joint test. This test is based on the idea that various misspecifications can lead to a violation of the assumption that the expected value of the disturbances is zero. Possible misspecifications include omitted variables, a wrong functional form, dependence between the regressors and the disturbances, etc. In this sense, the RESET test is also a joint test.

In this section we consider joint test of subgroups of slope parameters. As we indicated in the discussions in Sects. 4.3 and 5.2.6, it is possible for F- and t-test results to be in conflict. If the collinearity is concentrated in a subset of the predictor variables, then it is possible that two or more slope coefficients are insignificant based on individual t-tests but provide significant incremental explanatory power as a subgroup.

Tests of subgroups of variables are also of interest in case the model builder has arguments for the inclusion or exclusion of two or more predictor variables jointly. These joint tests are based on the incremental explanatory power of a group of at least two predictor variables. We use Eq. (5.26), a demand equation for the product class sales of detergents, to provide an illustration. We show the OLS-results of (5.26) in Table 5.3.

$$\hat{Q}_t = 202 - 0.29\frac{Inc_t}{CPI_t} - 29.5\frac{\bar{p}_t}{CPI_t} + 0.31\frac{a_{t-1}}{CPI_t} - 0.29t \tag{5.26}$$

where

Table 5.3 Multiple regression results (5.26)

	Predictor variable	Regression coefficient	Estimated standard error	t-value
1	(Intercept)	202	41	4.93
2	Income (deflated by Price Index)	−0.29	10.1	0.03
3	Average price (deflated by Price Index)	−29.5	11.7	2.52
4	Total advertising expenditures (deflated by Price Index)	0.31	0.15	2.07
5	Trend	−0.29	0.47	0.61
	$T = 32$, $R^2 = 0.47$			

$\hat{Q}_t$ = the estimated value of the product category sales (detergents) in thousands of kilos in time t,

Inc_t = the national income in nominal terms in t,

CPI_t = Consumer Price Index in t,

$\bar{p}_t$ = average price of detergents (average over n brands) in t in Euros,

a_{t-1} = the (lagged) advertising expenditure of all brands in thousands of Euros in t,

t = time index, $t = 1, \ldots, 32$,

T = the number of observations.

The F-test of the model as a whole indicates that the null hypothesis $\beta_1 = \cdots = \beta_K = 0$ can be rejected. From tables of the t-distribution we find that the critical value of the 5 % level of significance with $32 - 5 = 27$ degrees of freedom is 1.70 for a one-tailed test, and 2.05 for a two-tailed test. Thus, we cannot reject the null hypothesis that "Income" and "Trend" have no effect on the criterion variable, the demand for detergents at the product class level.

To test these two predictor variables' incremental explanatory power jointly, we estimate an alternative, *restricted*, model by eliminating Income and Trend. This restricted model has an R^2 of 0.43. The test of the null hypothesis that the parameters of Income (β_1) and Trend (β_4) are simultaneously equal to zero is performed using *an F-test* based on *incremental* R^2. Hence:

$$H_0 : \beta_1 = \beta_4 = 0$$

is tested with the statistic:

$$F = \frac{(R_F^2 - R_R^2)/(\mathrm{df}_R - \mathrm{df}_F)}{(1 - R_F^2)/\mathrm{df}_F} \tag{5.27}$$

where

R_F^2 = the unadjusted R^2 for the full model,

R_R^2 = the unadjusted R^2 for the restricted model,

df_R = the number of degrees of freedom left for the restricted model,

df_F = the number of degrees of freedom left for the full model.

In this example:

$$F = \frac{(0.47 - 0.43)/(29 - 27)}{(0.53)/27} \approx 1.02.$$

At the 5 % level, the critical F-value equals 3.35 ($F_{2,27}$). Thus we cannot reject the null hypothesis that parameters β_1 and β_4 are jointly zero. We conclude that the corresponding predictor variables are irrelevant (and that the insignificance of the individual t-tests is not due to collinearity between these two predictors).

The incremental R^2-test is one of the statistical tests that can be used for model selection. We discuss other statistical tests, for "nested models", in Sect. 5.6.

5.4.2 Pooling Tests

The F-test is also used to test whether data can be pooled or not across cross sections (see Sects. 2.7 and 4.5). To this end we define parameter homogeneity as our null hypothesis. To illustrate this for the Verhouten example: the null hypothesis specifies that

$$H_0 : \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix}$$

for all cross sections $i \neq j$ [Eq. (4.45)]. The alternative hypothesis is:

$$H_A : \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \neq \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix}$$

for at least some combination of cross sections i and j ($i \neq j$). The classical test is often referred to as the Chow test.[14] The Chow-test is an F-test with degrees of freedom ν_1 and ν_2:

$$F_{\nu_1,\nu_2} \sim \frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/\nu_1}{\mathrm{RSS}_2/\nu_2} \tag{5.28}$$

where

RSS_1 = residual sum of squares of the pooled regression [Eq. (4.45)],

RSS_2 = the sum of the residual sum of squares from the unit-by-unit regressions [Eq. (4.44)],

ν_1 = difference in degrees of freedom between the pooled regression and the unit-by-unit regressions,

[14] Chow (1960); Fisher (1970).

ν_2 = total degrees of freedom unused from the unit-by-unit regressions.

If the null hypothesis of parameter homogeneity is not rejected, pooling the observations is statistically justified. The degrees of freedom can be calculated as the total number of observations minus the total number of parameters. Assuming that there are n cross sections, each with T observations, the total number of observations equals $n \times T$. For fully pooled model with K explanatory variables and one constant, we estimate in total $K+1$ parameters. Hence the degrees of freedom for a fully pooled model equals $nT - K - 1$. The total degrees of freedom for the unit-by-unit models (ν_2) can be calculated as follows. The number of degrees of freedom per cross-sectional unit equals $T - K - 1$, so that in total there are $\nu_2 = n(T - K - 1)$ degrees of freedom for the unit-by-unit approach. Consequently, we have for ν_1:

$$\nu_1 = (nT - K - 1) - n(T - K - 1) = (n-1)(K+1). \tag{5.29}$$

If pooling all data and assuming homogeneity of all parameters is rejected, we may want to compare the estimation of unit-by-unit models (4.44) with partially pooled models, e.g. OLSDV models (4.47). The Chow-test can also be used in this case.

A comparison of OLSDV against the estimation of separate parameter vectors results in the following degrees of freedom for the numerator in (5.28):

$$\nu_1 = (nT - n - K) - n(T - K - 1) = (n-1)K. \tag{5.30}$$

In this case, RSS_1 is the residual sum of squares of the OLSDV model (4.47).

Another test to determine whether pooling of "individual" outcomes is allowed or not is the Roy–Zellner test (Roy 1957; Zellner 1962). This test is recommended in cases with a possibility of heteroscedasticity (Baltagi 2008, p. 61). We return to an application in Sect. 7.3.2.3.

5.4.3 *Causality Tests*

Causality tests[15] are used to:

- distinguish between causal and noncausal relations or associations;
- establish the direction of causality when variables are related;
- reduce a large set of potential predictors so that multivariate models become identifiable (based on bivariate tests).

[15]This section is based on Bult et al. (1997). For an extensive discussion of causality tests see the special issues of the *Journal of Econometrics*, vol 39 (1–2). For overviews of causality tests applied in marketing see Hanssens et al. (2001, p. 314). We note that we use the concept of "causality" that can be tested with statistical methods. This concept is not based on cause and effect in a strict philosophical sense (see Judge et al. 1985, p. 667).

We distinguish bivariate and multivariate causality tests. Although multivariate tests are preferred for multivariate models, both bivariate- and multivariate tests can be used to identify causal relations and to distinguish causality from association. However, if the objective is to reduce the set of potential predictor variables (given insufficient degrees of freedom) only bivariate tests can be used. In this section we introduce several bivariate tests.[16]

Modern tests of causality are based on the following definition of causality proposed by Granger (1969):

> "A variable x is said to cause another variable y, with respect to a given information set containing x and y, if future y-values can be predicted better using past values of x and y than using the past of y alone."

To formalize, suppose a marketing system is defined by the two-variable information set (x, y). The variable x is said to Granger cause y if the (one-step) expected quadratic forecasting error Q of using the bivariate model, is smaller than the Q of the univariate model, for at least one t (Judge et al. 1985, p. 667).

$$Q(y_{t+1} \mid y_1, \ldots, y_{t-P}, x_t, \ldots, x_{t-P^*}) < Q(y_{t+1} \mid y_t, \ldots, y_{t-P}) \tag{5.31}$$

where P and P^* are positive integers, indicating the maximum memory length in x and y. If (5.31) is not true, y is not Granger caused by x. If x causes y and y causes x then (x, y) is a feedback system.

We now introduce five bivariate causality tests,[17] including two Granger tests and two Sims tests which are regression methods. The *Granger tests* are based on the following model:

$$y_t = \sum_{i=1}^{P} \pi_{11,i} y_{t-i} + \sum_{i=1}^{P} \pi_{12,i} x_{t-i} + \varepsilon_t, \quad t = P+1, \ldots, T \tag{5.32}$$

where

$$\begin{aligned} \pi_{11,i},\ \pi_{12,i} &= \text{parameters,} \\ \varepsilon_t &= \text{a serially independent random disturbance term} \\ &\quad \text{from a distribution with mean zero and covariance} \\ &\quad \text{matrix } \Sigma. \end{aligned}$$

It is assumed that all polynomials have the same order P. If x_t does not cause y_t, $\pi_{12,i} = 0$ for $i = 1, \ldots, P$, or

[16] Multivariate causality tests, such as the VARMA-model (Judge et al. 1985, p. 667), can be used to detect causality or to determine the direction of causality in a fully specified model. For an application of the VARMA-model see Boudjellaba et al. (1992).

[17] Analytical arguments for these tests can be found in for example Chow (1983).

$$y_t = \sum_{i=1}^{P} \pi_{11,i} y_{t-i} + \varepsilon_t^* \tag{5.33}$$

where

$$\varepsilon_t^* = \text{a random disturbance term.}$$

In both Granger tests (see below) the null hypothesis is that $\pi_{12,i} = 0$, $i = 1, \ldots, P$ in (5.32).

The *Granger–Sargent* test has the following form:

$$\text{GS} = \frac{(\text{RRSS} - \text{URSS})/P}{\text{URSS}/(T - 2P)} \tag{5.34}$$

where

RRSS = the residual sum of squares of the restricted relation (5.33),

URSS = the residual sum of squares of the unrestricted relation (5.32).

Under the null hypothesis and the assumption that the disturbances are normally distributed with mean zero and variance σ^2, the statistic GS is distributed as an F-random variable with P and $(T - 2P)$ degrees of freedom.

The *Granger–Wald* test has an asymptotic χ_P^2 distribution under the null hypothesis (and is the asymptotic equivalent of the *Granger–Sargent* F-test):

$$\text{GW} = T \frac{\hat{\sigma}^2_{\varepsilon_t^*} - \hat{\sigma}^2_{\varepsilon_t}}{\hat{\sigma}^2_{\varepsilon_t}} \tag{5.35}$$

where

$\hat{\sigma}^2_{\varepsilon_t^*}$ = the estimate of $Var(\varepsilon_t^*)$ in (5.33), and

$\hat{\sigma}^2_{\varepsilon_t}$ = the estimate of $Var(\varepsilon_t)$ in (5.32).

The *Sims Methods* regress x_t on past, current and future y_t's. Sims (1972) showed that under the hypothesis of no causality from x to y, the regression parameters corresponding to future y_t's are equal to zero. The significance of the coefficients is tested using:

$$x_t = \sum_{i=-M}^{N} v_i y_{t-i} + \varepsilon_t, \quad t = N+1, \ldots, T-M \tag{5.36}$$

where

$$\varepsilon_t = \text{a random disturbance term,}$$

$$M, N = \text{the maximum number of "future" and "past" } y_t\text{'s.}$$

The null hypothesis is $v_i = 0$, $i = -M, \ldots, -1$. Relation (5.36) is estimated in constrained ($v_i = 0$, $i = -M, \ldots, -1$) and unconstrained form. Under the null hypothesis and the usual assumptions the statistic SI is distributed as an F-random variable with M and $T - M - N - 1$ degrees of freedom:

$$\text{SI} = \frac{(\text{RRSS} - \text{URSS})/M}{\text{URSS}/(T - M - N - 1)}. \tag{5.37}$$

One difficulty with the SI test is that the disturbance term ε_t in (5.36) is in general serially correlated, and consequently (5.37) does not have the claimed distribution if the null hypothesis is true (e.g. Geweke et al. 1983). To circumvent this problem a *Modified Sims* (MS) test was developed. In MS lagged values of x_t and y_t are included in the equations:

$$x_t = \sum_{i=1}^{P} \gamma_i x_{t-i} + \sum_{i=-M}^{N+P} v_i y_{t-i} + w_t, \tag{5.38}$$
$$t = N + P + 1, \ldots, T - M$$

where

$$w_t = \text{a disturbance term.}$$

Relation (5.38) is also estimated in constrained ($v_i = 0, i = -M, \cdots, -1$) and unconstrained form. The statistic MS is distributed under the null hypothesis as an F random variable with M and $(T - 2P - M - N - 1)$ degrees of freedom:

$$\text{MS} = \frac{(\text{RRSS} - \text{URSS})/M}{\text{URSS}/(T - 2P - M - N - 1)}. \tag{5.39}$$

A fifth test is the double-prewhitening method or the *Haugh–Pierce* test (Haugh 1976; Pierce 1977; Pierce and Haugh 1977).[18] The direction of causality between y_t and x_t is established by cross correlating the residuals of the univariate models fitted to each. These residuals, say $\hat{u}_t$ and $\hat{v}_t$, are causally related in the same way

[18]This test has been used in a marketing context by, for example, Hanssens (1980a,b) and Leeflang and Wittink (1992).

as y and x. Therefore, causality can be detected by estimating the parameters in the regression of u_t on past, current and future v_t's in the same manner as would be done by regressing y_t on past, current and future x_t's. The difference is that the residuals $\hat{u}_t$ and $\hat{v}_t$ are estimated by applying Box–Jenkins techniques to y_t and x_t. These techniques are discussed in Volume II.

There may be little consistency between the outcomes of the different causality tests. Tests which are only asymptotically equivalent can give very different results in small samples. For one thing, the tests differ in the effective number of data points available for analyses. These consequences increase in severity the smaller the sample size. For more details and recommended tests see Bult et al. (1997).

5.5 Face Validity

Face validity relates to the believability of a model's structure and its output, or the validity at face value. Face validity is based on theoretical and common-sense expectations, and on broadly accepted previous empirical results. This *prior knowledge* can be put to work in various ways: in structuring the model, in selecting appropriate estimation methods and in benchmarking the results with new data.

The use of face validity as a criterion depends, of course, on the model builder's prior knowledge with regard to the phenomena under study. This knowledge should guide model specification (e.g. relevant variables, operationalization of measures, functional forms). The model structure that is ultimately subjected to estimation can therefore also be evaluated in terms of face validity. The bases for knowledge about marketing phenomena include theories developed in relevant disciplines, such as economics and psychology.[19] Additional theoretical knowledge is generated within the field of marketing. And a large body of empirical results, much of it generated within the marketing field, can be used to determine the face validity of new empirical results. Of course, prior knowledge should not stifle the acceptance of surprising results. Current thinking about (marketing) problems is incomplete and sometimes incorrect. One purpose of model building is to test the prevailing theories.

Consider a model of (unit) sales for a brand, as a function of predictor variables such as own-brand price and own-brand advertising. Except for unusual cases, for example if price is interpreted by consumers as an indicator of quality, we expect price to have a negative effect on sales. Advertising, on the other hand is expected to have a positive effect on sales. Thus, ordinarily we are suspicious of empirical results that show "wrong" signs for the (partial) slope coefficients.[20]

Given a relatively large number of empirical studies in which price and advertising effects have been reported, and the meta-analyses completed on estimated

[19] See for example Alba (2011).

[20] The criticism will be muted if a coefficient with a sign opposite to expectations is not statistically significant. If the sign is "wrong" and the coefficient is significantly different from zero, we would suspect model misspecification.

effects, the argument has been made that today newly developed results should be tested against the average effects computed from published studies. The focus then shifts to a statistical test of the difference from the prevailing average effect, and the face validity of a significant difference in one direction, as we discussed in Sect. 4.3.2. In this sense, the basis for face validity considerations about short run advertising effects can be the average short run advertising elasticity of 0.12 reported by Sethuraman et al. (2011).

5.6 Model Selection

5.6.1 Introduction

The face validation criteria discussed in the previous section may allow a model builder to reject some models. However, it is common for more than one model specification to produce plausible results. Thus, face validity considerations may not suffice when the desired result is to identify one "best" model. If neither theoretical arguments pertaining to model structure, nor the violation of error-term assumptions, nor face validity considerations allow the model builder to reject all but one of the alternative models, we use explicit model comparisons.

There are at least two ways in which explicit comparisons can be made between alternative models. In this section we discuss statistical methods that allow the model builder to test hypotheses regarding relative model performance in the estimation sample. We introduce nested model comparisons and applicable statistical tests in Sect. 5.6.2. In Sect. 5.6.3 we discuss nonnested model comparisons, which are especially relevant for tests of alternative functional forms.

5.6.2 Nested Models

Nesting is a means of comparing alternative specifications where the parameters of the so-called lower-order equations are contained within the parameter space of higher-order ones. We use the dynamic models discussed in Sect. 2.8 to illustrate the concept of "nesting".

We first show the structure of a higher-order model, viz. the Partial Adjustment Autoregressive model (PAA model), which is a partial adjustment model [see Eq. (2.43)] with autocorrelated disturbances. The model was calibrated by Leeflang et al. (1992) with sales and promotion data for a pharmaceutical brand in the British market. The data stem from the "hypnotics and sedatives" segment of the pharmaceutical market. The marketing activity consisted of face-to-face communications "detailing" and impersonal media activity ("advertising"). Sales figures were based on the volume of drugs sold on prescription by pharmacists. The detailing variable

measured the time spent by pharmaceutical sales representatives promoting the drugs. Magazine advertising was the cost of ads at the time of publication and direct mailing efforts based on the cost at the day of mailing.

In all models, the market share in period t (m_t) is related to shares of detailing effort (ds_t) and shares of journal and direct mail advertising (as_t). All variables were expressed in terms of deviations from the mean. As a consequence the models have zero intercepts. The PAA model has the following structure:

$$m_t = \beta_1 ds_t + \beta_2 as_t + \lambda m_{t-1} + u_t \tag{5.40}$$

where

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

u_t = a disturbance term,

ρ = the autocorrelation parameter, and

ε_t = an error term.

Introducing the backward shift operator L, where $L(m_t) = m_{t-1}$ and substituting this operator in (5.40) we get:

$$m_t - \lambda L(m_t) = \beta_1 ds_t + \beta_2 as_t + \frac{\varepsilon_t}{1 - \rho L} \tag{5.41}$$

or

$$m_t = \frac{\beta_1 ds_t + \beta_2 as_t}{(1 - \lambda L)} + \frac{\varepsilon_t}{(1 - \lambda L)(1 - \rho L)}. \tag{5.42}$$

The following dynamic models are nested in (5.42):

- Partial Adjustment model (PA) in which there is no autocorrelation, i.e.: $\rho = 0$ while $\lambda \neq 0$ [see (2.43)];
- the Current-Effects Autoregressive model (CEA), in which $\lambda = 0$ but $\rho \neq 0$ [compare (2.42)];
- the Current-Effects model (CE), in which $\rho = \lambda = 0$ [compare (2.40)].

We show the relations between the various models (CE, CEA, PA, and PAA) in a nesting scheme, in Fig. 5.5.

Table 5.4 contains the regression coefficients, t-statistics (in parentheses), the coefficients of determination (R^2), the Durbin–Watson statistics and Durbin's h-statistics for the Leeflang et al. (1992) application. The coefficients in all four models are statistically significantly different from zero. The much stronger explanatory power (R^2 value) and the relevance of the lagged endogenous variable suggest that the partial adjustment (PA and PAA) models are better than the current effect models (CE and CEA). The estimated adjustment rates $\hat{\lambda}$, in the partial adjustment models, suggest the presence of delayed response effects.

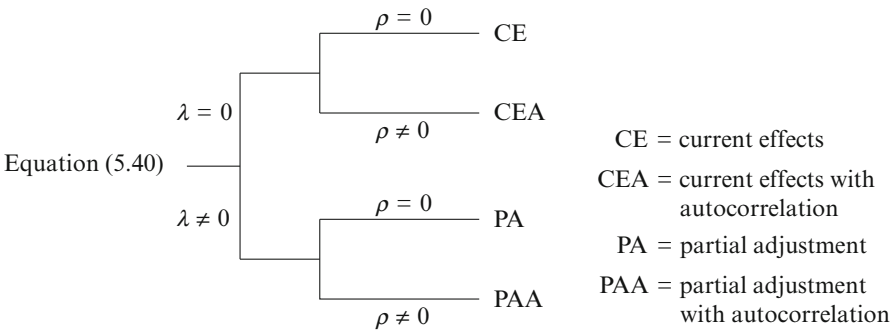

Fig. 5.5 Nesting scheme

Table 5.4 Model estimates and test results (*t*-statistics in parentheses)

Model	Detailing effort $\hat{\beta}_1$	Advertising effort $\hat{\beta}_2$	Lagged end.var $\hat{\lambda}$	Auto-corr. coeff $\hat{\rho}$	R^2	DW statistic	Durbin's *h*-statistic
CE	0.22	0.07	–	0.48	0.78	1.05	–
	(4.75)	(2.84)					
CEA	0.17	0.06	–	0.34	0.64	1.21	–
	(3.62)	(2.96)					
PA	0.07	0.04	0.68	0.06	0.94	2.01	−0.04
	(2.15)	(3.45)	(9.00)				
PAA	0.07	0.04	0.68	0.05	0.94	1.95	0.15
	(2.22)	(3.51)	(9.27)				

Source: Leeflang et al. (1992, p. 279)

It is interesting that between the two current effects models, the specification with autocorrelated errors (CEA) shows a substantial degree of first-order autocorrelation. Both models also have significant DW statistics. On the other hand, when the lagged endogenous variable is incorporated, the autocorrelation coefficient is very close to zero. Interestingly, the estimated current effects of both detailing and advertising are much smaller in the PA and PAA models, and the reduction in the detailing effect is especially large (both in an absolute and in a relative sense). The substantive conclusions differ greatly between, for example, the CEA and PA models.

We also note several other interesting aspects in Table 5.4. We have argued that the presence of first-order autocorrelation in the disturbance should rarely be treated through the addition of an autocorrelation parameter. If this were an appropriate remedy, then the Durbin–Watson (DW) statistic should be close to two in the CEA model. Since the DW statistic is not much different from its value in the CE model, it appears that the systematic pattern in the disturbances is not well captured by a first-order autocorrelated error structure.

The other aspect worth mentioning is that the R^2 value for CEA is lower than the value for CE. This may seem surprising given that the CEA model includes

an additional parameter (for the autocorrelation). The reason for this "anomaly" is that OLS maximizes R^2. Thus, even if there are statistical arguments that favor the use of an another estimation method (see Sect. 6.2) over OLS, the fit of the model, expressed in terms of the *original* criterion variable, cannot improve.

Although the results in Table 5.4 clearly favor the PA model (the autocorrelation coefficient is irrelevant) we now formalize the nested model comparisons. Each pair of nested models can be tested with a *likelihood ratio test* (statistic):

$$\eta = [\hat{\Sigma}_0/\hat{\Sigma}_1]^{-T/2} \tag{5.43}$$

where

$$\begin{aligned} \hat{\Sigma}_0 &= \text{the residual sum of squares of a restricted model} \\ &\quad \text{associated with the null hypothesis } (H_0), \\ \hat{\Sigma}_1 &= \text{the residual sum of squares of a less restricted model} \\ &\quad \text{associated with the alternative hypothesis } (H_1), \\ T &= \text{the number of observations.} \end{aligned}$$

The test statistic can be written as:

$$-2\ln\eta = T\ln\hat{\Sigma}_0 - T\ln\hat{\Sigma}_1 \tag{5.44}$$

which is asymptotically chi-squared distributed with $(p_1 - p_0)$ degrees of freedom (Greene 2012, pp. 566–567) and where:

$$\begin{aligned} p_0 &= \text{number of parameters in the restricted model, and} \\ p_1 &= \text{number of parameters in the "unrestricted" model.} \end{aligned}$$

Table 5.5 contains comparisons of the specifications for the Leeflang et al. (1992) application.

Under classical hypothesis testing a more complex (less restricted) model is not chosen unless there is statistically significant evidence in its favor.[21] The first pair of models in Table 5.5 has the current-effects (CE) model as the restricted model (H_0) and the current-effects autocorrelation model (CEA) as the less restricted alternative (H_1). The associated chi-squared statistic is significant at the 1 %-level, implying that the autoregressive current effects model is superior to the current effects model. By the same logic the second test result suggests that the autoregressive partial adjustment model (PAA) is superior to the autoregressive current-effect (CEA) model. However, the null hypothesis cannot be rejected based on the test of the

[21] Compare (Rust et al. 1995a, 1955b).

Table 5.5 Likelihood ratio test statistics

Restricted model (H_0)		Unrestricted model (H_1)		χ^2 test
Model	$-T\ln\hat{\Sigma}_0$	Model	$-T\ln\hat{\Sigma}_1$	$-2\ln\eta$
CE	-137.2	CEA	-144.3	7.1[a]
CEA	-144.3	PAA	-181.5	37.2[a]
PA	-181.4	PAA	-181.5	0.1

[a]Statistically significant at the 1 %-level
Source: Leeflang et al. (1992, p. 280)

partial adjustment model (PA) against its autoregressive counterpart (PAA). The partial adjustment (PA) model therefore emerges as the favored specification.

Other statistical tests for nested models include:

- the F-test of incremental explanatory power of an unrestricted model[22];
- the t-test (e.g. on λ, the lagged endogenous variable parameter in the PA model).

Both tests are appropriate if the predictor variables of the restricted model constitute a subset of the predictor variables of the unrestricted model (specifically one fewer in case of the t-test). If we apply this test on the model pair PA and CE, using the result in Table 5.5, we conclude that the PA model should be preferred over the CE model because $\hat{\lambda}$ is (highly) significant.

5.6.3 Non-nested Models

Non-nested models may include the same predictor variables, or they may involve some variables that are unique to each model. Suppose we consider the following multiplicative model:

$$\ln m_t = \alpha_1 \ln ds_t + \alpha_2 \ln as_t + \alpha_3 \ln m_{t-1} + v_t \tag{5.45}$$

where v_t = a disturbance term, and all other variables are defined above.

Models (5.40) and (5.45) use the same variables but have different functional forms. It is impossible to express model (5.45) as a constrained version of model (5.40), or vice versa, which renders the nested model tests inapplicable. For non-nested model comparisons and for models estimated with maximum likelihood, we can use information criteria such as the Akaike Information Criterion (AIC), the Schwarz criterion, Consistent AIC (CAIC), and a criterion developed by Allenby (1990). These criteria are full sample criteria (implying that there is no data splitting required as is often done for predictive validation as we discuss in Sect. 5.7). The information criteria seek to incorporate, in model selection, the divergent considerations of accuracy of estimation and the "best" approximation to reality.

[22]See, for example, Foekens et al. (1999) and the discussion in Sect. 5.4.1.

The statistics incorporate a measure of the precision of the estimate and a measure of parsimony in the parameterization of a model.

Akaike (1974) proposed a simple model comparison criterion, based on an information theoretic rationale. The precision of the estimated model can be represented by $\ln L$, the natural logarithm of the *likelihood* L. We discuss the likelihood concept in more detail in Sect. 6.4. For a given data set a higher $\ln L$ indicates a better fitting model. Similar to the adjusted R^2, Akaike proposed to penalize $\ln L$ for lack of parsimony, by subtracting the number of parameters from $\ln L$:

$$A = \ln L - (\text{number of parameters}) \tag{5.46}$$

where A = Akaike's criterion. A more common alternative expression of Akaike's criterion that embodies the same principle is Akaike's Information Criterion[23]:

$$AIC = -2A. \tag{5.47}$$

AIC is implemented by default in the output of many statistical software programs. When comparing different (nested or nonnested models), the model with lowest AIC is the preferred model.

Schwartz (1978) criticized AIC for being asymptotically nonoptimal. He proposed a revised form of the penalty function as follows:

$$SC = -2\ln L + \ln T \times (\text{number of parameters}) \tag{5.48}$$

where SC = Schwarz Criterion, T = the number of observations. SC is also known as the Bayesian Information Criterion (BIC). The parameter penalty of SC and BIC is larger, so that model selection based on this criterion leads to a preference for more parsimonious models.

Bozdogan (1987) proposed a criterion which penalizes overparametarization even more strongly. He proposed the Consistent AIC (CAIC), which is computed as:

$$CAIC = -2\ln L + (\text{number of parameters}) \times (\ln T + 1) \tag{5.49}$$

where T = the number of observations. Rust et al. (1955b) examined which of these and other model selection criteria perform best in selecting the current model based on simulated data. In their study the Schwarz criterion was the single best selection criterion. Due to the high correlation between the results for alternative model selection criteria they also suggest that the use of multiple model selection criteria may be unwarranted. However, as with all simulation studies, the generalizability of these results remains to be determined.

Another option for the selection of a functional form is the Box and Cox (1964) transformation, which includes linearity as a special case. The Box–Cox functional

[23] Other reformulations are found in Akaike (1981).

form can be written as[24]:

$$y_t^{(\lambda)} = \alpha + \beta_1 x_{1t}^{(\lambda)} + \beta_2 x_{2t}^{(\lambda)} + \cdots + \beta_K x_{Kt}^{(\lambda)} + \varepsilon_t \tag{5.50}$$

where

$$y_t^{(\lambda)} = \begin{cases} \dfrac{y_t^{\lambda} - 1}{\lambda} \text{ for } \lambda \neq 0 \\ \ln y_t \text{ for } \lambda = 0, \end{cases}$$

$$x_{1t}^{(\lambda)}, \ldots, x_{Kt}^{(\lambda)} = \text{predictor variables defined similarly,}$$

$$\varepsilon_t = \text{a disturbance term.}$$

Relation (5.50) includes as special cases the models:

$$y_t = \alpha^* + \beta_1 x_{1t} + \cdots + \beta_K x_{Kt} + \varepsilon_t, \quad \text{for } \lambda = 1 \tag{5.51}$$

where

$$\alpha^* = 1 + \alpha - \sum_{k=1}^{K} \beta_k$$

and

$$\ln y_t = \alpha + \beta_1 \ln x_{1t} + \cdots + \beta_K \ln x_{Kt} + \varepsilon_t, \quad \text{for } \lambda = 0. \tag{5.52}$$

Although a family of functions is defined by (5.50), a conventional likelihood ratio hypothesis test of linearity involves the null hypothesis $H_0: \lambda = 1$ versus the alternative $H_1: \lambda \neq 1$.

In cases where the models are non-nested *and* differ in the set of variables, other tests are required. Examples are the J-test, the P–E-test and a likelihood ratio test developed by Vuong (1989). These tests are described in, for example, Cameron and Trivedi (2009, pp. 279–284).

5.7 Predictive Validity

We have argued that if a model is built primarily for descriptive purposes, it should have face validity. If its primary purpose is for predictions, one might argue that face validity is not required. It is well known, for example, that models lacking

[24] See also Sect. 5.2.4.

descriptive value can provide accurate forecasts. However, the marketing models that are the topic in this book should be especially useful for *conditional* forecasting purposes. Ideally, a model user specifies alternative marketing programs, and for each possible program obtains an accurate, model-based forecast of consumers' purchase behavior. This reflects the idea that a marketing manager's environment is to some extent controllable.

In the marketing literature, it is very common for model builders to use some form of predictive validity as a statement of the model's usefulness. However, there is no established standard for the conduct of predictive validation. In this section we provide a perspective on this popular form of model validation.

Econometricians have derived the statistical properties of model-based forecasts, given error-term assumptions. For example, if the simple, linear model applies:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \quad t = 1 \ldots, T \tag{5.53}$$

we can use the estimated parameters and a specific predictor variable value, say x_0, to obtain the (conditional) forecast:

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0. \tag{5.54}$$

If the error term assumptions hold, we can construct a confidence interval for the unknown value y_0, given x_0[25]:

$$\hat{y}_0 \pm t_c^{T-2} s_{\hat{y}_0} \tag{5.55}$$

where

$$s_{\hat{y}_0} = s \sqrt{1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum_{t=1}^{T} (x_t - \bar{x})^2}},$$

t_c^{T-2} = the tabulated value of the t distribution, corresponding to the desired degree of confidence (c) and the model's degrees of freedom ($T-2$), and

s = the estimated value of the standard deviation of y.

If the model is accurate, and generally applicable, then the confidence intervals so constructed should contain the actual y_0 values a specified percentage of the time (the percentage being equal to the degree of confidence c).

If the model is estimated with *cross-sectional data*, it is common for model builders to split the data randomly into two samples. The first (analysis or

[25] See Wittink (1988, p. 47).

estimation) sample is used to estimate parameters, to test the error term assumptions, etc. The second (holdout or validation) sample is used to quantify the predictive validity (we discuss predictive validity measures below).

The question we pose is whether randomly splitting cross-sectional data into estimation and validation samples provides a true opportunity for the model builder to predictively validate the model. If we have built a model that is missing a critical variable so that one (or more) parameter estimates is (are) biased, will the predictive validity results be much poorer than the estimation results (e.g. will the predictive validity results suggest that the estimated model is inadequate)? Unfortunately, this is very unlikely in the case of random splitting. The reason is that estimation and validation samples are expected to have the same data characteristics. The bias in the effect of one or more included predictor variables will not reduce the model's predictive performance (relative to the model fit in estimation), if the correlation between included and excluded variables is the same in the estimation and validation samples. Thus, in case of randomly splitting the data there is no opportunity to *invalidate* the estimated model.

If the model builder uses *time-series data*, the situation is somewhat improved. Specifically, it is rare for model builders to split time-series randomly. There are several reasons for this. One is that the model may contain a lagged criterion variable in which case it is important to maintain the time sequence in the data. Another is that the time sequence in the data can be exploited in tests of autocorrelation in the error term. In addition, it is useful to examine a model's predictive validity to future time periods. Thus, if there are 2 years of, say, weekly data, the model builder may use the first year for estimation and the second year for validation. Time-series data can then provide an (implicit) opportunity to the model builder to check whether the results apply to a new time period. However, if the 2 years are very similar in data characteristics, for example if the market environment has not changed, then the validation exercise resembles the random splitting of cross-sectional data procedure. Thus, the larger the changes in the environment over time, the more powerful the validation exercise. At the same time, the larger the changes, the more likely it is that weaknesses in the model reduce the predictive validity. This suggests that model builders should at least report how the validation sample characteristics differ from the estimation sample.

Time-series data also provide other useful options. For example, model users may insist on evidence that a proposed model outperforms some benchmark. Brodie and De Kluyver (1987), Alsem et al. (1989) and Brodie et al. (2001) compared the performance of marketing-mix models to that of a naive model, which predicts next period's value for the criterion variable to be this period's actual value. One might argue that little faith should be placed in the parameter estimates, if the proposed model does not outperform a (naive) model that lacks structural characteristics consistent with marketing traditions. As shown by Foekens et al. (1994), the relative performance also depends on the extent to which the *characteristics of the data change between estimation* and *validation samples*.

Foekens et al. (1994) argue the following:

- It is well known that the uncertainty of individual predictions is influenced by the distance between the *predictor variables' values* and the *sample means*. For a new set of data this idea is captured by the change in average values in the predictor variables between estimation and validation sample data.
- Another way in which the structural similarity of estimation and validation data can be determined is by the correlation matrix for the predictor variables. For example, the forecasting accuracy of a misspecified model may not be reduced if the correlations remain the same. The greater the change in *correlations* the more likely it is to obtain poor predictive accuracy for misspecified models.
- One can determine the extent to which the same model, applied separately to estimation and validation data, produces different *parameter estimates*. Substantial differences may occur for several reasons. One is that the true parameters have changed. Another is that the estimated parameters differ because the model is misspecified, and this misspecification differentially affects the parameter estimates in the two samples.
- Additionally, the predictive validity will be affected by a difference in error variance between estimation and validation samples.

We now consider accuracy measures that can be used to assess predictive validity. To formalize, suppose we have T observations in total, and use the first T^* observations for estimation, leaving $(T - T^*)$ for validation. Thus, the unknown parameters $\alpha, \beta_1, \ldots, \beta_K$ in a multiple regression model are estimated using T^* observations. Substituting the estimates $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_K$ and using the values of $x_{1t}, x_{2t}, \ldots, x_{Kt}$, for $t = T^* + 1, T^* + 2, \ldots, T$, the following predicted values of y_t are obtained:

$$\hat{y}_t = \hat{\alpha} + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \cdots + \hat{\beta}_K x_{Kt}, \quad t = T^* + 1, \ldots, T. \tag{5.56}$$

Comparing the predicted values $\hat{y}_t$ with the actual values of y_t, $t = T^* + 1, \ldots, T$, the predictive validity of the relation can be determined. To test for a lack of bias we can use the *Average Prediction Error* (APE):

$$\text{APE} = \frac{\sum_{t=T^*+1}^{T} (y_t - \hat{y}_t)}{T - T^*}. \tag{5.57}$$

Note that the denominator in (5.57) is the number of observations in the validation sample. Also, positive and negative errors are allowed to offset each other. Thus, we can test the null hypothesis that the mean prediction error is zero, based on the t-test for the mean. Inability to reject the null hypothesis means that there is no evidence of *bias in the predictions*.

We note that this measure of bias is not as useful as one might think at first glance. A result that APE = 0 means only that on average the actual and predicted

values are the same. Nevertheless, it is quite possible for some or all predictions to systematically deviate from the actual values. For example, a model with the wrong functional form will not produce a positive APE value as long as the positive and negative prediction errors in the validation sample offset each other. This will occur if the estimation- and validation sample data characteristics are the same.

To measure the predictive performance of the model, which depends on bias and variance, we can use the "*Average Squared Predictor Error*" (ASPE), also known as "*Mean Squared Error*" (MSE):

$$\text{ASPE} = \frac{\sum_{t=T^*+1}^{T} (y_t - \hat{y}_t)^2}{T - T^*}. \tag{5.58}$$

The use of squared terms means that large prediction errors are weighted more heavily than small errors. This measure is consistent with the least squares principle of regression analysis. However, it has the drawback that it summarizes the prediction errors in squared units. To obtain a value in the units of measurement for the criterion variable, we can take the square root of ASPE, which is denoted by RASPE, and is also known as the *Root Mean Squared Error* (RMSE):

$$\text{RASPE} = \sqrt{\frac{\sum_{t=T^*+1}^{T} (y_t - \hat{y}_t)^2}{T - T^*}}. \tag{5.59}$$

The value for RASPE can be compared to the standard deviation of residuals in the estimation sample. In general we expect the value of RASPE to be greater than the standard deviation of the residuals, the actual difference being a function of the factors we have mentioned earlier.

A predictive validity measure that is dimensionless, easy to relate to, and potentially useful if one wants to make comparisons of forecast accuracy across different settings, is the *Mean Absolute Percentage Error* (MAPE):

$$\text{MAPE} = \frac{1}{T - T^*} \sum_{t=T^*+1}^{T} \frac{|\, y_t - \hat{y}_t \,|}{y_t} \times 100\,\%. \tag{5.60}$$

In this measure absolute, rather than squared, errors are computed, and each absolute error is expressed relative to the actual value, for observation t in the validation sample. If one believes that the magnitude of an error should be considered relative to the corresponding actual value, MAPE may be a suitable measure.

Other measures express the predictive performance of a given model relative to a benchmark model. As we have mentioned earlier, one naive (benchmark) model is to predict next period's value for the criterion variable with this period's actual value. The following Relative Absolute Error (RAE) measure incorporates this idea:

$$\text{RAE} = \frac{\sum_{t=T^*+1}^{T} |\, y_t - \hat{y}_t \,|}{\sum_{t=T^*+1}^{T} |\, y_t - y_{t-1} \,|}. \tag{5.61}$$

If RAE is less than one, the model outperforms the benchmark represented by a naive model that predicts the current value by the previous value (i.e. $\hat{y}_t = y_{t-1}$.)

A measure that is conceptually similar to *RAE*, but uses squared prediction errors instead of absolute ones, is Theil's U-statistic:

$$U = \sqrt{\frac{\sum_{t=T^*+1}^{T} (y_t - \hat{y}_t)^2}{\sum_{t=T^*+1}^{T} (y_t - y_{t-1})^2}}. \tag{5.62}$$

As with the RAE measure, if Theil's U-statistic is less than one, the model generating $\hat{y}_t$ outperforms the naive model.

Theil (1965) shows how ASPE in (5.58) can be decomposed:

$$\text{ASPE} = \text{APE}^2 + (s_{\hat{y}} - r s_y)^2 + (1 - r^2) s_y^2 \tag{5.63}$$

where

$$\begin{aligned} \text{APE} &= \text{the Average Prediction Error (5.57)}, \\ s_{\hat{y}} &= \text{the standard deviation of the predicted values}, \\ s_y &= \text{the standard deviation of the actual values}, \\ r &= \text{the correlation coefficient between actual and} \\ &\quad \text{predicted values}. \end{aligned}$$

The first term in (5.63) captures the squared bias, while the second and third terms together account for the prediction error due to unreliability (variance). For a model that is linear in the original variables, both the first- and the second term are zero in the *estimation* sample. The second term captures the difference in variability for the predicted values ($s_{\hat{y}}$) and the variability for the actual values (s_y) multiplied by the correlation between actual and predicted values. The third term is the proportion of the variance in the criterion variable in the *validation* sample that is not attributable to the estimated relation.

The advantage of using such a decomposition of the prediction errors is that the model builder can diagnose the source(s) of the errors. It is, for example, very useful to distinguish between bias and variance. Consider a comparison of predictive

validity between two models, one being the "preferred" model, the other being a simplified version. One would expect that the "preferred" model has less bias but potentially more variance. The decomposition allows the model builder to separate a difference in overall performance into differences due to bias and due to other components. Importantly, the more the validation data characteristics differ from the estimation data, the greater the expected contribution of the bias component to ASPE. Once decompositions of prediction errors have been made, the natural question becomes to what extent the (validation) data provided an opportunity for the "preferred" model to show better performance than the benchmark model. For example, in a stable environment in which marketing activities show little variation, it may be difficult to beat a benchmark model that predicts next period's value to be this period's actual value. On the other hand, if there is a substantial amount of variation in marketing activities in the validation sample, it should be possible to "beat" the benchmark model. Another relevant aspect, as we mentioned earlier, is the extent to which validation sample characteristics differ from the estimation sample. The greater this difference, the stronger the opportunity to falsify a (wrong) model. These considerations suggest that model builders should at least report central tendency and dispersion measures, for both estimation and validation samples. With access to this information a user can make a judgment about two relevant aspects:

- the extent to which the validation sample allows for model falsification (performance in the validation sample relative to the estimation sample);
- the extent to which the validation sample allows the model to outperform a naive model.

5.8 Model Validation for the Verhouten Case

In this section we return to the Verhouten case. In Sect. 4.4 we presented the estimation results for a linear additive and a multiplicative model. The outcomes indicated that many of the competitive variables are not significant. Therefore, we continue in this session with a reduced model, that contain only own marketing variables. Furthermore, to avoid repetition, we will focus on the multiplicative model in this section. The steps for the linear additive model are very similar and are left as an exercise. The model that we will work with in this session is a reduced version of (2.27) and is specified as follows:

$$S_{1t} = \theta\left(P_{1t}^{\gamma_1}\gamma_2^{F_{1t}}\gamma_3^{D_{1t}}\gamma_4^{FD_{1t}}\right)\varepsilon_{1t}. \tag{5.64}$$

First we will test the six basic assumptions underlying the linear model in Sect. 5.8.1. Thereafter, in Sect. 5.8.2, we will assess the predictive validity of the model. Because our validity assessment needs a hold-out sample, we split the available data into a calibration sample (the first 60 observations) and in a validation

sample (the last 8 observations). We now cannot longer use the results in Sect. 4.4, because these were based on the full set of 68 observations.

5.8.1 Testing the Six Assumptions for the Verhouten Case

To validate the six assumption for Eq. (5.64), we first estimate the following linearized version of the model using the first 60 observations:

$$S^*_{1t} = \theta^* + \gamma_1 P^*_{1t} + \gamma^*_2 F_{1t} + \gamma^*_3 D_{1t} + \gamma^*_4 FD_{1t} + \varepsilon^*_{1t} \tag{5.65}$$

where all the variables are defined in (4.40) on p. 112. The outcomes are presented in Table 5.6. In order to test the first assumption: $\mathrm{E}(\varepsilon_t) = 0$ we plot the residuals against each predictor variable.[26] As an illustration, we plot the residuals against the Price of Verhouten in Fig. 5.6. The graph does not seem to indicate that for certain price values the residuals differ systematically from zero. To assess this assumption a bit more carefully, we employ the RESET test. We compute $\hat{y}_t$ and add the second, third, and fourth power as explanatory variables to the model in Eq. (5.65). The regression does not indicate significance for any of these terms, so that we conclude that there is no evidence of misspecification.

In order to test the second assumption: $\mathrm{Var}(\varepsilon_t) = \sigma^2$ for all t, we take another look at the plots where the residuals are depicted against each predictor variable, but now with the aim to detect changes in the variability of the residuals. The plot in Fig. 5.6 shows a somewhat larger range of the residuals at higher prices, compared to residuals at lower prices. However, because we observe more residuals at higher prices, it is also likely that we observe more extreme values and consequently, a larger range. Hence, we need to test for heteroscedasticity more formally. We cannot

Table 5.6 Estimation results of the multiplicative model for Verhouten

	Parameter estimates	Standard error	t-value	p-value
Intercept ($\hat{\theta}^*$)	6.11	0.15	39.81	0.00
ln of Price Verhouten ($\hat{\gamma}_{11}$)	−3.75	0.34	−10.94	0.00
Feature-only Verhouten ($\hat{\gamma}^*_{21}$)	0.05	0.21	0.23	0.82
Display-only Verhouten ($\hat{\gamma}^*_{31}$)	0.01	0.42	0.02	0.98
Feature and Display Verhouten ($\hat{\gamma}^*_{41}$)	1.16	0.77	1.50	0.14
Number of observations = 60,	$R^2 = 0.87$,	$R^2_a = 0.86$,	$\hat{\sigma} = 0.17$	
RSS = 1.64, ESS = 10.78,	F-value = 90.25			

[26] Note that the sum of the residuals is equal to zero by construction. Hence, any test on the average of the residuals is not very informative.

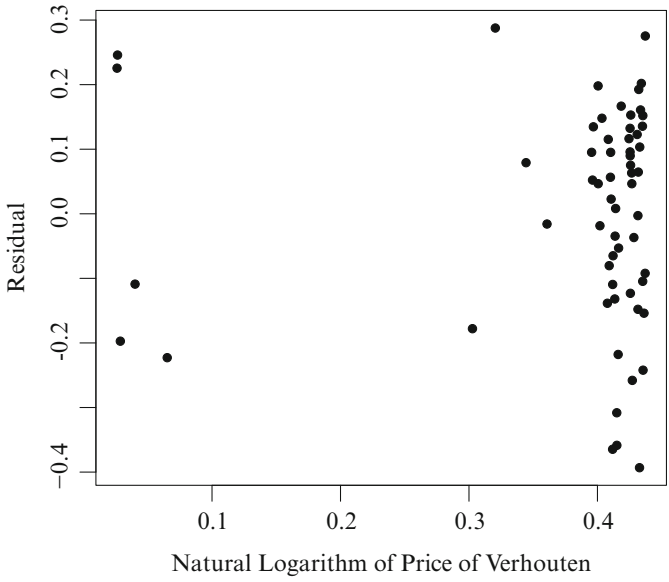

Fig. 5.6 Plot of residuals against Price of Verhouten

employ the Goldfeld–Quandt test to identify differences in variance of lower prices and higher prices, because that would involve estimating (5.65) on five observations. Instead, we utilize the Breusch–Pagan test to detect heteroscedasticity. We regress the squared residuals on the explanatory variables that appear in (5.65), and find that the regression is insignificant (the p-value associated with the test in (5.6) equals 0.68). Hence, we conclude that heteroscedasticity is not an issue the Verhouten case.

To assess autocorrelation, we plot the residuals over time, see Fig. 5.7. The residuals in Fig. 5.7 show shorter and longer runs on either side of the mean value. As this might indicate positive autocorrelation, further investigation is warranted. The value of the Durbin–Watson statistic equals 0.851, and the values for d_L and d_U with 60 observations and four regressors are $d_L = 1.444$ and $d_U = 1.727$. Since $DW < d_L$, we conclude that the residuals are positively autocorrelated.

We assess normality of the residuals graphically by comparing a histogram of the residuals with a normal curve in Fig. 5.8. The histogram indicates that the right tail is heavier than we would expect for a normal distribution. This is confirmed by the normal probability plot in Fig. 5.9. To test for nonnormality of the residuals, we employ the Kolmogorov–Smirnov test and find that the associated p-value equals 0.02. However, the Shapiro–Wilk test returns a p-value of 0.07, and Jarque–Bera's p-value equals 0.21. We conclude that there might be a mild issue with the normality of the residuals. Here we will not investigate transformations to overcome this issue. Instead, we will interpret the p-values of the outcomes with care.

As we did not yet discuss tests for endogeneity, we do not test for the fifth assumption.

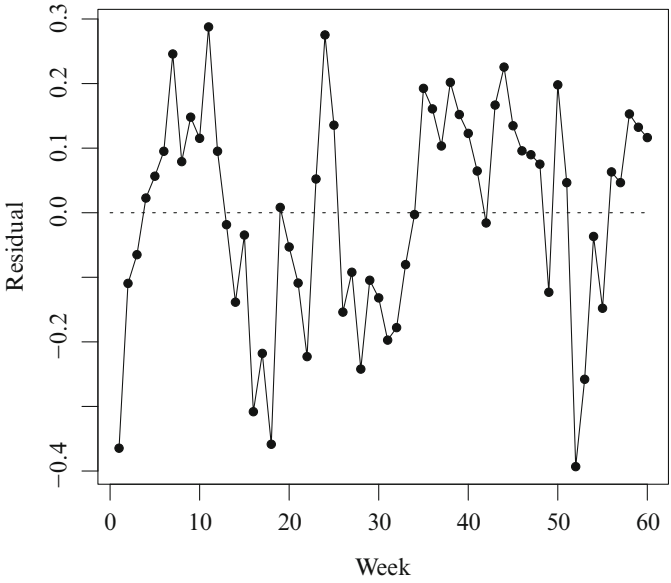

Fig. 5.7 Plot of residuals against time

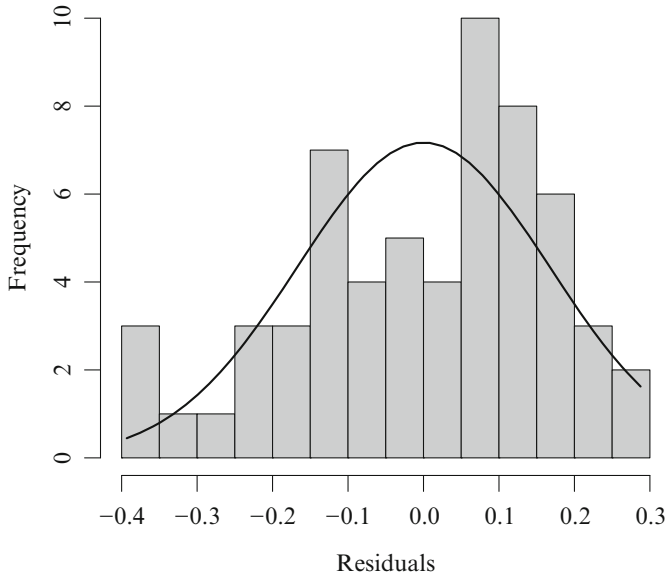

Fig. 5.8 Histogram with normal curve of the residuals

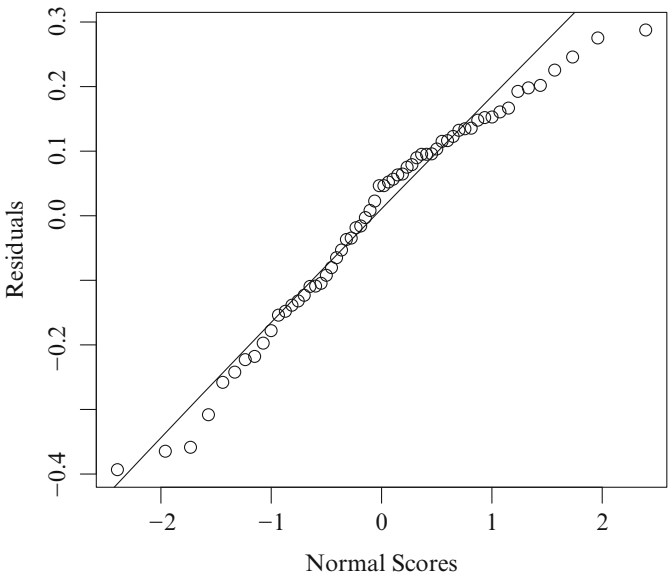

Fig. 5.9 Normal probability plot of the residuals

For the sixth assumption we inspect the correlation matrix of the explanatory variables in (5.65) to find indications for multicolinearity:

$$\begin{array}{c} \\ P^*_{it} \\ F_{it} \\ D_{it} \\ FD_{it} \end{array} \begin{array}{c} \begin{array}{cccc} P^*_{it} & F_{it} & D_{it} & FD_{it} \end{array} \\ \begin{pmatrix} 1 & -0.75 & -0.10 & -0.04 \\ -0.75 & 1 & 0.23 & 0.26 \\ -0.10 & 0.23 & 1 & 0.92 \\ -0.04 & 0.26 & 0.92 & 1 \end{pmatrix} \end{array}. \tag{5.66}$$

The correlation matrix indicates that several variables, notably P^*_{it} and F_{it}, as well as D_{it} and FD_{it} are highly correlated. This is confirmed when we investigate VIF values of the explanatory variables, see Table 5.7. The VIF values and the correlation matrix reveal that there is a multicollinearity issue with several of the independent variables in Eq. (5.65). This might be the cause of the non-significance of some of the parameters in Table 5.6. We return to this issue in Sect. 6.3.

Table 5.7 VIF values of the explanatory variables in (5.65)

	VIF
ln of Price Verhouten (P^*_{it})	2.72
Feature-only Verhouten (F_{it})	2.84
Display-only Verhouten (D_{it})	7.30
Feature and Display Verhouten (FD_{it})	7.84

5.8.2 *Assessing Predictive Validity for the Verhouten Case*

We now turn to the validation sample, which consists of the last eight observations for the Verhouten case. We enter the new observations for P^*_{1t}, F_{1t}, D_{1t} and FD_{it} together with the corresponding estimated parameters (Table 5.6) in Eq. (5.65) to compute out-of-sample predictions for S^*_{1t}. Figure 5.10 depicts these predictions (dashed line), together with the observed values for S^*_{1t}. The predictions match the observed values quite closely, indicating that the predictive validity of our model is satisfactory.

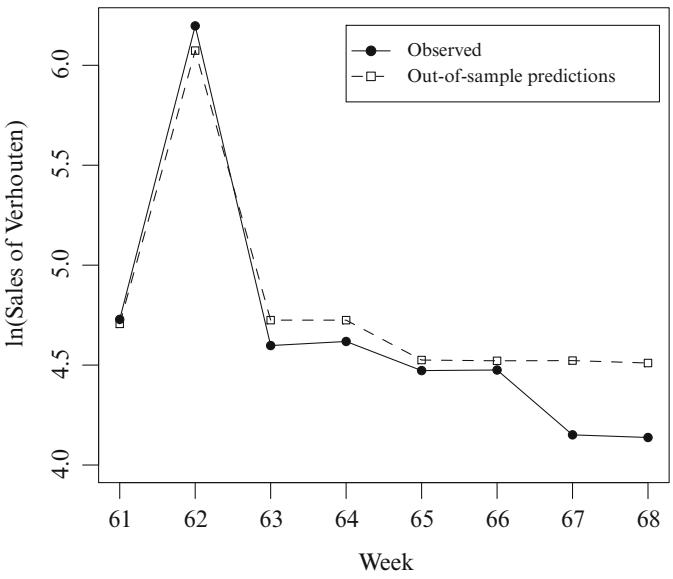

Fig. 5.10 Observations and out-of-sample predictions for S^*_{1t} in Eq. (5.65)

To assess the predictive validity of our model further we compute the APE, which equals −0.1164. This means that, on average, S^*_{1t} is slightly over-predicted in the validation sample. ASPE is computed as 0.0406, and RASPE = 0.2016. This value for RASPE indicates that, as expected, the standard deviation of the residuals in the validation sample is slightly larger than the standard deviation of the residuals in the estimation sample (compare $\hat{\sigma} = 0.17$ in Table 5.6). The MAPE value for the out-of-sample predications equals 15.3 % so that we conclude that on average the predictions for S^*_{1t} deviate about 15.3 % from the corresponding observed values.

Finally, we compare our model to a naive model where the next observation for S^*_{1t} is predicted by the current observation. A value of RAE = 0.3329 implies that our model outperforms this benchmark model. This is confirmed by Theil's U, which is computed as 0.1176.

References

Akaike, H.: A new look at statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)

Akaike, H.: Likelihood of a model and information criteria. J. Econom. **16**, 3–14 (1981)

Alba, J.W.: Consumer Insights: Findings from Behavioral Research. MSI Relevant Knowledge Series, Cambridge (2011)

Allenby, G.M.: Cross-validation, the Bayes theorem, and small sample bias. J. Bus. Econ. Stat. **8**, 171–178 (1990)

Alsem, K.J., Leeflang, P.S.H., Reuyl, J.C.: The forecasting accuracy of market share models using predicted values of competitive marketing behaviour. Int. J. Res. Mark. **6**, 183–198 (1989)

Baltagi, B.H.: Econometric Analysis of Panel Data, 4th edn. Wiley, Chichester (2008)

Baron, R.M., Kenny, D.A.: The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. **51**(6), 1173–1182 (1986)

Belsley, D.A., Kuh, E., Welsh, R.E.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York (1980)

Bera, A.K., Jarque, C.M.: An Efficient Large-Sample Test for Normality of Observations and Regression Residuals. Working Papers in Econometrics, Australian National University, No. 40, Canberra (1981)

Bera, A.K., Jarque, C.M.: Model specification tests: A simultaneous approach. J. Econom. **20**, 59–82 (1982)

Boudjellaba, H., Dufour, J.M., Roy, R.: Testing causality between two vectors in multivariate autoregressive moving average model. J. Am. Stat. Assoc. **87**, 1082–1090 (1992)

Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. Ser. B **26**, 211–252 (1964)

Bozdogan, H.: Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika **52**, 345–370 (1987)

Breusch, T.S., Pagan, A.R.: A simple test for heteroskedasticity and random coefficient variation. Econometrica **47**, 987–1007 (1979)

Brodie, R.J., Danaher, P., Kumar, V., Leeflang, P.S.H.: Econometric models for forecasting market share. In: J.S. Armstrong (ed.) Principles of Forecasting. Kluwer Academic, Hingham/Dordrecht (2001)

Brodie, R.J., De Kluyver, C.A.: A comparison of the short-term forecasting accuracy of econometric and naive extrapolation models of market share. Int. J. Forecast. **3**, 423–437 (1987)

Bronnenberg, B.J., Mahajan, V.: Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. Mark. Sci. **20**, 284–299 (2001)

Bult, J.R., Leeflang, P.S.H., Wittink, D.R.: The relative performance of bivariate causality tests in small samples. Eur. J. Oper. Res. **97**, 450–464 (1997)

Cameron, A.C., Trivedi, P.K.: Microeconometrics Using Stata. Stata Press (2009)

Card, D.: The causal effect of education on earnings. In: O. Ashenfelter, D. Card (eds.) Handbook of Labor Economics, vol. 3A. Elsevier, Amsterdam (1999)

Card, D.: The effect of unions on wage inequality in the U.S. labor market. Ind. Labor Relat. Rev. **54**(2), 296–315 (2001)

Chow, G.C.: Tests of equality between sets of coefficients in two linear regressions. Econometrica **28**, 591–605 (1960)

Chow, G.C.: Econometrics. McGraw-Hil, New York (1983)

De Vaus, D.: Surveys in Social Research, 6th edn. Allen & Unwin, Crows Nest (2013)

Durbin, J.: Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. Econometrica **38**, 410–429 (1970)

Durbin, J., Watson, C.S.: Testing for serial correlation in least squares regression, I. Biometrika **37**, 409–428 (1950)

Durbin, J., Watson, C.S.: Testing for serial correlation in least squares regression, II. Biometrika **38**, 159–178 (1951)

Ebbes, P.: Latent Instrumental Variables. Unpublished Ph.D. thesis, University of Groningen, Groningen (2004)
Ebbes, P., Wedel, M., Böckenholt, U.: Frugal IV alternatives to identify the parameters for an endogenous regressor. J. Appl. Econom. **24**, 446–468 (2009)
Ebbes, P., Wedel, M., Steerneman, T., Böckenholt, U.: Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. Quant. Mark. Econ. **3**, 365–392 (2005)
Farrar, D.E., Glauber, R.R.: Multicollinearity in regression analysis: The problem revisited. Rev. Econ. Stat. **49**, 92–107 (1967)
Fisher, F.M.: Test of equality between sets of coefficients in two linear regressions: An expository note. Econometrica **38**, 361–366 (1970)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation. Int. J. Forecast. **10**, 245–261 (1994)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: Varying parameter models to accommodate dynamic promotion effects. J. Econom. **89**, 249–268 (1999)
Friedman, R.: Multicollinearity and Ridge regression. Allgemeines Statistisches Archiv. **66**, 120–128 (1982)
Geweke, J., Meese, R., Dent, W.: Comparing alternative tests of causality in temporal systems. J. Econom. **21**, 161–194 (1983)
Goldfeld, S.M., Quandt, R.E.: Some tests for homoscedasticity. J. Am. Stat. Assoc. **60**, 539–547 (1965)
Goodman, L.A.: On the exact variance of products. J. Am. Stat. Assoc. **55**, 708–713 (1960)
Granger, C.W.J.: Investigating causal relations by econometric models and cross spectral methods. Econometrica **37**, 424–438 (1969)
Greene, W.H.: Econometric Analysis, 7th international edn. Prentice-Hall, Upper Saddle River (2012)
Hanssens, D.M.: Bivariate time series analysis of the relationship between advertising and sales. Appl. Econ. **12**, 329–340 (1980a)
Hanssens, D.M.: Market response, competitive behavior and time series analysis. J. Mark. Res. **17**, 470–485 (1980b)
Hanssens, D.M., Parsons, L.J., Schultz, R.L.: Market Response Models: Econometric and Time Series Analysis. Kluwer Academic, Boston (2001)
Hanssens, D.M., Pauwels, K., Srinivasan, S., Vanhuele, M., Yildirim, G.: Consumer attitude metrics for guiding marketing mix decisions. Mark. Sci. **33**, 534–550 (2014)
Haugh, L.D.: Checking the independence of two covariance stationary time series: A univariate residual cross correlation approach. J. Am. Stat. Assoc. **71**, 378–385 (1976)
Hausman, J.A.: Specification tests in econometrics. Econometrica **46**(6), 1251–1271 (1978)
Hayes, A.F., Scharkow, M.: The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? Psychol. Sci. **24**, 1918–1927 (2013)
Hinich, M.J., Talwar, P.P.: A simple method for robust regression. J. Am. Stat. Assoc. **70**, 113–119 (1975)
Horváth, C., Leeflang, P.S.H., Wieringa, J.E., Wittink, D.R.: Competitive reaction- and feedback effects based on VARX models of pooled store data. Int. J. Res. Mark. **22**, 415–426 (2005)
Huber, P.J.: Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Stat. **1**, 799–821 (1973)
Judge, G.G., W. E. Griffiths R. C. Hill, H.L., Lee, T.C.: The Theory and Practice of Econometrics, 2nd edn. Wiley, New York (1985)
Kaul, A., Wittink, D.R.: Empirical generalizations about the impact of advertising on price sensitivity and price. Mark. Sci. **14**, G151–G160 (1995)
Kenkel, J.L.: Some small sample properties of Durbin's test for serial correlation in regression models containing lagged dependent variables. Econometrica **42**, 763–769 (1974)
Kenny, D.: Mediation. http://davidakenny.net/cm/mediate.htm (2013) [Online; Accessed 7 Apr 2014]

Kristensen, K.: Hedonic theory, marketing research, and the analysis of complex goods. Int. J. Res. Mark. **1**, 17–36 (1984)
Kumar, T.K.: Multicollinearity in regression analysis. Rev. Econ. Stat. **57**, 365–366 (1975)
Kuskov, E., Villas-Boas, J.M.: Endogeneity and individual consumer choice. J. Mark. Res. **45**, 702–714 (2008)
Leeflang, P.S.H., Mijatovic, G.M., Saunders, J.: Identification and estimation of complex multivariate lag structures: A nesting approach. Appl. Econ. **24**, 273–283 (1992)
Leeflang, P.S.H., Wittink, D.R.: Diagnosing competitive reactions using (aggregated) scanner data. Int. J. Res. Mark. **9**, 39–57 (1992)
MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., Sheets, V.: A comparison of methods to test mediation and other intervening variable effects. Psychol. Methods **7**(1), 99–128 (2002)
MacKinnon, D.P., Lockwood, C.M., Williams, J.: Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivar. Behav. Res. **39**, 99–128 (2004)
Mahajan, V., Sharma, S., Wind, Y.: Parameter estimation in marketing models in the presence of influential response data: Robust regression and applications. J. Mark. Res. **21**, 268–277 (1984)
Mason, C., Perreault, Jr., W.D.: Collinearity, power and interpretation of multiple regression analysis. J. Mark. Res. **28**, 268–280 (1991)
Massy, W.F.: Principal component regression in exploratory statistical research. J. Am. Stat. Assoc. **60**, 234–256 (1965)
Park, S., Gupta, S.: Simulated maximum likelihood estimator for the random coefficient logit model using aggregate data. J. Mark. Res. **46**, 531–542 (2009)
Park, S., Gupta, S.: Handling endogenous regressors via joint estimation using Copulas. Mark. Sci. **31**, 567–586 (2012)
Petrin, A., Train, K.: A control function approach to endogeneity in consumer choice models. J. Mark. Res. **47**, 3–13 (2010)
Pierce, D.A.: Relationships - and the lack thereof - between economic time series, with special references to money and interest rates. J. Am. Stat. Assoc. **72**, 11–22 (1977)
Pierce, D.A., Haugh, L.D.: Causality in temporal systems. J. Econom. **5**, 265–293 (1977)
Ramsey, J.B.: Tests for specification errors in classical linear least squares regression analysis. J. R. Stat. Soc. Ser. B **31**(2), 350–371 (1969)
Ramsey, J.B.: Classical model selection through specification tests. In: P. Zarembka (ed.) Frontiers in Econometrics. Academic, New York (1974)
Rao, V.R., Wind, Y., DeSarbo, W.S.: A customized market response model: Development, estimation and empirical testing. J. Acad. Mark. Sci. **16**, 128–140 (1988)
Roy, S.N.: Some Aspects of Multivariate Analysis. Wiley, New York (1957)
Rust, R.T., Lee, C., Valente, Jr., E.: Comparing covariance structure models: A general methodology. Int. J. Res. Mark. **12**, 279–291 (1995a)
Rust, R.T., Simester, D., Brodie, R.J., Nilikant, V.: Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. Manag. Sci. **41**, 322–333 (1995b)
Schwartz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Sethuraman, R., Tellis, G.J., Briesch, R.: How well does advertising work? Generalizations from a meta-analysis of brand advertising elasticity. J. Mark. Res. **48**, 457–471 (2011)
Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality. Biometrika **52**, 591–611 (1965)
Sharma, S.: Applied Multivariate Techniques. Wiley, New York (1996)
Sims, C.A.: Money, income and causality. Am. Econ. Rev. **62**, 540–552 (1972)
Sobel, M.E.: Asymptotic confidence intervals for indirect effects in structural equation models. In: S. Leinhardt (ed.) Sociological methodology. American Sociological Association, Washington, DC (1982)
Sridhar, S., Srinivasan, R.: Social influence effects in online product ratings. J. Mark. **76**(5), 70–88 (2012)

Srinivasan, S., Vanhuele, M., Pauwels, K.: Mind-set metrics in market response models: An integrative approach. J. Mark. Res. **27**, 672–684 (2010)
Stewart, J.: Econometrics. Philip Allan, Hemel Hemstead (1991)
Theil, H.: The analysis of disturbances in regression analysis. J. Am. Stat. Assoc. **60**, 1067–1079 (1965)
Thomas, J.J., Wallis, K.F.: Seasonal variation in regression analysis. J. R. Stat. Soc. Ser. A **134**, 67–72 (1971)
Van Dijk, A., Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Similarity-based spatial models to estimate shelf space elasticities. Quant. Mark. Econ. **2**(3), 257–277 (2004)
Vuong, Q.H.: Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica **57**, 307–333 (1989)
White, H.: A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica **48**, 817–838 (1980)
Wittink, D.R.: The Application of Regression Analysis. Allyn and Bacon, Boston (1988)
Wooldridge, J.M.: Introductory Econometrics: A Modern Approach, 5th edn. Cengage Learning, Mason (2012)
Zellner, A.: An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J. Am. Stat. Assoc. **57**, 348–368 (1962)

Chapter 6
Re-estimation: Introduction to More Advanced Estimation Methods

6.1 Introduction

In this chapter we consider methods and procedures for the estimation of model parameters in cases where the basic assumptions of the general linear model are violated. When this is the case we need either other specifications and/or other estimation methods: "re-estimation". In Sect. 6.2 we introduce Generalized Least Squares (GLS) estimation methods. These models, amongst others, account for:

- heteroscedasticity (violation of assumption 2 in Sect. 4.2.3);
- autocorrelation (violation of assumption 3 in Sect. 4.2.3).

In Sect. 6.3 we revisit the Verhouten case and demonstrate how one may account for multicollinearity, autocorrelation and heteroscedasticity. In Chap. 4 and Sect. 6.2 we concentrate on least squares methods. In Sect. 6.4 we discuss other methods to obtain estimates, viz. Maximum Likelihood (ML) methods. These methods may also account for the above mentioned violations of the general linear model assumptions and for violations of the normality assumptions of the disturbance term (non-normality).

While in both Sects. 6.2 and 6.3 we assume one-way causality, we consider simultaneous causality in Sect. 6.5. The construction of a simultaneous system of equations offers opportunities to account for endogeneity. Other approaches how endogeneity can be approached such as Instrumental Variables (IV) estimation methods are discussed in Sect. 6.6. In Sect. 6.7, we discuss several tests for endogeneity. Finally, we discuss Bayesian estimation in Sect. 6.8.

In Volume II we discuss other "re-estimation" methods such as:

- non-linear estimation;
- non-parametric and semi-parametric estimation;
- generalized method of moments estimation, and
- structural equation models (SEM).

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_6

6.2 Generalized Least Squares

6.2.1 Introduction

Several of the assumptions for the application of OLS involve restrictive assumptions about the disturbance term. In this section, we discuss procedures, known as Generalized Least Squares-(GLS)-methods, that allow for more general disturbance characteristics. Specifically, GLS can accommodate violations of at least one of the assumptions 2 and 3 in Sect. 4.2.3.

Consider again the K-variable model, expressed in matrix notation, as in (4.15):

$$y = X\beta + u \tag{6.1}$$

where y and X are defined in a similar way as in Sect. 4.2.2, and we used u instead of ε to denote the disturbances to indicate that the assumptions may not all be satisfied. The variance-covariance matrix of the disturbances is now defined as:

$$\mathrm{E}(uu') = \Omega \tag{6.2}$$

where

$$\Omega = \begin{pmatrix} \omega_{11} & \cdots & \omega_{1T} \\ \vdots & \ddots & \vdots \\ \omega_{T1} & \cdots & \omega_{TT} \end{pmatrix} = \sigma^2 \Omega^*$$

which is a positive definite symmetric $T \times T$ matrix with full rank T, and where the ω_{ij} are the covariances of the disturbances. Assumptions 2 and 3 are satisfied if $\Omega^* = I$, where I is a $T \times T$ identity matrix. However, if this is not the case, we can obtain an expression for the generalized least squares estimator of β as follows.

Let the matrix V be a nonsingular $T \times T$ matrix, such that[1]:

$$V'V = (\Omega^*)^{-1} \quad \text{or} \quad (V'V)^{-1} = \Omega^*. \tag{6.3}$$

We premultiply both sides of (6.1) by V:

$$Vy = VX\beta + Vu. \tag{6.4}$$

The variance-covariance matrix of the disturbances (Vu) of (6.4) is:

$$\mathrm{E}((Vu)(Vu)') = \sigma^2 V \Omega^* V'. \tag{6.5}$$

[1]Since Ω, and thus Ω^*, is symmetric and positive definite, so are Ω^{-1} and $(\Omega^*)^{-1}$, and hence a matrix V satisfying (6.3) exists. See Appendix A.7.

Substituting (6.3) in (6.5) we obtain[2]:

$$E[(Vu)(Vu)'] = \sigma^2 V (V'V)^{-1} V' = \sigma^2 I. \tag{6.6}$$

This shows that if we transform the variables by the V-matrix in (6.4), the disturbance terms satisfy the error-term assumptions 2 and 3, i.e. they are homoscedastic and uncorrelated over time. Thus, the OLS-method could be applied to the transformed variables in (6.4). The Generalized Least Squares estimator[3] is:

$$\hat{\beta}_{\text{GLS}} = \left(X' (\Omega^*)^{-1} X\right)^{-1} X' (\Omega^*)^{-1} y \tag{6.7}$$

which can also be written as:

$$\hat{\beta}_{\text{GLS}} = \left(X' \Omega^{-1} X\right)^{-1} X' \Omega^{-1} y \tag{6.8}$$

since $\Omega = \sigma^2 \Omega^*$. This model and estimation method are "generalized" because other models can be obtained as special cases. The ordinary least squares estimator is one such special case in which $\Omega = \sigma^2 I$. We discuss other special cases below. The variance-covariance matrix of the GLS-estimator $\hat{\beta}_{\text{GLS}}$ is:

$$\text{Var}\left(\hat{\beta}_{\text{GLS}}\right) = \left(X' \Omega^{-1} X\right)^{-1} = \sigma^2 \left(X' (\Omega^*)^{-1} X\right)^{-1}. \tag{6.9}$$

If Ω is unknown, as it is in empirical work, we replace Ω by $\hat{\Omega}$ and use an Estimated Generalized Least Squares (EGLS) estimator (also called Feasible Generalized Least Squares (FGLS) estimator). This estimator is usually a two-stage estimator. In the first stage, the OLS-estimates are used to define residuals, and these residuals are used to estimate Ω. This estimate of Ω is used in the second stage to obtain the EGLS estimator denoted by $\hat{\hat{\beta}}_{\text{GLS}}$.

6.2.2 GLS and Heteroscedasticity

We now consider one special case in which the disturbances are *heteroscedastic*. Suppose that the first T^* disturbances have variance σ_1^2, and that disturbances $T^* + 1, \ldots, T$ have variance σ_2^2. Such a setting is often encountered when data from different cross-sections data are used. In such a case, we have:

[2]To see that the last equality holds, we define the matrix A as: $A = V(V'V)^{-1} V'$. If we multiply A by V, we have that $AV = V(V'V)^{-1} V'V = V$, which only holds if $A = I$.

[3]First derived by Aitken (1935), and for that reason also known as the Aitken estimator.

$$\mathrm{E}(uu') = \Omega = \begin{pmatrix} \sigma_1^2 & & & & \\ & \ddots & & 0 & \\ & & \sigma_1^2 & & \\ & & & \sigma_2^2 & \\ & 0 & & & \ddots \\ & & & & \sigma_2^2 \end{pmatrix} \tag{6.10}$$

which is a diagonal matrix with $\sigma_1^2,\ldots,\sigma_1^2,\sigma_2^2,\ldots,\sigma_2^2$ as diagonal elements. This special case of GLS is referred to as Weighted Least Squares (WLS), because GLS can be interpreted as OLS applied to a transformed model [compare (5.8) in which the transformation involves a predictor variable] with variables:

$$\tilde{y}_t = \frac{y_t}{\sigma_1} \text{ and } \tilde{x}_{kt} = \frac{x_{kt}}{\sigma_1} \quad (k = 1,\ldots,K) \qquad \text{for } t = 1,\ldots,T^*,$$

$$\tilde{y}_t = \frac{y_t}{\sigma_2} \text{ and } \tilde{x}_{kt} = \frac{x_{kt}}{\sigma_2} \quad (k = 1,\ldots,K) \qquad \text{for } t = T^*+1,\ldots,T.$$

Equation (5.8) suggests that the intercept of the model is also affected by this transformation. Hence, we need to apply the same transformation to a $(T \times 1)$-column of ones [see (4.14)] and add this to the set of explanatory variables. The model should then be estimated without an intercept.

Estimates for σ_1 and σ_2 can be obtained by taking the square root of the sample variance of the OLS-residuals in both groups[4]:

$$\hat{\sigma}_1 = \sqrt{\hat{\sigma}_1^2} = \sqrt{\frac{1}{T^*-1}\sum_{t=1}^{T^*}\hat{u}_t^2} \quad ,$$

$$\hat{\sigma}_2 = \sqrt{\hat{\sigma}_2^2} = \sqrt{\frac{1}{T-T^*-1}\sum_{t=T^*+1}^{T}\hat{u}_t^2} \quad .$$

In (5.7) the variance of the disturbance increases with the squared value of a predictor variable.[5] Prais and Houthakker (1955) suggested a variance proportional to the squared expected value of the criterion variable. Hence there are other opportunities to accommodate heteroscedasticity.

[4]Formally both estimates need to be pre-multiplied by $(T-1)/(T-K)$, but that does not affect the estimate for β.

[5]See Judge et al. (1985, pp. 439–441) for a more general expression.

6.2.3 GLS and Autocorrelation

A second special case of GLS is typical for time-series data. In this case, the covariances, $\text{Cov}(u_t, u_{t'})$, $t \neq t'$ differ from zero (but we assume that the disturbances are homoscedastic). We consider the case that the disturbances are generated by a first-order autoregressive scheme, also called a first-order stationary (Markov) scheme, as in (5.11):

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad t = 1, \ldots, T, \quad |\rho_j| < 1 \tag{6.11}$$

where the ε_t are independent normally distributed random variables with mean zero, and variance equal to σ_ε^2. We also assume ε_t to be independent of u_{t-1}. By successive substitution for u_{t-1}, u_{t-2} in (6.11) we obtain:

$$u_t = \rho^s u_{t-s} + \rho^{s-1}\varepsilon_{t-s+1} + \cdots + \rho^2\varepsilon_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t. \tag{6.12}$$

After multiplying both sides of (6.12) by u_{t-s} and taking expectations, we have:

$$\begin{aligned} \text{E}(u_t u_{t-s}) &= \rho^s \text{E}(u_{t-s}u_{t-s}) + \rho^{s-1}\text{E}(\varepsilon_{t-s+1}u_{t-s}) + \cdots \\ &\quad + \rho^2 \text{E}(\varepsilon_{t-2}u_{t-s}) + \rho\text{E}(\varepsilon_{t-1}u_{t-s}) + \text{E}(\varepsilon_t u_{t-s}) \\ &= \rho^s \sigma_u^2 \end{aligned} \tag{6.13}$$

since the ε_t are independent of u_{t-1} and u_t has variance σ_u^2.[6]

The variance-covariance matrix Ω now has the following form:

$$\text{E}(uu') = \Omega = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}. \tag{6.14}$$

An estimate for the autocorrelation coefficient ρ can be obtained as follows[7]:

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=2}^{T} \hat{u}_{t-1}^2} \tag{6.15}$$

[6] It can be shown that $\sigma_u^2 = \sigma_\varepsilon^2/(1-\rho^2)$.

[7] This is a least squares estimate of ρ. It differs slightly from the maximum likelihood estimator. See Greene (2012, pp. 966–967) for other estimators.

where the $\hat{u}_t$ are the OLS-residuals. Then by substituting $\hat{\rho}$ for ρ in (6.14), we obtain $\hat{\Omega}$, and we use the EGLS estimator.

To demonstrate that also in the presence of autocorrelation one can apply OLS to suitably transformed variables, we return to Eq. (4.13):

$$y_t = \alpha + \beta_1 x_{1t} + \cdots + \beta_K x_{Kt} + u_t, \quad t = 1, \ldots, T, \tag{6.16}$$

where u_t is defined as in (6.11). By lagging this expression one period, and by premultiplying with ρ, we obtain for $t = 2, \ldots, T$:

$$\rho y_{t-1} = \rho\alpha + \rho\beta_1 x_{1,t-1} + \cdots + \rho\beta_K x_{K,t-1} + \rho u_{t-1}. \tag{6.17}$$

Subtracting (6.17) from (6.16) and using (6.11) we get for $t = 2, \ldots, T$:

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \beta_1(x_{1t} - \rho x_{1,t-1}) + \cdots + \beta_K(x_{Kt} - \rho x_{K,t-1}) + \varepsilon_t. \tag{6.18}$$

If $\hat{\rho}$ is computed as in (6.15), where the residuals are obtained by applying OLS to (6.16), and $\hat{\rho}$ is substituted for ρ in (6.18), OLS can be applied to the transformed variables. Estimating the parameters in (6.18) is known as Cochrane–Orcutt estimation (Cochrane and Orcutt 1949). Note that this procedure does not have a data point for $t = 1$. Prais–Winsten estimation (Wooldridge 2012, p. 411) adds the following data point for $t = 1$ to Eq. (6.18):

$$\sqrt{1-\rho^2}y_1 = \sqrt{1-\rho^2}\alpha + \beta_1\sqrt{1-\rho^2}x_{1,1} + \cdots + \beta_K\sqrt{1-\rho^2}x_{K,1} + \sqrt{1-\rho^2}u_1.$$

Asymptotically, it makes no difference whether or not the first observation is used, particularly because many time series are substantially large. In practice, both the Cochrane–Orcutt and Prais–Winsten estimation methods are used in an iterative scheme. Once the EGLS-estimate of ρ is found using (6.15), $\hat{\rho}$ can be substituted in (6.18), we can find a new set of residuals, obtain a new estimate of ρ, transform the data using the new estimate of σ, and estimate (6.18) again, etc. The process can be repeated until the estimate of ρ differs very little from the previous estimate.

6.2.4 Using Generalized Least Squares with Panel Data

In Sects. 2.7 and 4.5 we discussed the situation where time series of multiple cross-sections were available. As we indicated in Sect. 6.2.2, heteroscedasticity is a concern in such situations. Consequently, pooling cross sections is likely to suffer from the problems indicated in the second row of Table 5.1. Given that there are data available over time for each of the cross sections, assumption 3 might also be violated simultaneously. An additional problem might arise if the autoregressive nature is different across the cross-sections (i.e. when each cross section has a

different value of ρ). In this subsection, we discuss how to apply GLS in this case. Our starting point is the K-variable model in Eq. (6.1) for each cross section, indexed by i:

$$y_i = X_i\beta_i + u_i. \tag{6.19}$$

We allow the disturbances in (6.19) to be simultaneously cross-sectionally heteroscedastic and time-wise autoregressive, but we assume that $\text{Cov}(u_{it}, u_{jt}) = 0$, for $i \neq j$ and for all t (but relax this assumption below). Suppose that we have N cross sections, so that we have N vectors of disturbances of size $T \times 1$: $u_1, u_2, \ldots, u_N$. Let us now stack these vectors, and let $\underline{u}$ denote the result:

$$\underline{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}. \tag{6.20}$$

The vector $\underline{u}$ has dimensions $NT \times 1$. Given the violations discussed above, the variance-covariance matrix of $\underline{u}$ has the following structure:

$$\text{E}(\underline{u}\,\underline{u}') = \underline{\Omega} = \begin{pmatrix} \sigma_1^2 P_1 & & & 0 \\ & \sigma_2^2 P_2 & & \\ & & \ddots & \\ 0 & & & \sigma_N^2 P_N \end{pmatrix}. \tag{6.21}$$

$\underline{\Omega}$ is a block-diagonal $NT \times NT$ matrix, where P_i, $i = 1, \ldots, N$ are $T \times T$ matrices that allow for a different autoregressive parameter ρ_i for each cross-section i. They are defined as

$$P_i = \begin{pmatrix} 1 & \rho_i & \rho_i^2 & \cdots & \rho_i^{T-1} \\ \rho_i & 1 & \rho_i & \cdots & \rho_i^{T-2} \\ \rho_i^2 & \rho_i & 1 & \cdots & \rho_i^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_i^{T-1} & \rho_i^{T-2} & \rho_i^{T-3} & \cdots & 1 \end{pmatrix}. \tag{6.22}$$

All other elements of $\underline{\Omega}$ in (6.21) are zero.

To obtain estimates of the parameters in (6.19) and (6.21), we proceed in several steps. First we stack the data and express the relations in (6.19) in a system of equations as follows[8]:

[8]Because the N sets of equations in (6.23) do not seem to be related, this structure is referred to as "Seemingly Unrelated Regressions (SUR)", see Zellner (1962).

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} X_1 & & 0 \\ & X_2 & \\ & \ddots & \\ 0 & & X_N \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \quad \text{or} \tag{6.23}$$

$$\underline{y} = \underline{Z}\underline{\beta} + \underline{u} \tag{6.24}$$

where

$$\begin{aligned} \underline{y} &= \text{ a } NT \times 1 \text{ vector,} \\ \underline{Z} &= \text{ a } NT \times KN \text{ matrix,} \\ \underline{\beta} &= \text{ a } KN \times 1 \text{ vector,} \\ \underline{u} &= \text{ a } NT \times 1 \text{ vector.} \end{aligned}$$

OLS is applied to all NT observations in (6.24) from which residuals $\hat{u}_{it}$ are obtained. We then estimate the autocorrelation parameters and incorporate these in (6.22) to obtain estimates for P_i ($i = 1, \ldots, N$). We also estimate the error variance for each cross section. In this case the relation between the estimated variance of the autocorrelated disturbance u_{it} and the variance of the error term ε_{it} [compare Eq. (6.11)] is:

$$\hat{\sigma}^2_{u_i} = \frac{\sigma^2_{\varepsilon_i}}{1 - \hat{\rho}_i^2}. \tag{6.25}$$

We use these estimates in (6.21), to obtain an estimate for $\underline{\Omega}$, and we apply GLS [Eq. (6.8)] to estimate $\underline{\beta}$, accounting for cross-section specific (first-order) autocorrelated disturbances and heteroscedasticity.

Next we relax the assumption that the disturbances are independent between the cross sections, i.e. the assumption that $\text{Cov}(u_{it}, u_{jt}) = 0$ for $i \neq j$.

We assume

$$\text{E}(u_{it}^2) = \sigma_i^2 \quad \text{for all } t, \tag{6.26}$$

$$\text{Cov}(u_{it}, u_{js}) = 0 \quad \text{for all } t \neq s \text{ and all } i \text{ and } j, \tag{6.27}$$

$$\text{Cov}(u_{it}, u_{jt}) = \sigma_{ij} \quad \text{for all } t \quad (= \sigma_i^2 \text{ if } i = j) \tag{6.28}$$

which together imply that there is *contemporaneous* correlation between the disturbances for different cross sections, that there is heteroscedasticity across the cross sections, but that there is no autocorrelation (this assumption is relaxed shortly). Using the assumptions in (6.26) through (6.28), we obtain:

$$\mathrm{E}(\underline{u}\,\underline{u}') = \underline{\Omega} = \begin{pmatrix} \sigma_1^2 I & \sigma_{12} I & \cdots & \sigma_{1N} I \\ \sigma_{21} I & \sigma_2^2 I & \cdots & \sigma_{2N} I \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} I & \sigma_{N2} I & \cdots & \sigma_N^2 I \end{pmatrix} \tag{6.29}$$

and

$$\hat{\underline{\beta}} = (\underline{Z}'\underline{\Omega}^{-1}\underline{Z})^{-1}\underline{Z}'\underline{\Omega}^{-1}\underline{y}. \tag{6.30}$$

Zellner (1962) proposed the following estimation procedure.[9] First estimate (6.23) by OLS. Then estimate the elements of $\underline{\Omega}$ from the OLS residuals:

$$\hat{\sigma}_{ij} = \frac{\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt}}{T-K}, \text{ for all } i \text{ and } j \tag{6.31}$$

where $\hat{\sigma}_{ii} = \hat{\sigma}_i^2$. Next, the EGLS estimator is created by substituting $\hat{\sigma}_{ij}$ for σ_{ij} in (6.29). By iterating this process it is possible to obtain better estimates of σ_{ij} (i.e. when the process converges).

We note that in one application, the $\hat{\sigma}_{ij}$ values were used to find explanations for the nature of competition between brands i and j.[10] This is conceptually akin to the idea that in empirical research the presence of autocorrelated (or contemporaneously correlated) residuals should often be interpreted as indicating that the model is misspecified. The use of Generalized Least Squares to accommodate such systematic patterns in the disturbances will not result in improved slope parameter estimates if the patterns are due to, for example, missing predictor variables or incorrect functional forms. We emphasize that model builders must be convinced that they have used the best possible model specification. And model builders should have a logical, substantively meaningful reason for the disturbances to be correlated when the model is otherwise assumed to be complete.

Finally, we show the structure of the variance-covariance matrix of the disturbances in case of contemporaneous correlation, heteroscedasticity and autocorrelation, assuming an autoregressive scheme as described in (6.11) but now with cross-sectional specific autocorrelation terms $\rho_i \neq \rho$ for $i = 1, \ldots, N$. We then have:

$$\mathrm{E}(\underline{u}\,\underline{u}') = \underline{\Omega} = \begin{pmatrix} \sigma_1^2 P_{11} & \sigma_{12} P_{12} & \cdots & \sigma_{1N} P_{1N} \\ \sigma_{21} P_{21} & \sigma_2^2 P_{22} & \cdots & \sigma_{2N} P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} P_{N1} & \sigma_{N2} P_{N2} & \cdots & \sigma_N^2 P_{NN} \end{pmatrix} \tag{6.32}$$

[9] See Zellner (1962); Kmenta (1971, pp. 517–519). See also Leeflang (1974, pp. 124–127).

[10] See Clarke (1973). For an asymmetric, non-hierarchical market share model, Carpenter et al. (1988) use the $\hat{\sigma}_{ij}$ values to identify potential cross-effects.

where

$$P_{ij} = \begin{pmatrix} 1 & \rho_j & \cdots & \rho_j^{T-1} \\ \rho_i & 1 & \cdots & \rho_j^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_i^{T-1} & \rho_i^{T-2} & \cdots & 1 \end{pmatrix} \quad \text{for } i, j = 1, \ldots, N.$$

In sum-constrained models we have that

$$\sum_{i=1}^{N} \hat{u}_{it} = 0$$

for each $t = 1, 2, \ldots, T$. By definition there is contemporaneous correlation of the disturbances in these models. As a consequence, the contemporaneous variance-covariance matrix (6.29) is singular, because the elements of each row total to zero:

$$\sum_{j=1}^{N} \hat{\sigma}_{ij} = \frac{\sum_{j=1}^{N} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt}}{T - K} = \frac{\sum_{t=1}^{T} \hat{u}_{it} \sum_{j=1}^{N} \hat{u}_{jt}}{T - K} = 0. \tag{6.33}$$

To avoid singularity, one equation is deleted. If the matrix $\underline{\Omega}$ is *known*, the resulting parameter estimates are invariant to which equation (which cross-sectional unit) is deleted.[11] When the variance-covariance matrix is *unknown*, as usually is the case, the parameter estimates depend on which equation is deleted.[12]

6.3 The Verhouten Case Revisited

In Sect. 5.8 we discussed validation of the Verhouten model. One of the most revealing problems is the high degree of multicollinearity of the independent variables. The following variables are highly correlated:

- (natural logarithm of) price of Verhouten and feature only of Verhouten;
- display-only of Verhouten and feature and display of Verhouten.

In this section we demonstrate that *reformulation* of the model through recoding variables offers opportunities to reduce multicollinearity. We then treat a second issue that we encountered in Sect. 5.8: autocorrelation of the disturbance term. Finally, we illustrate how to resolve heteroscedasticity of the disturbances.

[11] See McGuire et al. (1968). See also Hanssens et al. (2001).

[12] See Gaver et al. (1988).

6.3.1 Multicollinearity

In the Verhouten case we reformulate the variables and ultimately the model specification to reduce multicollinearity. This is done in a number of steps:

1. Replace the price of Verhouten by a price index:

$$\text{price index in } t = \frac{\text{price in week } t}{\text{regular price}} = \frac{P_t}{RP} = PI_t \tag{6.34}$$

   We take the median of all prices from week 1 to 40 as the regular price: $RP = €\,1.51$.
2. We define four price variables:

   - PF_t = price index if there is feature only support;
   - PD_t = price index if there is display only support;
   - PFD_t = price index if there feature & display support;
   - PWO_t = price index if there is no (feature and/or display) support;

   This definition is based on the outcomes of empirical studies such as Van Heerde et al. (2002, 2004) who found different price effects depending on the feature/display support condition. We also define new feature and display variables:

   - FWO_t = feature only, but without (wo) a price cut;
   - DWO_t = display only, without a price cut;
   - $FDWO_t$ = feature and display support, but no price cut.

   Given that the F_t, D_t and FD_t variables express the support in relative numbers, we need to define a threshold that distinguishes between occasions when price discounts are sufficiently supported by featuring and/or displays and when we consider them to be not supported. We subjectively chose to set the threshold at 40 %, which means that we consider prices to be supported if at least 40 % or more stores use some form of non-price promotions. Specifically:

   - $PF_t = PI_t$ if $F_t \geq 0.4$, and 1 otherwise[13];
   - $PD_t = PI_t$ if $D_t \geq 0.4$, and 1 otherwise;
   - $PFD_t = PI_t$ if $FD_t \geq 0.4$, and 1 otherwise;
   - $PWO_t = PI_t$ if $D_t < 0.4$ and $F_t < 0.4$ and $FD_t < 0.4$, and 1 otherwise.

   The non-price promotion variables *FWO*, *DWO* and *FDWO* are defined in a similar way. Here we assume that if the price index is larger than 0.8 we deal with *FWO*, *DWO* and *FDWO*. Hence,

   - $FWO_t = F_t$ if $PI_t > 0.8$ and 0 otherwise;
   - $DWO_t = D_t$ if $PI_t > 0.8$ and 0 otherwise;
   - $FDWO_t = FD_t$ if $PI_t > 0.8$ and 0 otherwise.

[13] Which means that PF_t does not pick up any price discounts in these weeks.

3. We specify and calibrate (estimate and validate) a new model, based on (5.64), where we use the new variables:

$$S_{1t} = \alpha\left(PWO_t^{\beta_1} PF_t^{\beta_2} PD_t^{\beta_3} PFD_t^{\beta_4} \beta_5^{FWO_t} \beta_6^{DWO_t} \beta_7^{FDWO_t}\right)\varepsilon_{1t} \tag{6.35}$$

where

S_{1t} = sales of Verhouten (brand 1) at Albert Heijn in week t,

ε_{1t} = a disturbance term, and

all other variables have been defined above. Remember that all variables refer to brand 1 (Verhouten).

4. We check whether the (new) variables fluctuate sufficiently. We find that PD_t, PFD_t, DWO_t and $FDWO_t$ do not vary sufficiently over time. We delete these variables from the model and obtain:

$$S_{1t} = \alpha\left(PWO_t^{\beta_1} PF_t^{\beta_2} \beta_3^{FWO_t}\right)\varepsilon_{1t}. \tag{6.36}$$

5. We estimate and validate (6.36) using the first 60 observations in the data set, and obtain the results which are shown in Table 6.1. From Table 6.1 we deduce that

Table 6.1 Parameter estimates and statistics of (6.36) (OLS)

Predictor variables	Regression coefficient	Estimated standard error	t-value	p-value	VIF-value
Intercept ($\ln(\alpha)$)	4.54	0.03	148.35	0.00	–
PWO (β_1)	−3.33	0.48	−6.99	0.00	1.01
PF (β_2)	−4.00	0.27	−15.21	0.00	1.01
FWO ($\ln(\beta_3)$)	0.38	0.20	1.93	0.06	1.01
Number of observations = 60,		$R^2 = 0.83$,	$R_a^2 = 0.82$,		$\hat{\sigma} = 0.19$
RSS = 2.09,	ESS = 10.34,	F-value = 92.49,		DW-value = 1.22	

there is hardly any collinearity between the independent variables which is due to the redefinition of the variables which are by definition (almost) independent. Comparing $\hat{\beta}_1$ with $\hat{\beta}_2$ confirms the findings in other studies (Van Heerde et al. 2002, 2004) that the effect of supported price is larger in absolute value than the effect of unsupported price.

In Sect. 5.8, we also encountered the problem that the residuals are positively autocorrelated. Let us now focus on this issue. Just as in Sect. 5.8, the value of the Durbin–Watson statistic in Table 6.1 indicates that there is (positive) autocorrelation, and the autocorrelation coefficient is estimated using (6.15) as 0.368. We accommodate autocorrelation in the next step.

6.3.2 Autocorrelation

We remove autocorrelation following the procedure outlined in Sect. 6.2.3 by transforming the variables in Eq. (6.36) as follows:

$$\ln S_{1t} - 0.368 \ln S_{1,t-1} ,$$
$$\ln PWO_t - 0.368 \ln PWO_{t-1} ,$$
$$\ln PF_t - 0.368 \ln PF_{t-1} ,$$
$$FWO_t - 0.368 FWO_{t-1} .$$

Applying OLS to these variables leads to parameter estimates and statistics that are shown in Table 6.2. We may conclude from Table 6.2 that the Durbin–Watson statistic indicates that there is no autocorrelation.

Table 6.2 Parameter estimates and statistics of (6.36) (accounting for autocorrelation)

Predictor variables	Regression coefficient	Estimated standard error	t-value	p-value	VIF-value
Intercept ($\ln(\alpha)/(1-\rho)$)	2.88	0.02	115.17	0.00	–
PWO (β_1)	−3.66	0.42	−8.73	0.00	1.07
PF (β_2)	−3.98	0.23	−17.67	0.00	1.08
FWO ($\ln(\beta_3)$)	0.29	0.18	1.60	0.12	1.03
Number of observations = 59, $R^2 = 0.86$, $R_a^2 = 0.85$, $\hat{\sigma} = 0.17$					
RSS = 1.67, ESS = 10.17, F-value = 111.7, DW-value = 2.01					

As pointed out in Sect. 5.2.3, the parameter estimates are unbiased if there is autocorrelation. From a comparison of the corresponding parameter estimates in Tables 6.1 and 6.2, it appears, however, that accounting for autocorrelation also leads to somewhat different parameter estimates, but these differences are not significant. The fact that the t-values are higher in Table 6.2 than in Table 6.1 indicates that the GLS correction leads to more efficient estimation.

Finally, we observe that the R^2 in Table 6.2 differs substantially from the R^2 based on OLS (Table 6.1). However, these R^2's should not be compared, because the R^2 in Table 6.2 is based on the relation between transformed dependent variables on transformed independent variables (Wooldridge 2012, p. 412).

6.3.3 Heteroscedasticity

Heteroscedasticity also reduces the efficiency of the (ordinary least squares) parameter estimates. Although heteroscedasticity occurs especially if cross-sectional data are used for estimation (see Sect. 5.2.2), we account for heteroscedasticity in our

example which is based on time-series data. To this end we make a distinction between observations which correspond with weeks with substantial promotions ($P_{1t} < €1.45$) and "non-promotional" weekly observations. Hence, we divide our sample of 59 observations that remain after the transformation to account for autocorrelation into two subsamples: one with 9 observations and one with 50 observations respectively. We apply the Goldfeld–Quandt test which is based on the residual sum of squares of the two subsamples:

$$F = \frac{\text{RSS}_1/(9-3)}{\text{RSS}_2/(50-3)} = 1.972 \tag{6.37}$$

The critical F-value equals $F^6_{47}(0.95) = 2.299$, which is larger than the value of the F-statistic so that we conclude that we find *no* evidence of significant heteroscedasticity.[14]

Although it is not necessary we may apply the transformation of Sect. 6.2.2 to the intercept and to all the variables of the model. The standard deviation of the 9 promotional residuals is estimated as 0.228, and the standard deviation of the 50 residuals that correspond to the non-promotional weeks equals 0.167. The resulting parameter estimates deviate very little from the estimates in Table 6.2, which is expected because a correction for heteroscedasticity was not needed.

6.4 Maximum Likelihood Estimation

There are different ways to obtain estimators. In the preceding sections we concentrated on the least squares method. In this section we discuss other methods to obtain estimators, viz. Maximum Likelihood (ML) methods. In Sect. 6.4.1 we provide a brief review of the ML estimation method for the simplest case of estimating the parameters of a distribution, and present a small synthetic example. We discuss the large sample properties of ML estimators in Sect. 6.4.2. In Sect. 6.4.3 we extend simple ML estimation to cases where we have explanatory variables. We summarize some well-known statistical tests based on the likelihood in Sect. 6.4.4 and discuss an empirical example in Sect. 6.4.5.

6.4.1 Maximizing the Likelihood

The principle of Maximum Likelihood, due to Fischer (1922), provides a statistical framework for assessing the information available in the data. The principle of maximum likelihood is based on distributional assumptions about the data.

[14] We note that this might be due to the small number of observations in the promotional group.

The likelihood principle is to choose as estimate of the parameter (parameter vector) θ_0 that value of θ that maximizes the likelihood of observing the actual sample. In the discrete case this likelihood is the *probability* obtained from the probability mass function; in the continuous case this is the density. If one value of θ implies that the probability of the observed data occurring is 0.0012, whereas a second value of θ gives a higher probability of 0.0014, then the second value of θ is a better estimate (Cameron and Trivedi 2009, p. 139).

Suppose that we have N random variables $\{Y_1, \ldots, Y_N\}$ with observations that are denoted as $\{y_1, \ldots, y_N\}$, such as purchase frequencies for a sample of N subjects. Let $f(y_i \mid \theta)$ denote the probability density function for Y_i, where θ is a parameter characterizing the distribution (we assume θ to be a scalar for convenience).

The Maximum Likelihood principle is an estimation principle that finds an estimate for one or more unknown parameters (say θ) such that it maximizes the likelihood of observing the data $y = \{y_1, \cdots, y_N\}$. The Likelihood of a model (L) can be interpreted as the probability of the observed data y, given that model. A certain parameter value θ_1 is more likely than another, θ_2, in light of the observed data, if it makes observing those data more probable. In that case, the θ_1 will result in a larger value of the likelihood than θ_2: so that $L(\theta_1) > L(\theta_2)$.

The probability of observing y_i is provided by the *pf* or *pdf*: $f(y_i \mid \theta)$. When the variables for the N subjects are assumed independent, the joint density function of all observations is the product of the densities over i. This gives the following expression for the likelihood:

$$L(\theta) = \prod_{i=1}^{N} f(y_i \mid \theta). \tag{6.38}$$

Note that we present the likelihood as a function of the unknown parameter θ; the data is considered as given.

This formulation holds for both discrete and continuous random variables. Discrete random variables are for example 0/1 choices, or purchase frequencies, while market shares and some ratings on scales can be considered as continuous random variables. Important characteristics of these random variables are their expectations and variances. In the purchase frequency example usually a (*discrete*) Poisson distribution is assumed (see Chap. 8):

$$Y_i \sim f(y_i \mid \lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}. \tag{6.39}$$

The expectation of the Poisson variable in (6.39) can be shown to be $\mathrm{E}(Y_i) = \lambda$, and its variance is $\mathrm{Var}(Y_i) = \lambda$.

One of the well-known *continuous* distributions is the exponential distribution:

$$Y_i \sim f(y_i \mid \mu) = \mu e^{-\mu y_i}. \tag{6.40}$$

The mean and variance of the exponential random variable in (6.40) are $\mathrm{E}(Y_i) = 1/\mu$ and $\mathrm{Var}(Y_i) = 1/\mu^2$. This distribution is frequently used for interpurchase times.[15] The functions in (6.39) and (6.40) are known as the probability function (*pf*) and probability density function (*pdf*), respectively. Both the exponential and the Poisson distributions belong to the Exponential Family, which is a general family of distributions that encompasses both discrete and continuous distributions.[16] Also the normal and the Bernoulli distributions belong to the exponential family. The distributions in this class share some common properties that facilitates studying these distributions simultaneously.

If f belongs to the exponential family, the expression for the likelihood in (6.38) simplifies considerably after taking the natural logarithm. The product in (6.38) is replaced by a sum:

$$l(\theta) = \sum_{i=1}^{N} \ln f(y_i \mid \theta). \tag{6.41}$$

Since the natural logarithm is a monotonic function, maximizing the log-likelihood $l(\theta)$ in Eq. (6.41) yields the same estimates as maximizing the likelihood $L(\theta)$ in Eq. (6.38).

In the Poisson example, the log-likelihood takes a simple form, as the Poisson distribution belongs to the exponential family:

$$l(\lambda) = \ln\left(\prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right) \tag{6.42}$$

$$= -N\lambda + \ln\lambda \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \ln(y_i!). \tag{6.43}$$

The ML estimator of λ is obtained by setting the derivatives of the log-likelihood equal to zero:

$$\frac{\partial l(\lambda)}{\partial \lambda} = -N + \frac{1}{\lambda}\sum_{i=1}^{N} y_i = 0. \tag{6.44}$$

Solving (6.44) provides the Maximum Likelihood Estimator (MLE) for λ:

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} y_i\,, \tag{6.45}$$

which is the sample mean.

[15] Gupta (1991). See also Sect. 8.4.

[16] Cameron and Trivedi (2009, pp. 147–149).

Similarly, in the example of the exponential distribution, the log-likelihood is:

$$l(\mu) = N \ln \mu - \mu \sum_{i=1}^{N} y_i. \tag{6.46}$$

Setting the derivative of (6.46) with respect to μ to zero, and solving for μ yields the estimator:

$$\hat{\mu} = \frac{N}{\sum_{i=1}^{N} y_i}, \tag{6.47}$$

which is the inverse of the sample mean.

A graph of the log-likelihood against the corresponding parameter yields a concave function for the Poisson and the exponential distribution, each with a single maximum, indicating that the MLE is unique in both cases. This property holds for all members of the exponential family.

To illustrate ML estimation, we use a small synthetic example. Assume we have data on yearly purchase frequencies for 10 subjects. The data, generated from a Poisson distribution with a mean of $\lambda = 25$ are shown in the second column of Table 6.3. Applying Eq. (6.45) we obtain a MLE of $\lambda = 23.4$, which is quite close to the true value of 25.

Table 6.3 Data for the illustration of ML estimation

Subject i	y_i
1	25
2	22
3	31
4	21
5	24
6	26
7	20
8	27
9	24
10	14

6.4.2 Large Sample Properties of the MLE

One of the main benefits of the maximum likelihood approach is that it has attractive large sample properties. Small sample properties of the MLE are usually not known, except for cases when the likelihood is based on the normal distribution. Asymptotic

properties of MLEs are obtained if the sample size tends to infinity: $N \to \infty$. Under fairly general conditions, MLEs:

1. are consistent;
2. have asymptotically minimum variance;
3. are asymptotically normal.

The first important property of the MLE is *consistency*. $\hat{\theta}$ is said to be consistent for θ if the probability that the estimate differs from the true value by less than any arbitrarily small number δ, approaches zero as $N \to \infty$:

$$\lim_{N\to\infty} P(|\,\hat{\theta} - \theta\,| > \delta) = 0. \tag{6.48}$$

This property implies that MLEs tend to their true value in probability for large samples.

Second, the ML approach yields asymptotically *efficient* estimators. An estimator is efficient if it has the lowest possible variance among all estimators in a particular class, and thus has the highest precision.

Under fairly general conditions, the Cramér-Rao theorem states that[17] the variance of an unbiased estimate for θ is at least as large as $(I(\theta))^{-1}$, where $I(\theta)$ is defined as:

$$I(\theta) = \mathrm{E}\left(\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2\right). \tag{6.49}$$

If θ is a single parameter, $I(\theta)$ is called the *Information number*. When θ is a vector of parameters, $I(\theta)$ is known as the *Information matrix*.[18] For an intuitive interpretation of $I(\theta)$, one can view $I(\theta)$ as the variance of the first derivative of the log-likelihood in θ. If this variance is relatively large, the data result in relatively large changes in the slope of the log-likelihood around θ, which means that the log-likelihood is relatively peaked. Consequently, compared to situations where $I(\theta)$ is smaller and the log-likelihood is thus flatter, the data are more *informative* for finding a good estimate for θ. It can be shown that the asymptotic variance of the MLE equals $(I(\theta))^{-1}$, which makes the MLE asymptotically an efficient estimate.

Another useful expression for $I(\theta)$ relates $I(\theta)$ to H, the second derivative of $l(\theta)$:

$$I(\theta) = -\mathrm{E}(H)\,, \tag{6.50}$$

where H is defined as (using the notation for the case where θ is a vector of parameters):

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}. \tag{6.51}$$

[17] See e.g. Greene (2012, p. 476).

[18] When θ is a vector of parameters, the difference between the covariance matrix and $(I(\theta))^{-1}$ is a nonnegative matrix.

H is known as the *Hessian* of the log-likelihood.

In the Poisson example above, the asymptotic variance (AVar) of the MLE for λ can be computed by differentiating Eq. (6.44) once more with respect to λ. Using the result in (6.50) gives:

$$\begin{aligned} \text{AVar}(\hat{\lambda}) &= \left(-\text{E}\left(-\frac{1}{\lambda^2} \sum_{i=1}^{N} y_i \right) \right)^{-1} = \left(\frac{1}{\lambda^2} \text{E}\left(\sum_{i=1}^{N} y_i \right) \right)^{-1} . \\ &= \lambda / N \end{aligned} \tag{6.52}$$

A final important property of the ML estimator is that it is *asymptotically normal*.[19] This property of the ML estimator allows for statistical inference for the parameters in θ, based on the normal distribution:

$$\hat{\theta} \sim N(\theta, \text{AVar}(\hat{\theta})) , \tag{6.53}$$

where $\text{AVar}(\hat{\theta}) = (I(\theta))^{-1}$.

The above asymptotic properties of the likelihood hold under certain regularity conditions (Lindsey 1996, p. 187). Although a full discussion of these regularity conditions is beyond the scope of this monograph, the following aspects may be useful in practice. The log-likelihood function is said to be regular if in an open neighborhood of the true parameter value, it can be approximated by a quadratic function. Such an approximation breaks down in situations where the true value lies on the boundary of the parameter space so that the quadratic approximation is inappropriate, or when the number of parameters to be estimated increases with the number of observations. The latter situation occurs if in our Poisson example a parameter exists for each individual i, i.e. λ_i for $i = 1, \ldots, N$.

6.4.3 MLE with Explanatory Variables

In the preceding subsections we discussed MLE in the simplest case of estimating the parameters in an distribution. Marketing models, however, usually require the estimation of the effects of explanatory variables, whether we estimate these models by OLS, GLS or MLE. Fortunately, ML estimation is quite straightforwardly extended to cases where we use explanatory variables. In such cases, the log-likelihood has the following form:

$$l(\theta) = \sum_{i=1}^{N} \ln f(y_i | x_i, \theta) \tag{6.54}$$

[19] Compare Lindsey (1996, p. 199).

where x_i is an explanatory variable. Expression (6.54) is called a *conditional likelihood function* since it reflects the behavior of the y_i given/conditional on x_i. As these models become more complex, the likelihood equations do not yield closed-form expressions in all cases, and MLEs are determined using methods that maximize the likelihood numerically. These results extend to multivariate data, systems of equations, and panel data by replacing the scalars y_i and x_i by vectors.

To illustrate, let us continue our Poisson example that we started in Sect. 6.4.1. Let us consider again yearly purchase frequencies, but now assume that these are driven by household size. Values for y_i, the purchase frequency of household i, are generated by random draws from a Poisson distribution, where the mean (λ_i) depends on x_i, household size of household i as follows: $\lambda_i = \exp(1.5 + 0.5x_i)$. This is a two-parameter Poisson regression model with $\mu = 1.5$ and $\beta = 0.5$. Consequently, θ is now a (2×1) vector:

$$\theta = \begin{pmatrix} \mu \\ \beta \end{pmatrix}.$$

The simulated values for purchase frequency and household size for ten households are shown in Table 6.4. Because there are no closed-form solutions for the two parameters, we apply Newton's algorithm, to maximize the log-likelihood numerically.[20] Given a set of starting values, $\hat{\theta}_r$, the estimate for θ in iteration r, is determined by:

$$\hat{\theta}_r = \hat{\theta}_{r-1} - gH(\hat{\theta}_{r-1})^{-1} S(\hat{\theta}_{r-1}) \tag{6.55}$$

Table 6.4 Data for Poisson example with an explanatory variable

Subject i	y_i	x_i
1	7	1
2	7	1
3	18	3
4	12	2
5	6	1
6	1	1
7	31	3
8	15	2
9	57	5
10	29	4

[20] See Scales (1985); Eliason (1993).

where

$$
\begin{aligned}
g &= \text{the step length,} \\
H(\hat{\theta}_{r-1}) &= \text{the Hessian (the matrix of second-order derivatives of the} \\
&\quad \text{log-likelihood) evaluated at } \hat{\theta}_{r-1}, \text{ and} \\
S(\hat{\theta}_{r-1}) &= \text{the vector of first-order derivatives of the log-likelihood} \\
&\quad \text{(as in (6.44)) evaluated at } \hat{\theta}_{r-1}.
\end{aligned}
$$

At iteration r, both $H(\hat{\theta}_{r-1})$ and $S(\hat{\theta}_{r-1})$ are evaluated at the previous estimate $\hat{\theta}_{r-1}$. Table 6.5 shows the iteration process. As starting values for both parameters we take zero. The algorithm is said to converge if the first derivative of the log-likelihood (6.44) changes less than 10^{-5}. In this case this took six iterations. Table 6.5 shows that the ML estimates are close to the true parameter values.

Table 6.5 The iteration process from Newton's algorithm

r	μ	β	g	$l(\theta)$
1	0.00	0.00	0.00	−392.06
2	−0.81	0.88	0.10	−152.71
3	0.01	0.70	0.10	−108.10
4	1.35	0.53	0.45	−32.63
5	1.57	0.54	1.00	−29.12
6	1.55	0.54	1.00	−29.07

For specifying models for a wide range of data types the following method can be used.[21] First choose a distribution that is appropriate for the dependent variable y. Compare in this respect Sect. 3.4 and Table 3.1. Then parametrize the parameter(s) of that distribution in terms of the regressors and their parameters. Some commonly used distributions and parameterizations are given in Table 6.6.

Table 6.6 Maximum Likelihood: commonly used densities

Density	Range of y	Density $f(y)$	Common parametrization
Normal	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu = X\beta$ $\sigma^2 = \sigma^2$
Bernoulli	0 or 1	$p^y(1-p)^{1-y}$	$p = \frac{\exp(X\beta)}{1+\exp(X\beta)}$
Exponential	$(0, \infty)$	$\lambda \exp(-\lambda y)$	$\lambda = \exp X\beta$, or $\frac{1}{\lambda} = \exp X\beta$
Poisson	$0, 1, 2, \ldots$	$\frac{\exp(-\lambda)\lambda^y}{y!}$	$\lambda = \exp(X\beta)$

Source: Cameron and Trivedi (2009, p. 140)

[21] Based on Cameron and Trivedi (2009, p. 140).

Table 6.6 is an extension of Table 3.1. We briefly indicate which distributions are used for which kind of dependent variable[22]:

- For *continuous data* on $(-\infty, \infty)$, the normal distribution is the standard distribution. The classical linear regression model parameterizes $\mu = X\beta$ and assumes σ^2 is constant;
- For *discrete binary data* taking values 0 or 1, the density is always the Bernoulli, a special case of the binomial with one trial. The usual parameterizations for the Bernoulli density leads to the logit model, or the probit model. These models are discussed in Chap. 8;
- For *positive continuous data* on $(0, \infty)$, the default distribution is the exponential, as indicated by Table 6.6. For specific cases, notably for duration data considered in Chap. 8, the richer Weibull, gamma and log-normal distributions are often used;
- For *integer-valued count data* taking values $0, 1, 2, \cdots$ the Poisson distribution that was introduced in Sect. 6.4.1 is the default. Another possibility is to use the richer negative binomial distribution.

6.4.4 Statistical Tests

In this subsection we introduce some useful inferential tools for ML estimation. First let us assume that θ consists of a single parameter. With the asymptotic distribution of $\hat{\theta}$ in Eq. (6.53), we have asymptotically:

$$z = \frac{\hat{\theta} - \theta}{\sqrt{\text{AVar}(\hat{\theta})}} \sim N(0, 1) \tag{6.56}$$

which converges in distribution to a *standard* normal distribution. This is a very useful result that allows statistical tests for hypotheses formulated for θ, for example $H_0 : \theta = 0$, to be conducted using the z-test.

As an illustration, we obtained the following Hessian for the Poisson regression model in the previous subsection:

$$H = \begin{bmatrix} -17.90 & -47.50 \\ -47.50 & -134.93 \end{bmatrix}. \tag{6.57}$$

Inverting the Hessian after multiplying all values by -1, and taking the square root of the diagonal elements of the resulting matrix gives us the asymptotic standard errors (ASE) of the estimates. These are shown in Table 6.7. The t-values in Table 6.7 show that the null hypotheses that the mean and regression parameters

[22] We closely follow Cameron and Trivedi (2009, p. 140).

are zero is strongly rejected. Since the data were actually generated with non-zero parameter values, the results of the t-tests are consistent with our knowledge of the true value of the parameters. However, the conditions for the asymptotic approximations are unlikely to be valid in this example, as we have only 10 observations. These calculation only serve as an illustration.

Table 6.7 Estimation results of the synthetic data example

Parameter	Estimate	ASE	t-value
μ	1.55	0.29	5.31
β	0.54	0.11	5.11

Equation (6.56) assumes a scalar parameter θ. If, however, θ is a $(K \times 1)$ vector of parameters, then tests on the entire parameter vector (or sub vectors) can be conducted using the Wald-test, say for $H_0 : \theta = \theta_0$:

$$W = (\hat{\theta} - \theta_0)' \left(\text{AVar}(\hat{\theta})\right)^{-1} (\hat{\theta} - \theta_0) \tag{6.58}$$

which converges in distribution to a χ^2 variable with K degrees of freedom under the null hypothesis. The advantage of the Wald-test is that it allows for tests on several parameters of the model, without the need to re-estimate (a restricted version of) the model. In the Poisson regression example, the Wald test for jointly testing $\mu = 0$ and $\beta = 0$ yields a value of 161.87, with 2 degrees of freedom, which is highly significant.

Another frequently used test is the Likelihood Ratio (LR) test. The LR test is used to investigate two models that are nested and chooses that model that has the highest likelihood, given the observed data. A more detailed discussion of this test is given in Sect. 5.6.2. Two models are estimated, yielding log-likelihood values $l_1(\hat{\theta})$ and $l_2(\tilde{\theta})$, respectively, where we assume the latter model to be more restricted, for example because one or more parameters are set to zero. Due to the fact that the two models are nested, minus twice the difference in their log-likelihood values is asymptotically distributed as χ^2 under the null hypothesis that the restrictions are valid:

$$\text{LR} = -2(l_1(\hat{\theta}) - l_2(\tilde{\theta})) \sim \chi^2_{\text{df}} \tag{6.59}$$

where df is the difference in the number of parameters in $\hat{\theta}$ and $\tilde{\theta}$. The LR test requires estimation of two models, and is thus computationally more intensive than the Wald test.

If we re-estimate the model for the two-parameter Poisson regression synthetic data in Table 6.4 with the restriction $\beta = 0$, we obtain a log-likelihood of $l(\theta) = -44.68$ (Newton's algorithm converged in 6 iterations to a parameter value of $\hat{\mu} = 2.89$). Thus in this case the LR statistic for testing the models with and without (see Table 6.7) the restriction equals $\text{LR} = -2(-44.68 + 29.07) = 31.22$, which is highly significant at one degree of freedom (the difference in the number

of parameter values for the two models). This is of course expected since these synthetic data were generated with a nonzero value of β. As a cautionary note we mention that the asymptotic χ^2-distribution for the Wald- and LR tests are unlikely to hold given the small sample size.

If the models to be compared are not nested, the LR test does not apply. Information criteria are then commonly used to identify the most appropriate model, see Sect. 5.6.3.

6.4.5 *MLE with Explanatory Variables: An Example*

In this section we illustrate the use of MLE with explanatory variables. Van Nierop et al. (2011) study the effect of the introduction of an informational website by a large retailer on off-line customer buying behavior. Figure 6.1 depicts the average monthly number of store visits of registered website users and non-users over time. The website was introduced during period 15. The figure indicates that the average registered user of the site visits the store less than the average non-user.

Van Nierop et al. (2011) model the effect of the introduction of the website on V_{it}, the number of store visits by customer i in month t. As suggested by Table 6.6,

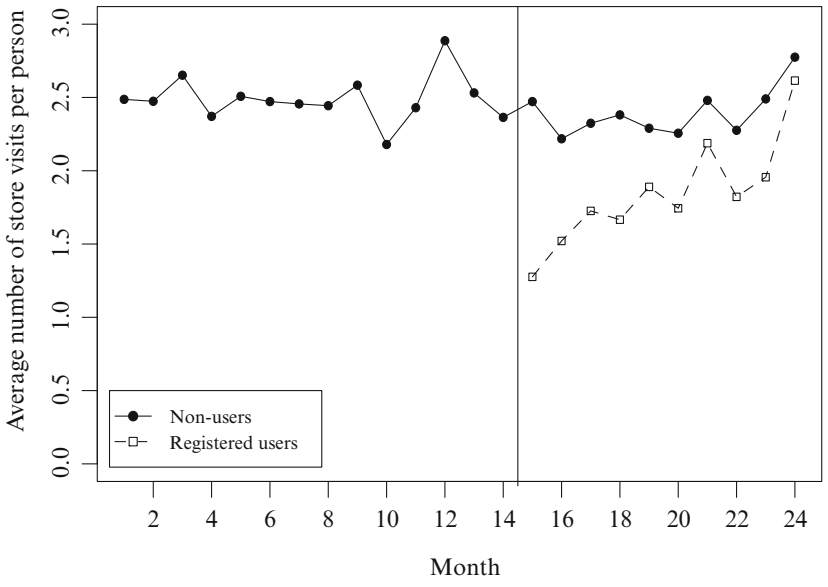

Fig. 6.1 Average number of store visits per person. *Source*: Van Nierop et al. (2011)

the Poisson regression model is used for this dependent variable. The probability that individual i visits the store v_{it} times in month t is written as[23]:

$$P(V_{it} = v_{it}) = \frac{e^{-\lambda_{it}} \lambda_{it}^{v_{it}}}{v_{it}!}. \tag{6.60}$$

The Poisson parameter λ_{it} reflects the expected number of store visits for individual i in month t, and is explained by K regressors $z_{i1t}, z_{i2t}, \ldots, z_{iKt}$ for individual i at time t:

$$\ln \lambda_{it} = \sum_{k=1}^{K} \theta_k z_{ikt} \tag{6.61}$$

The regressors consist of an intercept, a variable that capture the online behavior of that customer number, a dummy that indicates when the web site was introduced, dummies for own and competitive promotions, a trend variable, and individual-specific customer characteristics. The parameters $\theta_1, \ldots, \theta_K$ describes the effects of these explanatory variables.

Van Nierop et al. (2011) calibrated Eqs. (6.60) and (6.61) using data of a sample of 436 customers, consisting of both registered users (209) and non-users (227). All customers in the sample visited the store or the website at least once during the two periods before and after the introduction of the website. In total, the 436 customers generated 4,572 store visits.

Table 6.8 contains the estimation results for the parameter $\theta_1, \ldots, \theta_K$ of the Poisson model. The number of website visits affects the number of store visits negatively ($\theta_2 = -0.302, p < 0.001$). This means that for an individual customer, an informational website is expected to cause a reduction in off-line store visits. Combined with the after-introduction dummy for non-users, this result is striking. The dummy indicates that after correcting for possible other variables, non-users of the website visit the store more after its introduction than do registered users ($\theta_3 = 0.178$, $p < 0.001$). This effect is also reflected in Fig. 6.1.

Table 6.8 also shows that the number of store visits is positively influenced by the holiday shopping season ($\hat{\theta}_4 = 0.164$, $p < 0.002$) and the general promotion ($\hat{\theta}_5 = 0.106$, $p < 0.035$). The fashion promotion does not have a significant effect on the number of store visits ($p = 0.180$). The effects of the first two promotional activities by competitor 1 are negative ($\hat{\theta}_7 = -0.177$, $p = 0.015$ and $\hat{\theta}_8 = -0.241$, $p = 0.006$). The negative coefficient ($\hat{\theta}_{11} = -0.137$, $p = 0.003$) for the log of the time since introduction illustrates a trend effect that can be explained partly by a macro-economic decline. The distance to the closest store has a negative influence on the number store visits ($\hat{\theta}_{12} = -0.018$, $p < 0.001$), indicating that the farther away customers live, the fewer store visits they make. Older people visit the store more frequently than do younger people, though this effect is barely significant

[23] We closely follow the study by Van Nierop et al. (2011).

Table 6.8 Poisson parameter estimates for the number of store visits (N = 4,572)

Variable	Coefficient	p-value[a]
Intercept (θ_1)	1.155	<0.001
Number of website visits (θ_2)	−0.302	<0.001
Dummy user/non-user[b] (θ_3)	0.178	<0.001
Holiday shopping season dummy (θ_4)	0.164	0.002
General promotion dummy (θ_5)	0.106	0.035
Fashion promotion dummy (θ_6)	0.079	0.180
Competitor 1 starts webstore (θ_7)	−0.177	0.015
Competitor 1 major TV advertisement (θ_8)	−0.241	0.006
Competitor 1 advertisements in 2001 (θ_9)	0.002	0.971
Competitor 2 introduces magazine (θ_{10})	0.002	0.971
(Log) months since introduction (θ_{11})	−0.137	0.003
Distance to closest store in miles (θ_{12})	−0.018	<0.001
Age in years (θ_{13})	0.003	0.067
Gender (0 = male, 1 = female) (θ_{14})	−0.023	0.470
Higher education (0 = no, 1 = yes) (θ_{15})	0.057	0.047

[a]This columns shows p-values for two-sided t-tests
[b]This variable equals 0 before the introduction of the informational website; after the introduction, it equals 1 for non-users and 0 for registered users
Source: Van Nierop et al. (2011, p. 159).

($\hat{\theta}_{13} = 0.003$, $p = 0.067$). There is no difference in the number of store visits between men and women ($p = 0.470$). Finally, there is a positive relation between people with a least a college education and the number of store visits ($\hat{\theta}_{15} = 0.057$, $p = 0.047$).

6.5 Simultaneous Systems of Equations

Model building in marketing can involve:

- a single equation;
- multiple equations (such as in the case of multiple brands);
- simultaneous equations.

So far we only considered single- and multiple-equation models. Simultaneous equations represent a special case of multiple equations, and a system of simultaneous equations is used in cases where more than one equation is needed to properly specify the relations in a market. In such a system there are multiple *endogenous* variables which are the variables to be explained by the equations. The remaining variables are predetermined, consisting of *exogenous* and (potentially) lagged endogenous variables. Exogenous variables are considered to be determined outside the system of equations, and are therefore taken as given. They are similar to the predictor variables used in a single-equation model.

The concept of simultaneity refers to the idea that the endogenous variables are "explained" jointly and simultaneously by the predetermined variables and the disturbances. Importantly, an endogenous variable may be used both to explain other variables and to be explained, in different equations. As a result, such an endogenous variable cannot be stochastically independent of *all* disturbance terms. On the other hand, the predetermined variables may be assumed to be independent of the disturbances.

The need for special estimation methods for simultaneous equations derives especially from the violation of assumption 5 for the disturbances (Sect. 4.2.3) which states that the predictors (x_{kt}) and the disturbances (ε_t) are required to be independent. Thus, in essence the presence of an endogenous variable which has both an explanatory role and which is to be explained by the system is one of the reasons for the violation of assumption 5.

We first discuss in further detail the example that we briefly mentioned in Sect. 5.2.5 to illustrate the fundamental problem (a violation of assumption 5 with regard to the error term) caused by simultaneity in the relationships between the variables that we have in our data set.

Suppose a brand manager is interested in determining the effect of own advertising expenditures on unit sales. The manager believes that sales are mainly driven by own and competitive advertising expenditures, and assumes a linear demand function:

$$q_t = \beta_0 + \beta_1 a_t + \beta_2 a_t^c + \varepsilon_t \tag{6.62}$$

where

q_t = brand sales in 10,000 units,

a_t = brand advertising expenditures in thousands of dollars,

a_t^c = competitive advertising expenditures in thousands of dollars, and

ε_t = an error term.

In this example we consider competitive advertising as an exogenous variable.

Let us assume that the brand manager's advertising spending decisions depend on the brand's sales level, and also on advertising spending for other brands in the firm's portfolio, due to budget restrictions:

$$a_t = \alpha_0 + \alpha_1 q_t + \alpha_2 a_t^P + \nu_t \tag{6.63}$$

where

a_t^P = advertising expenditures for other brands in the firm's portfolio, and

ν_t = an error term.

We assume that a_t^P is an exogenous variable.

Equations (6.62) and (6.63) jointly constitute a system of two equations that contains two endogenous variables: q_t and a_t. One fundamental question is whether a_t is stochastically independent of ε_t in (6.62). By substituting (6.62) into (6.63) we obtain:

$$a_t = \alpha_0 + \alpha_1(\beta_0 + \beta_1 a_t + \beta_2 a_t^c + \varepsilon_t) + \alpha_2 a_t^P + \nu_t$$

so that

$$a_t = \frac{\alpha_0 + \alpha_1\beta_0 + \alpha_1\beta_2 a_t^c + \alpha_2 a_t^P + \nu_t}{1 - \alpha_1\beta_1} + \frac{\alpha_1}{1 - \alpha_1\beta_1}\varepsilon_t. \tag{6.64}$$

It is clear from (6.64) that a_t depends on ε_t. Thus, error-term assumption 5, required for single-equation estimation of (6.62) using OLS, is violated.

So far, in the advertising example, we were interested in estimating a single equation. However, this equation turned out to be part of a system of equations, which hindered application of OLS to the demand function only. One might also consider joint estimation of the complete system of equations. In the general we can specify a system of equations as follows:

$$Y\Gamma + ZB = E \tag{6.65}$$

where

Y = a $T \times m$ matrix containing the endogenous variables,

Γ = a $m \times m$ matrix of parameters (and constraints),

Z = a $T \times L$ matrix with exogenous variables and a column of ones,

B = a $L \times m$ matrix of parameters (and constraints),

E = a $T \times m$ matrix of disturbances.

In the example above, $m = 2$ and $L = 2$. This general description makes it possible to consider a variety of possible conditions. For example, if Γ is diagonal, then none of the endogenous variables depends on other endogenous variables. This reduces the problem to one for which it is possible to estimate each equation separately (or treat it as a multiple equation problem if the disturbances are correlated across equations).

If Γ is triangular (for example, any endogenous variable depends only on other endogenous variables that exist in equations specified earlier), we have essentially recursive equations. For example, we could imagine a brand manager with a demand equation such as (6.62) but an advertising budget that does not depend on current sales. The adequacy of single-equation estimation procedures is then determined by the lack of correlation between the disturbances across these two equations.

When Γ is neither diagonal nor triangular (as is the case in the example above), we need to use special estimation methods. The simplest and perhaps most popular

one is the two-stage least squares method 2SLS. This method is a special case of the method of *instrumental variables* (IV) briefly introduced in Sect. 5.2.5, and discussed further in the next section.

But first we let us discuss an issue that is related to structure of the system of equations. To this end, let us consider a simplified version of the system of equations defined by (6.62) and (6.63) where $\beta_2 = 0$ and $\alpha_2 = 0$:

$$q_t = \beta_0 + \beta_1 a_t + \varepsilon_t \tag{6.66}$$

$$a_t = \alpha_0 + \alpha_1 q_t + \nu_t. \tag{6.67}$$

We may rewrite (6.67) as

$$q_t = \alpha'_0 + \alpha'_1 a_t + \nu'_t$$

where $\alpha'_0 = -\alpha_0/\alpha_1$, and $\alpha'_1 = 1/\alpha_1$, and $\nu'_t = -\nu/\alpha_1$. Hence, an equivalent representation of the simplified system of equations is:

$$q_t = \beta_0 + \beta_1 a_t + \varepsilon_t \tag{6.68}$$

$$q_t = \alpha'_0 + \alpha'_1 a_t + \nu'_t. \tag{6.69}$$

The demand equation in (6.68) has precisely the same structure as the rewritten budget restriction in (6.69). Consequently, there is nothing in the structure of these two equations that helps to identify that one is representing the demand function, and the other the budget restriction function. Hence, we cannot hope to estimate the four parameters in these equations reliably. This example illustrates the general phenomenon that the structure of the system of equations can be such that the parameters cannot be uniquely determined. If that is the case, then the system of equations is said to be *unidentified*.

Let us investigate whether the exogenous variables that appear in Eqs. (6.62) and (6.63) are helpful in identifying the system of equations. The dashed lines in Fig. 6.2 indicate the relationship between advertising and sales that follow from the demand function in (6.62). Different levels of competitive advertising (a_t^c) result in vertical shifts in these relationships. The dash-dotted lines in Fig. 6.2 represent the relationships between q_t and a_t that follow from the decision function in Eq. (6.63) for different levels of advertising expenditures for other brands in the firm's portfolio (a_t^P), which shift these functions horizontally. Given these relationships, it is likely that advertising-sales values are observed in the vicinity of the intersections of these functions, indicated by dots in Fig. 6.2.

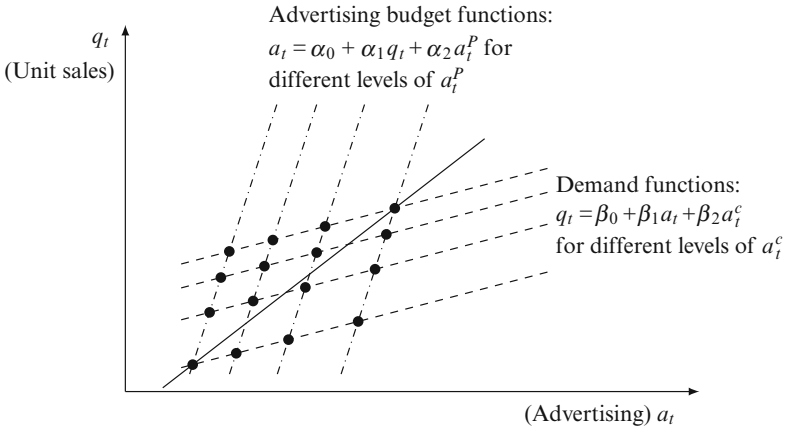

Fig. 6.2 Relations between advertising and unit sales for different values of a_t^c and a_t^P

If a relationship between advertising and sales is estimated with OLS using such data points, a regression line similar to the solid line in Fig. 6.2 will emerge, which does not resemble any of the functions depicted in Fig. 6.2. Specifically, the slope of the solid line differs from β_1, the slope of the demand functions, so that the effectiveness of advertising on sales is wrongly assessed.

We note that β_1 can be properly estimated if the regression equation controls for a_t^c (for example by fixing it to a certain level so that we are restricted to one of the dashed lines in Fig. 6.2). Subsequently, variation in a_t^P is then needed to determine the responsiveness of demand to advertising.

We conclude that adding a_t^c does not resolve the identification of the demand function. Surprisingly, it is the variation in a_t^P that allows us to trace the demand function; a_t^c merely serves as a control variable in the demand function. This is reversed when we consider the identification of the budget restriction function. The identification issue for this function is resolved by adding a_t^c to the system of equations in the demand function, and now a_t^P is only a control variable.

A full treatment of the identification of a system of equations is out of scope for this text. However, a simple condition might be helpful in making a first assessment of possible identification issues. The so-called *order condition* requires that for each equation in the system:

> "the number of exogenous variables that appear elsewhere in the equation system must be at least as large as the number of endogenous variables in this equation" (Greene 2012, p. 365).

The *order condition* is a *necessary* condition for identification. A *sufficient* condition is the *rank condition*. A detailed discussion of the rank condition requires (a lot of) matrix algebra will not be covered here (we refer to e.g. Wooldridge 2010, Chapter 5). However, in the present case, where we have two equations, the rank condition requires that at least one of the exogenous variables excluded from the first equation must have a nonzero population coefficient in the second equation, and vice versa.

We conclude that the system, consisting of Eqs. (6.62) and (6.63) is identified. Equation (6.62) is identified by the inclusion of the exogenous variable a_t^P in Eq. (6.63), and Eq. (6.63) is identified by the inclusion of another exogenous variable, a_t^c, in Eq. (6.62).

Now that we have established that the structure of the system of equations allows us, in principle, to obtain unique parameter estimates, we return to the issue of resolving the endogeneity bias due to simultaneity in the next section.

6.6 Instrumental Variables Estimation

In the preceding section we mentioned that 2SLS is a special case of IV-estimation. The IV approach was developed around the following "estimation strategy" (Greene 2012, p. 262). Let us first assume that in the classical model (4.13):

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t$$

all of the variables x_{kt}, $k = 1, \ldots, K$, are correlated with ε_t (later in this section we will consider the case where some of the explanatory variables are exogenous). And imagine that there exists a set of variables $z_{\ell t}$, $\ell = 1, \ldots, L$, where $L \geq K$, such that the $z_{\ell t}$ are correlated with x_{kt} but *not* with ε_t. Then we can construct a consistent estimator $\beta' = (\beta_1, \ldots, \beta_k)$ based on the relationships between the x_{kt}, $z_{\ell t}$ and ε_t. In this manner the set of original regressors is *replaced* by a new set of regressors, that are based on so-called *instrumental variables*. Instrumental variables (or: instruments) are variables that are correlated with the endogenous regressors but uncorrelated with the disturbances. Instruments that satisfy the latter requirement are *valid*, and instrumental variables that are highly correlated with the endogenous regressors are *strong* instruments. Conversely, *weak* instruments are not strongly correlated with the endogenous variables.

To introduce IV estimation a bit more formally, we return to the K-variable model (4.15):

$$y = X\beta + \varepsilon.$$

The IV method is used when the assumption $\mathrm{E}(X'\varepsilon) = 0$ is violated. The matrix X may be substituted by a matrix Z such that $\mathrm{E}(Z'\varepsilon) = 0$. Thus, every column of the new matrix Z is uncorrelated with ε, and every linear combination of the columns of Z is uncorrelated with ε. If Z has the same number of predictor variables as X (i.e. $K = L$), the IV estimator is:

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'y. \tag{6.70}$$

In cases where the number of instruments L is larger than K, the number of variables in X, the matrix multiplications in Eq. (6.70) break down, and Z needs to be replaced

by a matrix with K columns. Several options to choose such a matrix can be found in the literature. One approach is to choose K linear combinations of the columns of Z. Of all the different linear combinations of Z that we might choose, $\hat{X} = Z(Z'Z)^{-1}Z'X$ turns out to be the most efficient:

$$\begin{aligned}\hat{\beta}_{IV} &= (\hat{X}'X)^{-1}\hat{X}'y \\ &= \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'y. \end{aligned} \tag{6.71}$$

It can be shown that this expression can also be written as:

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y. \tag{6.72}$$

Equation (6.72) expresses that this IV estimator for β is determined in two steps. First, the columns in X are fitted on the columns of Z. That is: $\hat{X}$ is computed as: $\hat{X} = Z(Z'Z)^{-1}Z'X$. In step 2, the estimate for β is determined as in the well-known OLS formula for β, see Eq. (4.19), with X replaced by $\hat{X}$. Because $\hat{X}$ is fitted on Z, it can be considered as a function of Z, so that $\mathrm{E}(\hat{X}'\varepsilon) = 0$, which can easily be verified. This procedure to obtain an IV estimator is also known as Two-Stage Least Squares, and is usually abbreviated to 2SLS or TSLS.

Let us now consider the case where some of the explanatory variables are uncorrelated with ε (i.e. some of the explanatory variables are exogenous). All of the steps to determine the 2SLS estimator still apply if the matrix Z in Eq. (6.72) is constructed in such a way that it contains all of the exogenous variables, next to the instrumental variables. This means that the matrix Z has a column for each instrument that is needed to resolve the problems associated with the endogenous explanatory variables, and it also contains a column for every exogenous explanatory variable in X. By constructing Z in this way, step 1 of the 2SLS procedure returns a matrix $\hat{X}$ that contains a column for every exogenous explanatory variable, and a column for every endogenous explanatory variable. Moreover, the columns in $\hat{X}$ associated with the exogenous variables are identical to the corresponding columns in X, because fitting an exogenous variable to (a set of variables including) itself results in a perfect "fit". Another interesting observation is that, next to the instrumental variables, also the exogenous variables are used to instrument the endogenous variables.

In many cases the instruments for the endogenous explanatory variables will be "new" variables that are not included in the original equation(s).[24] In time-series analyses quite often a lagged endogenous variable is chosen as an instrumental variable.

To illustrate the 2SLS procedure we return to the advertising example in the previous section. In Fig. 6.2 we observed a_t^P is helpful in tracing the demand function

[24] See, for example, Fischer et al. (2010); Hui et al. (2013).

(in fact, we discussed that a_t^P identifies the demand equation). Consequently, this variable is an obvious instrument for a_t. In step 1 we fit a_t to a_t^P:

$$\hat{a}_t = \hat{\pi}_0 + \hat{\pi}_1 a_t^P, \tag{6.73}$$

where $\hat{\pi}_0$ and $\hat{\pi}_1$ are the OLS estimates of the coefficients in a simple linear regression of a_t on a_t^P. In step 2 we determine the 2SLS estimators for the coefficients in Eq. (6.62) by applying OLS to:

$$q_t = \beta_0 + \beta_1 \hat{a}_t + \beta_2 a_t^c + \varepsilon_t. \tag{6.74}$$

Note that applying step 1 to a_t^c leaves a_t^c unaltered, as this is assumed to be an exogenous variable.

Tests for significance of individual coefficients that are the results of IV estimation can be performed by the conventional t-tests. An F-test can also be performed along the lines of Sect. 4.3.2, albeit that in the numerator of expression (4.33), the IV residuals have to be weighted. We refer to Heij et al. (2004, pp. 406–407) for a more profound discussion.

An IV approach that is equivalent to 2SLS in case of a linear specification, is the *control function approach*. The control function approach also consists of two steps. Similarly to the first step in the 2SLS approach, we also regress the endogenous variables on the instruments in the first step of the control function approach. However, now the residuals and not the fitted values are taken to step 2. In step 2, the residuals of the first step are simply added as explanatory variables to the original equation (i.e. the one containing the endogenous explanatory variable(s)), which is then estimated as if all variables were exogenous.

In the advertising example, we determine $\hat{\xi}_t = a_t - \hat{a}_t = a_t - \hat{\pi}_0 - \hat{\pi}_1 a_t^P$ in step 1 [compare Eq. (6.73)], and estimate in step 2 an extended version of Eq. (6.62) where $\hat{\xi}_t$ is added as explanatory variable:

$$q_t = \beta_0 + \beta_1 a_t + \beta_2 a_t^c + \beta_3 \hat{\xi}_t + \varepsilon_t. \tag{6.75}$$

Adding $\hat{\xi}_t$ to Eq. (6.62) 'controls' for the part of a_t that is related to ε_t, so that the estimate for β_1 is not subject to endogeneity bias.

When the control function approach is applied to nonlinear models, the first step does not change. The residuals of the first stage regression are added to the set of explanatory variables and the model is estimated using the same non-linear regression approach that would have been used if all explanatory would have been exogenous. In nonlinear models the control function approach performs much better than other IV methods (Luan and Sudhir 2010; Petrin and Train 2010).

Rossi (2014) notes that the control function has a lot of appeal for applied researchers, because all that is needed is a first stage linear regression in the instruments and simply add the residual (which are constructed variables) from this first stage to the linear model. He argues that, despite the ease of use of the

control function approach, the advantages are less clear from inference point of view, because the standard errors will be incorrectly estimated as they do not take into account that some of the variables are constructed.

One of the assumptions of the 2SLS-method is that the disturbances of the different equations of the system are independent. We know from the discussion about seemingly unrelated regressions in Footnote 8 of Sect. 6.2.4 that parameter estimates of the system are inefficient if we do not account for contemporaneous correlations of the disturbances. An estimator that makes use of the cross-equation correlations of the disturbances is more efficient. The techniques that are generally used for joint estimation of the simultaneous system of relations include: Three-Stage Least Squares (3SLS), maximum likelihood (Sect. 6.4) and GMM (Generalized Method of Moments). Thus the 3SLS method includes the application of GLS to the system of structural relations.

Briefly, 3SLS works as follows. The estimator of the elements of the variance-covariance matrix of the disturbances Ω are obtained by first applying 2SLS on the set of simultaneous equations. The 2SLS residuals are then used to compute $\hat{\Omega}$ which allows 3SLS estimates to be obtained with $\hat{\Omega}$ and the "new regressors" $z_{1t}, \ldots, z_{Lt}$. It is also possible to iterate the 3SLS estimation method: I3SLS.[25]

As an example we discuss the study by Carpenter (1987). Carpenter studied competitive marketing strategies consisting of product quality levels, promotional expenditures and prices. Using a simultaneous equation model he examined the interrelations between these marketing instruments. For example he let prices and promotional spending be signals of product quality and allowed promotional spending to influence prices. A model of interrelations between marketing instruments requires the use of simultaneous equation estimation methods. Carpenter used 3SLS (and 2SLS) on a cross section of business-level data.

Rossi (2014) draws attention to the limitations of instrumental variables estimators. He argues that validity of the instruments, which is a necessary condition for asymptotic unbiasedness of the estimators, is an unverifiable assumption. And even if the instruments are valid, he emphasizes that instrumental variable estimators can have poor sampling properties such as fat tails, high RMSE, and bias. He argues that in panel or time series data the consequences of misspecification of the functional form or distribution of the error term are likely to be much more problematic than potential endogeneity biases.

Consequently, one might wonder whether the use of instrumental variable estimation methods generates estimates that are preferred over, say, OLS estimates. Although there are examples of OLS estimates that are almost identical to the estimates generated by 2SLS or 3SLS,[26] the differences between estimates can be very large. These differences are quite often used for testing purposes. However,

[25]For applications, see Schultz (1971); Carpenter (1987); Tellis and Fornell (1988) and Hanssens et al. (2001, p. 90).

[26]Greene (2012, p. 372).

Rossi (2014) shows that invalid instruments can cause the estimates to differ even when there is no endogeneity bias, so that such differences cannot always be interpreted as evidence of endogeneity biases.

We conclude that IV estimators or other endogeneity corrections should not be applied by default. Instead, we advise to first consider whether these are needed (e.g. based on theory or earlier results), and subsequently test for the presence of endogeneity issues. If an endogeneity correction appears to be necessary, it is imperative that instruments are used that both valid and strong. Otherwise, the cure may be worse than the disease. At the end of the next section we discuss tests for assessing the validity and strength of instruments.

6.7 Tests for Endogeneity

There are several tests for endogeneity. Some authors prefer to test exogeneity, i.e. they test the null hypothesis $\mathrm{E}(X'\varepsilon) = 0$ (e.g. Heij et al. 2004, p. 409). Hence, the terms endogeneity test and exogeneity test are used interchangeably and depend on the specification of the null hypothesis.

Examples of endogeneity test are the Hausman test and the Wu test. We consider both tests in more detail below.

To introduce the Hausman test (Hausman 1978), we return to the 2SLS estimation method, a special case of the IV-method.[27] It can be proven that the 2SLS estimator is less efficient than OLS when the explanatory variables are exogeneous. Hence, it is useful to have a test for endogeneity of an explanatory variable that shows whether 2SLS is necessary.

We return to (6.62):

$$q_t = \beta_0 + \beta_1 a_t + \beta_2 a_t^c + \varepsilon_t$$

where a_t^c is exogeneous and a_t is a single "suspected" endogeneous variable, given that there may be relation such as (6.63). If a_t is uncorrelated with ε_t, it is optimal to estimate (6.62) using OLS. In case there is nonzero correlation, 2SLS is a better option. We can test this by directly comparing the OLS and 2SLS estimates, and determining whether the differences are statistically significant. The Hausman test is based on this idea, and utilizes a Wald statistic to test the difference between the two estimates:

$$H = \left(\hat{\beta}_{IV} - \hat{\beta}_{OLS}\right)' \hat{\Sigma}^{-1} \left(\hat{\beta}_{IV} - \hat{\beta}_{OLS}\right)' \tag{6.76}$$

where $\hat{\beta}_{IV}$ and $\hat{\beta}_{OLS}$ are the 2SLS and the OLS estimates, respectively and $\hat{\Sigma}$ is an estimate of the variance-covariance matrix of $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})$, which can be computed

[27] We closely follow Wooldridge (2012, pp. 512–513).

as $\hat{\Sigma} = \text{Cov}(\hat{\beta}_{IV}) - \text{Cov}(\hat{\beta}_{OLS})$. H follows a $\chi^2(K^*)$ distribution, where K^* is the number of endogenous explanatory variables in the regression. If the OLS and 2SLS estimates differ significantly we conclude that a_t must be endogeneous (maintaining that a_t^c is exogeneous).

Application of the Hausman test can be quite cumbersome, because the inverse of $\hat{\Sigma}$ may not exist. Therefore, in most practical cases, modelers use (asymptotically) equivalent tests that are easier to implement. For example, a simple regression test based on the control function approach can be used. The control function approach was discussed in the previous subsection, and significance of the estimate for β_3 in Eq. (6.75) provides evidence for the presence of endogeneity.

One can also test for endogeneity of multiple explanatory variables. This test is based on a joint test for the significance of including multiple "step 1 residuals" of the control function approach, one for each suspect endogenous explanatory variable. Joint significance indicates that at least one "suspected" explanatory variable is endogeneous.

Another well-known test for endogeneity is the test developed by Wu (1973). Wu proposes to apply an F-statistic to test the joint significance of the elements γ in the ("augmented") regression (6.77)

$$y = X\beta + \hat{X}^*\gamma + \varepsilon \tag{6.77}$$

where $\hat{X}^*$ are the fitted values in the regression of variables X^* on Z, where Z is the matrix that has been used in (6.70). X^* constitutes a matrix that contains L^* of the variables in X that are suspected to be endogenous. We refer for other issues regarding testing for endogeneity to, for example, Wooldridge (2012, pp. 534–535).

In the previous section we discussed that validity and strength are two important requirements for instrumental variables. We will now address how these requirements can be assessed.

Let us first consider the question whether the instrumental variables are valid, i.e. whether $\text{E}(\tilde{Z}'\varepsilon) = 0$, where $\tilde{Z}$ consists of the columns in Z that correspond to instrumental variables. As simple approach is to regress the estimated IV-residuals $e_{IV} = y - \hat{X}\hat{\beta}_{IV}$ on $\tilde{Z}$, the instrumental variables. The so-called Sargan-test on the validity of instruments is an appropriate test in this respect. The Sargan test is a Lagrange Multiplier (LM) test:

$$\text{LM} = nR^2$$

where R^2 results from the regression of the IV-residuals on the instrumental variables. Under the null hypothesis that the instrumental variables are exogenous, LM (asymptotically) is $\chi^2(r)$ distributed, where r is the number of instruments minus the number of endogenous variables. Failure to reject the Sargan test indicates that the instruments are not significantly correlated with the disturbance term, and

that there is no evidence that the instrument are invalid. Note that the Sargan test can only be applied when there are more instruments than the number of endogenous variables (Heij et al. 2004, p. 413).

Assuming that the validity requirement is met, let us now consider the strength of a set of instrumental variables. As we discussed in the previous section, instruments are strong if they are highly correlated with the endogenous variables, and tests have been developed that can be used to assess whether the relationship is strong enough. If there is only one endogenous explanatory variable, the first step of the 2SLS approach can be used to this end. In this first step, the endogenous variable is regressed on the instruments, and the F statistic that is normally used for evaluating the overall significance of the regression is now used as a test for the strength the instruments. An F-value larger than 10 indicates that the instruments are sufficiently strong (Stock et al. 2002).

If there are two or more endogenous variables one cannot simply apply this procedure sequentially to each endogenous variable because multicollinearity between the variables would invalidate the result (Greene 2012, p. 290). A procedure for calculating F that takes the correlations between the endogenous variables into account proceeds as follows (Greene 2012, p. 290). First we compute the matrices $A = (X'X)^{-1}$ and $B = (\hat{X}'\hat{X})^{-1}$, where $\hat{X}$ is the outcome of the first step of the 2SLS procedure. Then, for the k-th endogenous variable we determine

$$R_k^2 = \frac{A_{kk}}{B_{kk}} \tag{6.78}$$

where the indices indicate that we take the element in the kth row and the kth column. The F statistic is calculated as

$$F = \frac{\left(R_k^2\right)/(K)}{\left(1 - R_k^2\right)/(T - K - 1)} \tag{6.79}$$

where, as usual, T is the number of observations, and K is the number of explanatory variables.

6.8 Bayesian Estimation

6.8.1 Subjective Data

In Chap. 4 and in previous sections of this chapter we dealt with methods developed to extract parameter estimates purely from objective data, data that represent observed or observable quantities. In this subsection we return to the question that we raised in Sect. 3.5.7 how we can generate similar quantifications in the absence of objective data.

To justify the use of alternative bases for the quantification of relations, we refer to the arguments we have made in favor of model building (see Sect. 1.4). For example, by formalizing relations between variables we give ourselves, and others, an opportunity to reflect on the quantified expressions. We can voice our opinions about what we believe to be incorrect aspects and we start an internal discussion about the relations in question. Perhaps most importantly, the model builders can document the performance of models. Such documentation provides us with an opportunity to determine a model's actual performance. Without the formalization of predictions, conditional upon marketing activities, it is impossible to track the true quality of decisions made by managers. Thus, an important advantage associated with model use is that managers will apply a systematic procedure to test the accuracy of model-based predictions. This can be done for a single model by itself, for multiple models or for models in alternative objective or subjective bases for predictions.

In Sect. 3.5.7.1, we introduced the concept of "models of man". There is a large body of research on the success of "models of man". For example, a regression model of an admission director's judgments of academic performance for MBA students (as a function of their GMAT scores, undergraduate GPA's, undergraduate institution qualities, etc.) predicts actual performances better than the very same judgments on which the model is estimated. The reason for this result is simple: the model is consistent. It gives exactly the same prediction today, tomorrow or any time given a set of values for the predictors. The admission director, however, makes judgments that are subject to noise (or to conditions that do not relate to the academic performance of the students). If the admission director's task were to admit the applicants who are expected to have the strongest academic performance, the model of the director's judgments will tend to generate predicted values with greater accuracy than the judgments. This illustrates that predictions produced by a model based on subjective data generally outperforms predictions based on the subjective data itself.

Of course, we may argue that predictions from a model estimated with *objective data* can do even better. This should be true as long as the MBA program content and other aspects stay relatively constant over time. In that case data from students who have graduated can be used to obtain the parameter estimates that best explain their actual academic performance. The estimated model can then be used to predict the performance of future applicants. This model would of course also give exactly the same prediction any time it is used for a given set of values for the predictors. And this "model of outcomes" will outperform the "model of man" as long as the bases for actual performance remain the same.

However, in the MBA application, past data may be insufficient if the students' characteristics change dramatically or the curriculum and/or the requirements for good performance are very different than in earlier times. In general, data insufficiency may also be due to a lack of variability in one or more predictors, excessive covariation between predictors or other severe problems. Also, the competitive environment for marketing decisions can change dramatically, for example, after a new brand enters the market. Thus, historical data sometimes do not provide insight into the relations that pertain to the new environment. Most econometric models are

"static" in the sense that both the structure and the parameters are fixed (adaptive and time-varying parameter models are a notable exception). Thus, subjective estimation may not just be attractive in the absence of objective data but also to overcome the limitations of historical data.

6.8.2 *Combining Objective and Subjective Data: Bayes' Theorem*

As a result of the above discussed benefits, the combination of subjective and objective data has seen much attention in (marketing) modeling literature. Blattberg and Hoch (1990) suggest that model-based systems for decision making should be a hybrid system of 50 % judgment (manager) and 50 % model (customer). They find that this mix does better than either model or manager alone. A hybrid system may correct for human shortcomings in information processing while the human picks up on cues, patterns, and information not incorporated in the model (Bucklin et al. 1998, p. 6).

Combining objective and subjective data and estimates can be accomplished by informal analysis or formal analysis. Informal analysis can be used to adjust or update empirically determined coefficients. For example, parameter estimates from sample data can be adjusted by multiplying them by subjectively estimated indices. Alternatively, we can start with judgmental parameters. Little (1975) does this based on the idea that people tend to overinterpret patterns in historical data.

Combining subjective and objective information in a formal way is achieved by Bayesian Analysis.[28] We show a general framework for Bayesian analysis in Fig. 6.3. Suppose a firm wants to estimate the trial rate (θ) for a new product. Based

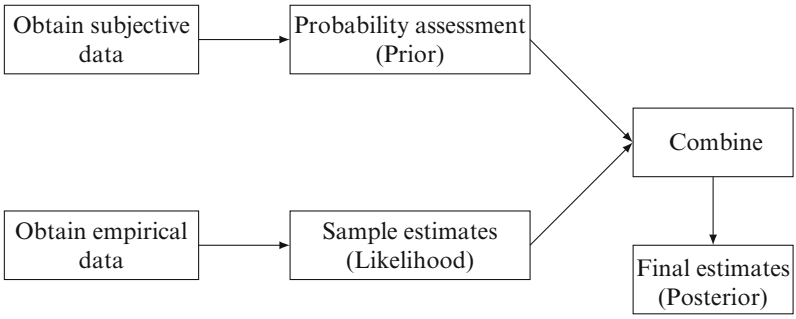

Fig. 6.3 A general framework for Bayesian analysis. *Source*: Naert and Weverbergh (1981, p. 107)

[28]For a much more detailed treatment of Bayesian methods we refer to Rossi et al. (2005). See also Congdon (2003, 2006).

on experience from launching other products in the same category, the trial rate is subjectively assessed. This subjective data called *prior* information and may be available in the form of a prior distribution $f(\theta)$.

The firm may also offer the product in a test market, thus obtaining sample data z. Let this sample evidence or objective data be available under the form of a sampling distribution $f(z \mid \theta)$. A decision to launch or not to launch would then be based on the posterior distribution $f(\theta \mid z)$ obtained from Bayes' Theorem:

$$f(\theta \mid z) = \frac{f(\theta) f(z \mid \theta)}{\int f(\theta) f(z \mid \theta) d\theta}. \tag{6.80}$$

Combining objective and prior information also applies to the estimation of response curves. Assume the following model:

$$y = X\beta + u \tag{6.81}$$

where β is a $K \times 1$ vector of unknown parameters. Without prior information β is estimated by, for example, ordinary least squares:

$$\hat{\beta} = (X'X)^{-1} X'y. \tag{6.82}$$

Suppose now that prior information is available on K' of the K parameters, i.e., a $K' \times 1$ vector of prior estimates $\hat{\beta}_p$. Prior information can be subjective, but also objective. It can originate from theory, expert judgments, analysis of other data sources, and/or from a meta-analysis of the existing empirical effect sizes (Sect. 1.3.2). So, for example, prior value of a price elasticity can be set equal to -2.62; i.e. the average value found in a meta-analysis by Bijmolt et al. (2005). Assuming these prior estimates to be unbiased, we have:

$$E(\hat{\beta}_p) = \beta_p. \tag{6.83}$$

The prior estimates are subject to error:

$$\hat{\beta}_{pk} = \beta_{pk} + \mu_k, \quad k = 1, 2, \ldots, K' \tag{6.84}$$

where μ_k is the error term for the kth estimate. Let the covariance matrix of the error term of the prior estimates be:

$$E(\mu\mu') = \Phi. \tag{6.85}$$

To formally combine objective and prior information, we write (6.84) in matrix form:

$$\hat{\beta}_p = A\beta + \mu, \quad \text{with } A = [I \quad O] \tag{6.86}$$

where I is an $K' \times K'$ identity matrix, and O is an $K' \times (K - K')$ matrix of zeros.[29] Combining (6.81) and (6.86) we obtain:

$$y^* = \begin{pmatrix} y \\ \beta_p \end{pmatrix} = \begin{bmatrix} X \\ A \end{bmatrix} \beta + \begin{pmatrix} u \\ \mu \end{pmatrix} \tag{6.87}$$

or

$$y^* = Z\beta + v$$

with

$$E(vv') = \begin{bmatrix} \sigma^2 I & O \\ O & \Phi \end{bmatrix}. \tag{6.88}$$

Equation (6.87) can be estimated by generalized least squares to obtain an estimate for β that combines prior information with objective information.

6.8.3 *Likelihood, Prior and Posteriors*

Key concepts of Bayesian estimation include the likelihood, the prior and the posterior.

To generalize from a sample to the population, both classic and Bayesian statisticians rely on the *likelihood principle*, i.e. all information in the sample is contained in the likelihood function. The likelihood function tells us how likely it is that we observe data z, given values for the parameters (see also Sect. 6.8). In general, the likelihood is defined by

$$L(\theta) = f(z|\theta) \tag{6.89}$$

where z represents the observed data, θ represents the unknown parameters, and f is some probability density function. Hence, we infer the likelihood of the parameter values θ, by assessing the probability of observing z, given those parameter values.

Instead of placing the uncertainty of such inference in the data, the Bayesian approach places the uncertainty in our lack of knowledge about *parameters* and operationalizes that lack of knowledge in terms of a (joint) probability distribution over all unknown quantities, that is, parameters. Before any data are observed, the Bayesian summarizes everything known about the model parameters in just such a single distribution, called the *prior distribution*. The prior distribution is

[29] We assume that prior information is available on the first K' estimators. The variables can be arranged such that this is the case.

then combined with (often called *updated by*) the likelihood, to yield the *posterior* distribution, which is then used for testing, prediction etc. A key Bayesian property is that the posterior distribution is proportional to the product of the likelihood and prior distribution [compare Eq. (6.80)]:

$$f(\theta|z) \propto f(z|\theta)f(\theta). \tag{6.90}$$

The prior distribution $f(\theta)$ reflects what we know (or do not know) about the parameters θ before consulting the data; the posterior distribution $f(\theta|z)$ reflects what we know about the parameters θ after combining both the observed data and the information contained in the prior. The shorthand notation of θ to denote the unknown parameters masks the strength of the Bayesian approach (Feinberg and Gonzalez 2012). Unknown quantities, such as missing data, can be construed as unknown parameters and included in θ.

6.8.4 Conjugate Priors

Bayesian analysis requires calculating and sampling from the posterior distribution. A key way to simplify this calculation is by selecting a prior distribution for which combining it with the likelihood yields the same form for the posterior distribution, i.e. a conjugate prior (Box and Tiao 1992). "Same form" means that prior and posterior belong to the same distributional family, differing only in their parameters. The most well-known example is that a Normal likelihood and a Normal prior will combine into a Normal posterior distribution. Likelihoods that belong to the exponential family have conjugate prior distributions, and often this is the Beta or Gamma. For example, in a model for count data with a Poisson likelihood (see Sect. 8.3), assuming a Gamma prior will result in a Negative Binomial distribution for the posterior.

6.8.5 Markov Chain Monte Carlo (MCMC) Estimation

Inferences within the Bayesian framework are based on the posterior distribution, which can be high-dimensional and very complex. Typically, researchers use summary measures such as expected value (means) and posterior standard deviation of the parameters θ. A key complication is that we can rarely sample from the desired posterior distribution immediately because this would require knowing approximately where the distribution is the largest, and computing the integral of the posterior distribution. These computations were very challenging before the development of simulation approaches such as Markov Chain Monte Carlo (MCMC) methods.

The key idea of MCMC is to sample each unknown parameter in turn, sequentially cycling through each unknown many times, always conditional on the latest draws for all the other parameters. Theorems show that such sampling will reach a stationary distribution for the parameters under rather general conditions. Practically, we choose a start point at random, let the simulation go, and, within several thousands of iterations reach a stationary distribution. Within MCMC, Gibbs sampling is popular when the distributions are known and easy to sample from, while Metropolis-Hastings sampling can handle "ill-behaved" distributions (Tanner 1996).

During MCMC estimation, diagnostic tests help identify when a stationary distribution is reached. If parameter estimates still "jump around" after hundreds of thousands of iterations, the researchers should check his data and model for sufficient variation in each variable and multicollinearity among the variables. Such multicollinearity is reflected in stretches of iterations where the model attributes an effect to a variable, followed by stretches of iterations where the model attributes that effect to another variable.

Parameter restrictions are easily incorporated into MCMC estimation, because Bayesian analysis is not concerned with the asymptotic properties of standard distributional forms (as frequentist statistical analysis is). For example, a diagonal covariance matrix can be imposed via the prior or within the sampling procedure by setting any parameter to a specific value (e.g. 0 for all off-diagonal elements of the covariance matrix), and sampling for all the other parameters conditional on the constraints.

6.8.6 Bayesian Analysis in Marketing

Marketing has seen numerous applications of Bayesian analysis on a very wide range of topics and various data types (see Rossi and Allenby 2003 for an overview till that year). Some of those are based on the science-philosophical debate of the Bayesian approach with classical or "frequentist" principles regarding the nature of statistical inference, the role of prior beliefs and claims about how one should update one's belief in the light of new information (Feinberg and Gonzalez 2012). Yet, Bayesian methods based on MCMC have become widely accepted in marketing science, and many studies have adopted or even developed Bayesian methods to study marketing phenomena. Practically, there are several conditions that suggest a Bayesian approach to marketing data and research questions. In our experience, three conditions stand out:

1. *Structurally missing data*: missing data are sometimes structural to the problem the researcher is interested in. For instance, during a product harm crisis, Kraft withdrew not just its affected brand, but also its flagship brand from the Australian peanut butter market for several months. Van Heerde et al. (2007) analyzed marketing effectiveness of Kraft's and the competing brands

before, during and after the withdrawal. Because both Kraft brands have zero sales during the withdrawal, a Bayesian estimation allows their prior (pre-crisis estimates) to go without updating while the competitors' estimates get updated, revealing an increased advertising impact.

2. *Shrinkage* of parameters over many cross-sections or other hierarchical data structures: often, marketing models need to be estimated over many cross-sections, such as consumers or stores of a retail chain. By chance, some of these section-specific estimates are bound to violate face validity, e.g. positive price elasticities for a fast moving consumer good at a retail store. Bayesian estimation allows shrinkage of the cross-section specific parameters towards the overall mean estimate (e.g. −2.62 across all stores of the chain). For an example, we refer to the next subsection.
3. *Varying parameters* based on systematic changes: while frequentist econometrics can handle time-varying parameters, Bayesian estimation shines when the parameter varies by a systematic factor, e.g. marketing spending. An example is the study by Ataman et al. (2008), where own marketing's effects on new brand base sales and price elasticities are determined using Bayesian estimation in a Dynamic Linear Model (West and Harrison 1997).

Thus, Bayesian estimation provides a general framework flexible enough to adapt to different types of data and research interests. The main costs of coding and computational complexity have been decreasing steadily over time.

6.8.7 Example: Bayesian Analysis of the SCAN*PRO Model

An interesting example of Bayesian analysis concerns the SCAN*PRO covered in Sect. 7.3.2.2. Andrews et al. (2008) investigate whether accounting for store-level heterogeneity in marketing mix effects (through Bayesian shrinkage) offers a substantial improvement over the original model, which assumes homogeneous effects. They consider a slight modification of the SCAN*PRO model[30] (Wittink et al. 2011):

$$q_{kjt} = \left[\prod_{r=1}^{n}\left(\frac{p_{krt}}{\bar{p}_{kr}}\right)^{\beta_{rj}}\prod_{\ell=1}^{3}\gamma_{\ell rj}^{D_{\ell krt}}\right]\left[\prod_{t=1}^{T}\delta_{jt}^{X_t}\right]\left[\prod_{k=1}^{K}\lambda_{kj}^{Z_k}\right]e^{\varepsilon_{kjt}} \tag{6.91}$$

where

q_{kjt} = unit sales (e.g. number of pounds) for brand j in store k in week t,

p_{krt} = unit price for brand r in store k, week t,

[30] See also Sect. 7.3.2.2.

$\bar{p}_{kr}$ = the median regular unit price (based on the non-promoted weeks) for brand r in store k,

D_{1krt} = an indicator variable for feature advertising: 1 if brand r is featured (but *not* displayed) by store k, in week t; 0 otherwise,

D_{2krt} = 0 an indicator variable for display: 1 if brand r is displayed (but *not* featured) by store k, week t; 0 otherwise,

D_{3krt} = an indicator variable for the simultaneous use of feature and display: 1 if brand r is featured *and* displayed; 0 otherwise,

X_t = an indicator variable (proxy for missing variables and seasonal effects): 1 if the observation is in week t; 0 otherwise,

Z_k = an indicator variable for store k : 1 if the observation is from store k; 0 otherwise

ε_{kjt} = a disturbance term for brand j in store k, week t.

The parameters in the model can be interpreted as follows:

β_{rj} = the own price (deal) elasticity if $j = r$, or cross-price elasticity if $j \neq r$,

γ_{1rj} = the own feature ad multiplier if $j = r$, or cross-feature ad multiplier if $j \neq r$,

γ_{2rj} = the own display multiplier if $j = r$, or a cross-display multiplier if $j \neq r$,

γ_{3rj} = the own display *and* feature multiplier if $j = r$, or a cross-display feature multiplier if $j \neq r$,

δ_{jt} = the (seasonal) multiplier for brand j in week t,

λ_{kj} = store k's regular (base) unit sales for brand j when there are no temporary price cuts and no promotion activities for any of the brands.

The size of model depends on:

n = the number of brands used in the competitive set,

K = the number of stores in the sample for a major market, and

T = the number of weeks.

The authors rewrite Eq. (6.91) as

$$q^*_{kjt} = X_{kjt}\beta_k + \varepsilon_{kjt} \tag{6.92}$$

where $q^*_{kjt} = \ln\left(q_{kjt}\right)$, the (log-transformed) marketing mix variables and weekly indicator variables are included in the matrix X_{kjt}, and the elasticities and log-transformed multipliers are included in the coefficient vector β_k.

In the Hierarchical Bayes (HB) formulation, the authors specify a multivariate normal population distribution to capture heterogeneity between stores as follows:

$$\begin{aligned} q^*_{kjt} &= X_{kjt}\beta_k + \varepsilon_{kjt} \\ \varepsilon_{kjt} &\sim N(0, \sigma_k^2) \\ \beta_k &\sim N(\bar{\beta}, \Lambda) \\ \bar{\beta} &\sim N(b_0, D_0) \\ \Lambda^{-1} &\sim W(\nu_0, S_0) \\ \sigma_k^{-2} &\sim G(a, b). \end{aligned} \tag{6.93}$$

The mean vector $\bar{\beta}$ represents the mean marketing mix elasticities and log multipliers in the population of stores, whereas the covariance matrix Λ captures the extent of heterogeneity and the correlation in marketing mix elasticities and multipliers between stores. The error variances σ_k^2 are store-specific. HB models require priors for the hyperparameters $\bar{\beta}$ and Λ and for the variances σ_k^2. It is customary to assume that σ_k^{-2} is distributed gamma $G(a,b)$, that $\bar{\beta}$ has a multivariate normal prior, $N(b_0, D_0)$, and that Λ^{-1} has a Wishart prior, $W(\nu_0, S_0)$. Andrews et al. (2008) found better fit when constraining Λ to be diagonal, indicating no covariances among parameters between stores. They chose the following values for the parameters of the priors: $a = 3$, $b = 1$, $b_0 = 0$, $D_0 = 10^3 I_r$, $\nu_0 = r + 2$, and $S_0 = (1/\nu_0) I_r$, where I_r is the identity matrix with dimension r. Because the univariate special case of the inverted Wishart distribution is the inverted gamma distribution, the diagonal elements of Λ^{-1} are assumed to have gamma priors. Given these 'well-behaved' distribution, the authors can use standard Gibbs sampling methods, allowing 3,000 iterations for burn-in.

Andrews et al. (2008) compare the fit of this Hierarchical Bayes model with that of a finite mixture model and the original Ordinary Least Squares (OLS) estimation. The empirical context is Dutch store-level scanner data in the shampoo product category. The estimation results show that the HB model has the best fit for all brands (both in Log Likelihood as in explanatory power. Compared to the original (OLS) SCAN*PRO model, the HB model has a 4–8 % higher R^2. As to predictive power, the HB estimation also beats the OLS estimation for most brands. Interestingly however, the HB estimation did not substantially improve parameter accuracy, so little may be lost by using the original SCAN*PRO model with Ordinary Least Squares estimation. The authors' explanation for this unexpected finding is that the large number of parameters required for the original model,

> "creating a difficult environment for recovering heterogeneity in parameters and resulting in decreasing returns from using additional parameters to explain more variation in sales" (Andrews et al. 2008, p. 31).

References

Aitken, A.C.: On least squares and linear combination of observations. Proc. R. Soc. Edinb. **55**, 42–48 (1935)

Andrews, R.L., Currim, I.S., Leeflang, P.S.H., Lim, J.: Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? Int. J. Res. Mark. **25**, 22–33 (2008)

Ataman, B.M., Mela, C.F., Van Heerde, H.J.: Building brands. Mark. Sci. **27**, 1036–1054 (2008)

Bijmolt, T.H.A., Van Heerde, H.J., Pieters, F.G.M.: Determinants of price elasticity: New empirical generalizations. J. Mark. Res. **42**, 141–156 (2005)

Blattberg, R.C., Hoch, S.J.: Database models and managerial intuition: 50% model + 50% manager. Manag. Sci. **36**, 887–899 (1990)

Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis, Wiley Classics Library edn. Wiley, New York (1992)

Bucklin, R.E., Lehmann, D.R., Little, J.D.C.: From decision support to decision automation: A 2020 vision. Mark. Lett. **9**, 234–246 (1998)

Cameron, A.C., Trivedi, P.K.: Microeconometrics Using Stata. Stata Press (2009)

Carpenter, G.S.: Modeling competitive marketing strategies: The impact of marketing-mix relationships and industry structure. Mark. Sci. **6**, 208–221 (1987)

Carpenter, G.S., Cooper, L.G., Hanssens, D.M., Midgley, D.F.: Modelling asymmetric competition. Mark. Sci. **7**, 393–412 (1988)

Clarke, D.G.: Sales-advertising cross-elasticities and advertising competition. J. Mark. Res. **10**, 250–261 (1973)

Cochrane, D., Orcutt, G.H.: Applications of least squares regression relationships containing autocorrelated error terms. J. Am. Stat. Assoc. **44**, 32–61 (1949)

Congdon, P.: Applied Bayesian Statistical Modelling. Wiley, Chichester (2003)

Congdon, P.: Bayesian Statistical Modelling, 2nd edn. Wiley, Chichester (2006)

Eliason, S.R.: Maximum Likelihood Estimation, Logic and Practice. Sage, Newbury Park (1993)

Feinberg, F.M., Gonzalez, R.: Bayesian modeling for psychologists: An applied approach. In: APA Handbook of Research Methods in Psychology, vol. 2. Research Designs. American Psychological Association (2012)

Fischer, M., Leeflang, P.S.H., Verhoef, P.C.: Drivers of peak sales for pharmaceutical brands. Quant. Mark. Econ. **8**(4), 429–460 (2010)

Fischer, R.A.: On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. A **222**, 309–368 (1922)

Gaver, K.M., Horsky, D., Narasimhan, C.: Invariant estimators for market share systems and their finite sample behavior. Mark. Sci. **7**, 169–186 (1988)

Greene, W.H.: Econometric Analysis, 7th international edn. Prentice-Hall, Upper Saddle River (2012)

Gupta, S.: Stochastic models of interpurchase time with time dependent covariates. J. Mark. Res. **28**, 1–15 (1991)

Hanssens, D.M., Parsons, L.J., Schultz, R.L.: Market Response Models: Econometric and Time Series Analysis. Kluwer Academic, Boston (2001)

Hausman, J.A.: Specification tests in econometrics. Econometrica **46**(6), 1251–1271 (1978)

Heij, C., De Boer, P.M.C., Franses, P.H., Kloek, T., Van Dijk, H.K.: Econometric Methods with Applications in Business and Economics. Oxford University Press, Oxford (2004)

Hui, S.K., Huang, Y., Suher, J., Inman, J.J.: Deconstructing the "first moment of truth": Understanding unplanned consideration and purchase conversion using in-store video tracking. J. Mark. Res. **50**, 445–462 (2013)

Judge, G.G., W. E. Griffiths R. C. Hill, H.L., Lee, T.C.: The Theory and Practice of Econometrics, 2nd edn. Wiley, New York (1985)

Kmenta, J.: Elements of Econometrics. Macmillan, New York (1971)

Leeflang, P.S.H.: Mathematical Models in Marketing, a Survey, the Stage of Development, Some Extensions and Applications. H.E. Stenfert Kroese, Leiden (1974)

Lindsey, J.K.: Parametric Statistical Inference. Clarendon Press, Oxford (1996)
Little, J.D.C.: BRANDAID: A marketing-mix model, part 2: Implementation, calibration, and case study. Oper. Res. **23**, 656–673 (1975)
Luan, Y.J., Sudhir, K.: Forecasting marketing-mix responsiveness for new products. J. Mark. Res. **47**, 444–457 (2010)
McGuire, T.W., Farley, J.U., Lucas, Jr., R.E., Ring, W.L.: Estimation and inference for linear models in which subsets of the dependent variable are constrained. J. Am. Stat. Assoc. **63**, 1201–1213 (1968)
Naert, P.A., Weverbergh, M.: Subjective versus empirical decision models. In: R.L. Schultz, A.A. Zoltners (eds.) Marketing Decision Models. North Holland, New York (1981)
Petrin, A., Train, K.: A control function approach to endogeneity in consumer choice models. J. Mark. Res. **47**, 3–13 (2010)
Prais, S.J., Houthakker, H.S.: The Analysis of Family Budgets. Cambridge University Press, Cambridge (1955)
Rossi, P.E.: Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. Mark. Sci. **34**, 655–672 (2014)
Rossi, P.E., Allenby, G.M.: Bayesian statistics and marketing. Mark. Sci. **22**, 304–328 (2003)
Rossi, P.E., Allenby, G.M., McCulloch, R.: Bayesian Statistics and Marketing. Wiley, Hoboken (2005)
Scales, L.E.: Introduction to Non-linear Optimization. MacMillan, London (1985)
Schultz, R.L.: Market measurement and planning with a simultaneous equation model. J. Mark. Res. **8**, 153–164 (1971)
Stock, J.H., Wright, J.H., Yogo, M.: A survey of weak instruments and weak identification in generalized method of moments. J. Bus. Stat. Econ. Stat. **1**(4), 518–529 (2002)
Tanner, M.A.: Tools for Statistical Inference, 3rd edn. Springer, New York (1996)
Tellis, G.J., Fornell, C.: The relationship between advertising and product quality over the product life cycle: A contingency theory. J. Mark. Res. **25**, 64–71 (1988)
Van Heerde, H.J., Helsen, K., Dekimpe, M.G.: The impact of a product-harm crisis on marketing effectiveness. Mark. Sci. **26**, 230–245 (2007)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: How promotions work: SCAN*PRO-based evolutionary model building. Schmalenbach Bus. Rev. **54**, 198–220 (2002)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Decomposing the sales promotion bump with store data. Mark. Sci. **23**, 317–334 (2004)
Van Nierop, J.E.M., Leeflang, P.S.H., Teerling, M.L., Huizingh, K.R.E.: The impact of the introduction and use of an informational website on offline customer buying behavior. Int. J. Res. Mark. **28**, 155–165 (2011)
West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models, 2nd edn. Springer, New York (1997)
Wittink, D.R., Addona, M.J., Hawkes, W.J., Porter, J.C.: SCAN*PRO: The estimation, validation and use of promotional effects based on scanner data. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in Honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)
Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data, 2nd edn. MIT, Cambridge (2010)
Wooldridge, J.M.: Introductory Econometrics: A Modern Approach, 5th edn. Cengage Learning, Mason (2012)
Wu, D.M.: Alternative tests of independence between stochastic regressors and disturbances. Econometrica **41**, 733–750 (1973)
Zellner, A.: An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J. Am. Stat. Assoc. **57**, 348–368 (1962)

Chapter 7
Examples of Models for Aggregate Demand

7.1 Introduction

In this chapter we give some examples of marketing models which have been estimated using the general linear model. Most of these models have been estimated using *aggregate* demand data. Aggregate demand refers to the demand across a sample of customers or households and can be measured at levels such as store, chain and market demand.

In Chap. 8 we discuss models that are estimated using other models than the general linear model. These models are briefly introduced in Sect. 3.4; see Table 3.1. Usually these models are calibrated using individual/household data and thus refer to what is called "individual demand". In this chapter we first briefly discuss relations between individual demand (models) and aggregate demand (models) in Sect. 7.2.

We continue to discuss descriptive/predictive models in Sect. 7.3. We discuss illustrations of three classes of demand functions: product category/product class/industry sales,[1] brand sales, and brand market share models. Such a classification is useful because model specification, both in terms of variables and mathematical form, can have features that are distinct for each of these three categories. In Sect. 7.3.1, we discuss product class sales models. Brand sales models can either be modeled separately as a function of decision- and environmental variables, or it can be obtained from the product of product class sales and market share. We illustrate both approaches in Sect. 7.3.2. A detailed analysis of market share models is given in Sect. 7.3.3, with particular attention to robust market share specifications.

Finally, we spend some attention to normative models in Sect. 7.4.

[1] We use the terms industry sales and product class (category sales) interchangeably.

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_7

7.2 An Introduction to Individual and Aggregate Demand

Aggregate demand models can either describe market behavior directly, or indirectly through individual behavior models from which outcomes are aggregated to determine market response. An aggregate response model, if postulated *directly*, is applied to aggregate data and has its own component of response uncertainty (represented by the properties of the error term). The probabilistic properties of the *indirectly* specified aggregate models are derived from the perspectives of the individual demand models.

The modeling of aggregate response from the addition of behavior across individuals ideally reflects heterogeneity across households in their intrinsic brand preferences and in their sensitivities to marketing variables (Chintagunta et al. 1991). Different approaches are available to account for household heterogeneity at the disaggregate level.[2] The question is how to accommodate differences across households in their brand preferences and sensitivities to marketing variables in models of aggregate data. Related aggregation questions apply to the aggregation of store data to the market level.[3] We will discuss these issues briefly below.

Some models have been developed which have about the same structure at the individual (micro) level and at the aggregate (macro) level. For example, Markov models are often used to accommodate the idea that the last brand chosen (in period t) affects the current purchase (in period $t+1$). A first-order Markov model applies when *only* the last purchase has an influence on the next one, i.e.

$$P(X_{t+1} = j \mid X_t = i, X_{t-1} = r, \cdots) = P(X_{t+1} = j \mid X_t = i) = p_{ijt} \tag{7.1}$$

where $P(X_{t+1} = j \mid X_t = i, X_{t-1} = r, \cdots)$ is the probability that the brand purchased at time $t+1$ is j, given that the brand purchased at t was i, at $t-1$ was $r, \cdots$. These probabilities are called *transition probabilities*. A *first-order* Markov model applies when only the most recent purchase has an influence on the current one. In a *stationary* first-order Markov model the transition probabilities are constant over time: $p_{ijt} = p_{ij}$. The transition probabilities are related to the (individual) brand choice probabilities ($\pi_{j,t+1}$) as:

$$\pi_{j,t+1} = \sum_{r=1}^{n} p_{rj}\pi_{rt}, \text{ for every } j = 1, \ldots, n, \quad t = 1, \ldots, T \tag{7.2}$$

[2] Guadagni and Little (1983); Kamakura and Russell (1989); Chintagunta (1992, 1993a); Gönül and Srinivasan (1993); Gupta and Chintagunta (1994); Rossi and Allenby (1994); Wedel and Kamakura (1998); Yang and Allenby (2000); Andrews and Currim (2009); Andrews et al. (2011).

[3] See Christen et al. (1997).

where

$\pi_{j,t+1}$ = probability brand j is chosen at $t+1$ by an individual consumer/household,

n = total number of brands.

The unconditional (π_{rt}) and conditional (p_{rj}) probabilities are distributed over the population of consumers. One may account for *heterogeneity* by making assumptions about these distributions.[4] Under the assumption of consumer homogeneity, i.e. consumers have the same p_{rj} and π_{rt} values, it can be demonstrated that relation (7.3) holds at the aggregate level:

$$m_{j,t+1} = \sum_{r=1}^{n} \tilde{p}_{rj} m_{rt}, \text{ for every } j = 1,\ldots,n, \quad t = 1,\ldots,T \tag{7.3}$$

where

$m_{j,t+1}$ = market share of brand j in period $t+1$,

$\tilde{p}_{rj}$ = fraction of consumers who buy brand r in t and brand j in $t+1$.

The fractions $m_{j,t+1}$, m_{jt}, $\tilde{p}_{rj}$ follow the multinomial distribution with means $\pi_{j,t+1}$, π_{jt} and p_{rj}. Although (7.2) and (7.3) have the same structure, the definitions of the variables are clearly different between the individual and the aggregate levels.

A second example of similarity in structure between individual- and aggregate models is the specification of the logit model (see Sects. 8.2.2 and 8.2.3). The multinomial logit model is a popular model in the marketing science literature since the pioneering research by Guadagni and Little (1983). We introduce the model specification at the individual/household level.[5] Consider an individual i confronted with a choice from a set of alternatives, CS_i, such as different brands in a product category. The utility that consumer i expects from alternative (brand) j is U_{ji}. This utility can be divided into two components, a systematic part (V_{ji}) and a random component (ε_{ji}). Thus:

$$U_{ji} = V_{ji} + \varepsilon_{ji}. \tag{7.4}$$

Given a specific set of alternatives, individual i chooses the option with the highest utility. The probability of choosing j is:

$$\pi_{ji} = P[U_{ji} > U_{ri}, \underset{r \neq j}{r, j} \in CS_i]. \tag{7.5}$$

[4] See Chap. 8.

[5] We closely follow Guadagni and Little (1983). See also Sect. 8.2.2.

If the ε_{ji} in (7.4) are independently distributed random variables with a double exponential distribution, then it can be shown[6] that individual i's choice probabilities have the (simple) form:

$$\pi_{ji} = e^{V_{ji}} / \sum_{r \in CS_i} e^{V_{ri}}. \tag{7.6}$$

The systematic, deterministic component of a consumer's utility for alternative j can be expressed as a linear function of observed variables relevant to j ($x_{\ell ji}, \quad \ell = 1, \ldots, L$):

$$V_{ji} = \alpha_{0j} + \sum_{\ell=1}^{L} \alpha_{\ell j} x_{\ell ji}. \tag{7.7}$$

Substituting (7.7) in (7.6) we obtain the expression for the multinomial logit model at the individual level:

$$\pi_{ji} = \exp\left(\alpha_{0j} + \sum_{\ell=1}^{L} \alpha_{\ell j} x_{\ell ji}\right) / \sum_{r \in CS_i} \exp\left(\alpha_{0r} + \sum_{\ell=1}^{L_r} \alpha_{\ell r} x_{\ell ri}\right) \tag{7.8}$$

where

$$L_r = \text{the number of predictor variables for alternative } r.$$

From the disaggregate logit model in Eq. (7.8), forecasts of aggregate demand and market shares can be obtained. The essence is that a prediction of the share of choices of a brand, needs to be computed from the individual-level choice probabilities. If the model would have been calibrated on the whole population of I consumers, this approach would be conceptually simple: the expected market share of brand j would equal the average of the choice probabilities of the I individuals in the population:

$$m_j = \sum_{i=1}^{I} \pi_{ji} / I. \tag{7.9}$$

However, in most cases the individual-level choice probabilities are not known for all individuals in the population, since the levels of the predictors are unknown outside the sample. Some researchers have developed solutions for this problem.[7]

The most important ways of representing heterogeneity in disaggregate models currently in use are through: (1) either a continuous or a discrete mixture distribution

[6] Theil (1969); McFadden (1974).

[7] See, for example, McFadden and Reid (1975).

of the parameters; see Volume II; (2) through the applications of Bayesian estimation methods[8]; (3) through the specification of a distribution for one or more of the parameters which describe consumer demand at the individual level: see Chap. 8.

7.3 Example of Descriptive/Predictive Models

7.3.1 Product Class Sales Models

Product class sales models explain the demand for a certain category through environmental variables and/or aggregated values of marketing instruments.

If product class sales (industry sales) is to be explained from *cross-sectional* data, the explanatory variables are often socio-economic and demographic variables, such as age, sex, education, occupation, income, location, family size, as well as marketing variables. However, the effects of marketing instruments cannot be obtained unless the marketing instruments vary across individuals, or groups of individuals, or cross sections. Many examples of these models can be found in the economic and econometric literature. See, e.g. Duesenberry (1949), Klein and Lansing (1955), Kapteyn et al. (1980), Kapteyn et al. (1997).

An example of a marketing model estimated with cross-sectional data is a model that captures differences in trade show effectiveness across industries, companies and countries. The authors of this model (Dekimpe et al. 1997) provide results about the effects of various show types and tactical variables (booth size, personnel) on observed performance.

In *time-series data*, the predictors often consist of both environmental and marketing variables. Examples of environmental variables are: population size, a weather index, and information on economic activity. To explain product class sales through marketing variables it is common for researchers to use aggregate values of marketing instruments, meaning that aggregation is performed over all brands that constitute the product class. Examples are total advertising expenditures, total number of retail outlets and average price.

An example of a product class model is a model that specifies the relation between industry sales and variables such as industry advertising expenditures in the West-German cigarette market. One of the *non-dynamic* relations is relation (7.10):

$$Q_t = e^{\alpha_0 + u_t} \left(\sum_{r=1}^{n} a_{rt} \right)^{\alpha_1} (C_t)^{\alpha_2} (QR_t)^{\alpha_3} (QP_t)^{\alpha_4} (QC_t)^{\alpha_5} \tag{7.10}$$

[8]See, for example, Sect. 6.8, Andrews et al. (2008) or Andrews et al. (2011).

where

Q_t = total number of cigarettes in t,

u_t = a disturbance term,

a_{rt} = advertising expenditures of brand r,

C_t = household consumption in t,

QR_t = industry sales of "roll your own" tobacco, in units in t,

QP_t = industry sales of pipe tobacco, in units in t, and

QC_t = industry sales of cigars, measured units in t.

Table 7.1 Parameter estimates and statistics of relation (7.10)

Variable	Parameter estimate	Standard error
Advertising (α_1)	−0.04[b]	0.02
Household consumption (α_2)	−0.60[a]	0.09
Roll your own tobacco (α_3)	−0.07[b]	0.03
Pipe tobacco (α_4)	−0.07	0.04
Cigars (α_5)	−0.08	0.05
R^2	0.85	

[a]Estimates significant at the 1 % level
[b]Estimates significant at the 5 % level
Source: Leeflang and Reuyl (1985a, p. 96)

All variables are defined per capita (per person over 15 years of age). After all variables were regressed against time (to remove trends), the estimated parameters for the (adjusted) per capita variables in (7.10), obtained from monthly observations, are those in Table 7.1. Advertising has a significant effect on industry sales. In another part of their analysis (not shown here) Leeflang and Reuyl demonstrate that this effect diminishes over time. The estimated coefficient for household consumption (per capita) indicates that the consumption of cigarettes is quite responsive to household consumption, albeit that a percentage increase in household consumption leads to a smaller percentage increase in cigarette consumption. The cross elasticities are all negative. The cross elasticity for roll your own tobacco is the only significant cross elasticity.

Several studies have investigated the effects of pharmaceutical marketing efforts on primary demand.[9] We discuss a primary demand model for a pharmaceutical category (k) developed by Fischer and Albers (2010). In this model primary demand

[9]See, for example, Narayanan et al. (2004); Chintagunta and Desiraju (2005); Manchanda et al. (2005); Venkataraman and Stremersch (2007); Kremer et al. (2008); Leeflang and Wieringa (2010); Wieringa and Leeflang (2013); Ding et al. (2014); Wieringa et al. (2014).

is explained by aggregate values of marketing instruments and environmental variables, such as seasonal dummy variables and the time that has been elapsed since the launch of the first brand in the category k. The model has the following structure for category k $(1, 2, \ldots, K)$ and time period t $(1, 2, \ldots, T)$:

$$\ln Q_{kt} = \alpha_0 + \alpha_1 \ln DET_{kt} + \alpha_2 \ln \sum_{r=1}^{n} a_{rkt} + \alpha_3 \ln DTC_{kt} + \alpha_4 \ln PRC_{kt}$$
$$+ \alpha_5 LCT_{kt} + \alpha_6 \ln Q_{k,t-1} + \sum_{h=1}^{3} \ln \beta_h S\,D_h t + \sum_{m=1}^{K-1} \ln \beta_{3+m} \delta_k + w_{kt} \qquad (7.11)$$

where

Q_{kt} = unit sales of category k in period t,

DET_{kt} = total expenditures on detailing in category k in period t,

a_{rkt} = journal advertising expenditures of brand r in category k in t,

DTC_{kt} = total expenditures on direct to consumer advertising in category k and period t,

PRC_{kt} = unit sales-weighted average price in category k in t,

LCT_{kt} = elapsed time since launch of category k in period t,

$S\,D_{ht}$ = seasonal dummy variable for quarter h and period t,

δ_k = category dummy variable for category k,

w_{kt} = a disturbance term which is assumed to be normally distributed: $N(0, \sigma_k^2)$, where σ_k^2 is a category specific error variance.

Fischer and Albers (2010) obtained quarterly data from IMS Health for 86 categories ($K = 86$) in the United States over a 5 year observation period. The IMS MIDAS database offers information about sales, expenditures on detailing (sales force efforts), professional journal advertising, Direct-To-Consumer (DTC) advertising and revenues. In addition, Fischer and Albers collected data about product launch dates. Using this data and other models they found mean short term elasticities of 0.033 for detailing and 0.004 for both journal advertising and DTC advertising. Long-term elasticities appear to be substantially larger, at least for journal advertising (0.036) and DTC advertising (0.041). The mean value of the long-term detailing elasticity is 0.047, which is only slightly larger than the corresponding short-term elasticity. These means have distributions with relatively large standard deviations. For a substantial number of categories the estimated elasticities are negative.

Many other examples of industry sales demand models can be found in the literature. See e.g. Leone and Schultz (1980); Lancaster (1984); Campo et al. (2000); Nijs et al. (2001); Ma et al. (2011).[10]

7.3.2 Brand Sales Models

7.3.2.1 Introduction

Brand sales can either be modeled directly or indirectly. Directly means that sales of brand j are explained as a function of marketing variables of brand j, marketing variables of competing brands, and environmental variables. Indirectly means that brand sales (q_{jt}) obtains from the product of category sales (Q_t) and market share of the brand of interest (m_{jt}). Specification of product class sales was discussed in the previous section. We discuss market share models in Sect. 7.3.3.

Arguments in favor of modeling brand sales *indirectly* are:

1. It is possible to distinguish between changes in q_{jt} that are caused by changes in market size, Q_t, and those that come from changes in the relative position of brand j in that market, expressed by market share, m_{jt}.
2. Using market share rather than sales as the dependent variable has the following advantages: environmental variables, and seasonal or cyclical factors causing expansion or contraction of the entire market need not be included.[11] The share model concentrates attention on the competitive interactions between brands in the product class.
3. By phasing variables over different levels we reduce the potential number of explanatory variables per equation, which reduces multicollinearity.

Arguments in favor of modeling brand sales *directly* are:

1. To the extent that marketing activities for individual brands influence product category sales, it is implausible that those marketing effects are the same for equivalent increases across the brands that belong to a product category.
2. Product category sales result from the aggregation of sales across brands belonging to the category. Since brands are heterogeneous in marketing activities and tend to have unique parameters relating marketing variables to sales, the interpretation of product category demand model parameters is unclear.

[10]Calls for advertising bans in different areas such as alcohol and cigarettes continue to echo around the world on a continuing basis. This explains why so many product class models have been developed in these areas. See, e.g. Duffy (1996); Franses (1991); Leeflang and Reuyl (1995); Luik and Waterson (1996); Nelson (2006); Capella et al. (2011).

[11]The assumption being that such variables affect demand for each brand equally. This assumption will often be quite reasonable. If not, however, environmental variables affecting brands differently should be included in the market share function (as well as in the direct estimation of brand sales).

7.3.2.2 Modeling Brand Sales Directly: The SCAN*PRO Model

We illustrate the direct approach with the well-known SCAN*PRO model (Wittink et al. 2011) that we introduced in Sect. 6.8.7. This model uses brand sales as the criterion variable. The SCAN*PRO model is a store-level model developed to quantify the effects of promotional activities implemented by retailers on a brand's unit sales. The model accommodates temporary price cuts, displays, and feature advertising. In addition, it includes weekly indicator variables to account for the effects of seasonality and missing variables (such as manufacturer television advertising and coupon distributions) common to the stores in a metropolitan area, and store indicator variables. This model has been used in over 3,000 different commercial applications in the United States, in Canada, in Europe, and elsewhere. ACNielsen provides data, parameter estimates and possible scenarios for marketing managers.

A slight modification of the original model is specified as follows, for brand j, $j = 1,\ldots,n$ in store $k = 1,\ldots,K$ in week $t = 1,\ldots,T$:

$$q_{kjt} = \left[\prod_{r=1}^{n}\left(\frac{p_{krt}}{\bar{p}_{kr}}\right)^{\beta_{rj}}\prod_{\ell=1}^{3}\gamma_{\ell rj}^{D_{\ell krt}}\right]\left[\prod_{t=1}^{T}\delta_{jt}^{X_t}\right]\left[\prod_{k=1}^{K}\lambda_{kj}^{Z_k}\right]e^{\varepsilon_{kjt}} \tag{7.12}$$

where the variables were defined in Sect. 6.8.7.

This model has been numerically specified in a study of the model's forecasting accuracy at different levels of aggregation (store, chain, market-level). We consider here the parameter estimates obtained from store-level data only. These estimates are obtained from scanner data provided by ACNielsen, for one large metropolitan area in the United States. Weekly data were available for three national brands competing in a frequently purchased food category. The average values of the significant parameter estimates are shown in Table 7.2. The averages are averages over three brands and the 40 stores in the sample.[12]

As expected, the own-brand price elasticity is negative, and the cross-brand elasticity is positive. The (promotion) multipliers with a value larger than 1 have a positive effect on unit sales, while values smaller than 1 have a negative effect. All cross effects, except feature, have the expected negative impact on q_{kjt}.

The SCAN*PRO model has been used in several studies in which aggregation effects are considered.[13] The model also constitutes the basis for the development of varying parameter models,[14] semiparametric models[15] (see Volume II) and models for the effects of dynamic lead- and lag effects.[16]

[12] See Foekens et al. (1994).

[13] Foekens et al. (1994); Gupta et al. (1996); Christen et al. (1997).

[14] Foekens et al. (1999).

[15] Van Heerde et al. (2001).

[16] Van Heerde et al. (2000); Andrews et al. (2008).

Store-level SCAN*PRO regression models have been developed to decompose the sales promotion bump into three parts: cross-brand effects, cross-period effects and category expansion effects. Across four different store-level datasets it is found that each of these three parts constitutes about one third on average (Van Heerde et al. 2004).[17] The SCAN*PRO model has been extended in different steps in a number of ways. These extensions are results of evolutionary model building.[18]

Table 7.2 Average values of parameter estimates of the SCAN*PRO model

Own-brand effects				Cross-brand effects			
		Feature				Feature	
Feature	Display	and display	Price	Feature	Display	and display	Price
1.63	2.25	3.07	−3.50	1.25	0.87	0.90	0.63

Source: Foekens et al. (1994, p. 260)

7.3.2.3 Modeling Brand Sales Directly: Models for Pharmaceutical Markets

We now consider examples of brand sales models which have been developed to model the effects of pharmaceutical marketing. They illustrate several modeling issues that have been discussed in the previous chapters.[19]

Rizzo (1999) is one of the first researchers who investigates the effects of marketing expenditures on:

- pharmaceutical brand sales, and on
- the price elasticity of demand.

The latter topic has received considerable scholarly attention in the past decade because it is postulated that marketing activities may reduce the price elasticity of demand, which allows firms to charge higher prices for their drugs. This has unwanted welfare effects and is sometimes referred to as the *persuasive effect* of marketing. Studies by Rizzo (1999) and Windmeijer et al. (2005) find that marketing reduces the price elasticity of demand and also increases demand. Rizzo (1999) estimates the effects of promotional efforts on the demand and price elasticity of demand of anti-hypertensive drugs with the following model:

$$\begin{aligned}\ln q_{it} &= \alpha_0 + (\alpha_1 + \alpha_2 \ln d_{it} + \alpha_3 \ln D_{it}) \ln p_{it} + \alpha_4 \ln d_{it} \\ &\quad + \alpha_5 (\ln d_{it})^2 + \alpha_6 \ln D_{it} + \alpha_7 (\ln D_{it})^2 + \beta X_{it} + u_{it} \end{aligned} \tag{7.13}$$

[17] See also Van Heerde et al. (2003); Leeflang et al. (2008).

[18] See Van Heerde et al. (2002).

[19] The following text is based on Wieringa and Leeflang (2013). See also Leeflang and Wieringa (2010).

where for pharmaceutical product i in year t:

q_{it} = quantity sold (defined in daily doses),

d_{it} = detailing efforts (flow) [see also Eq. (2.49)],

D_{it} = detailing stock, $D_{it} = \rho_d D_{i,t-1} + d_{it}$,
with ρ_d as the carryover coefficient for detailing,

p_{it} = wholesale price of an initial dosage, such as the dose
for 100 days' worth of therapy,

X_{it} = a vector of variables that include competitor prices
and detailing efforts, the length of time the drug has been
on the market, and other product-specific factors.

The parameter α_1 is known as the pre-marketing own-brand price elasticity. If $\alpha_2 > 0$, detailing flow lowers the price elasticity of demand, whereas if $\alpha_3 > 0$, the detailing stock lowers the price elasticity of demand. The model accommodates the indirect effects of detailing on demand through price elasticity with parameters α_2 and α_3, and allows for direct demand effects of both detailing flow and stock on sales with the parameters $\alpha_4 - \alpha_7$.

Rizzo (1999) estimates (non-brand specific) parameters of (7.13) using annual data about 46 antihypertensive drugs, including ACE inhibiters, calcium channel blockers, beta blockers, and diuretics. The sample consists of 222 observations. He finds that the pre-marketing own-brand price elasticity is negative and significant ($\hat{\alpha}_1 < 0$), for various specifications of the model, and that detailing efforts systematically lower price sensitivity, because $\hat{\alpha}_2 > \hat{\alpha}_3 > 0$. The results also suggest that sales increase directly with detailing efforts ($\hat{\alpha}_5 > \hat{\alpha}_6 > 0$).

Windmeijer et al. (2005) modify this model slightly for their study for the Netherlands' Bureau for Economic Policy Analysis. They calibrate a Rizzo-type model using monthly data from the Dutch market for 140 products from 11 therapeutic markets. Their findings are similar to those of Rizzo (1999), though not quite as strong. They also find a (slightly) negative average pre-marketing own-brand price elasticity, but it is significant only at the 10 % level. On average, Windmeijer et al. (2005) find that pharmaceutical marketing reduces price sensitivity. Pharmaceutical marketing (flow and stock variables) also increases demand, but this effect weakens at higher levels of marketing expenditures.

Through a critical evaluation of the frequently cited and often applied Rizzo model, we specify four concerns. First, the model pools data over brands and markets, which means that it assumes:

- the marketing expenditures of all brands in all categories have the same effect on demand;
- the marketing expenditures of all competitive brands have the same effect on demand (i.e. the competitive structure of each product category is the same).

Second, the assumption that pooling is allowed is not formally tested in applications of Rizzo's model. Third, the model does not account for varying parameters over the brand's life cycle. Fourth, some applications of the model (De Laat et al. 2002; Windmeijer et al. 2005) effectively assume that all marketing expenditures (e.g., direct mail, detailing, journal advertising) have the same effect on demand, because these expenditures are aggregated. To investigate the first two concerns initially, we apply Rizzo-type models at the *individual brand* level and test whether pooling is supported.

We consider five relevant submarkets of prescribed drugs in the Dutch market[20]: ulcers, hypertension, cholesterol, depression, and asthma. We use monthly data covering the period 1994–2001 for the brands in each submarket. We obtained data about expenditures on detailing efforts, medical journal advertising, and direct mail (in Euros), whereas Rizzo (1999) only considers expenditures on detailing efforts. Therefore, we replace d_{it} and D_{it} by mef_{it} (*m*arketing *e*xpenditures *f*low of brand i in month t) and mes_{it} (*m*arketing *e*xpenditures *s*tock of brand i in month t), respectively. Our results show that variation in the parameter estimates is substantial across the 84 brands, which is in line with findings of Venkataraman and Stremersch (2007). In Table 7.3, we summarize for each parameter the variation in brand-specific estimation results. The "mean" column indicates the mean value of the 84 parameter estimates, the columns labeled $p_{.05}$ and $p_{.95}$ show the 5th and the 95th percentiles of the set of parameter estimates, and the column labelled "fraction significant and correct sign" indicates the fraction of parameters that are statistically different from 0 and have the right sign.

Table 7.3 Brand-level results of Rizzo's (1999) model

Coefficient of	Mean	$p_{0.05}$	$p_{0.95}$	Fraction significant and correct sign
$\ln p_{it}$ (price)	−0.08	−2.60	3.18	0.13
$\ln mef_{it} \times \ln p_{it}$ (interaction flow)	−0.25	−1.79	0.96	0.05
$\ln mes_{it} \times \ln p_{it}$ (interaction stock)	0.20	−0.57	1.80	0.12
$\ln mef_{it}$ (marketing flow)	0.001	−0.85	0.94	0.08
$(\ln mef_{it})^2$ (marketing flow)^2	−0.01	−0.04	0.02	0.12
$\ln mes_{it}$ (marketing stock)	−0.29	−2.75	1.04	0.10
$(\ln mes_{it})^2$ (marketing stock)^2	0.02	−0.02	0.14	0.07

Source: Wieringa and Leeflang (2013, p. 3392)

To determine whether pooling of individual brand-level outcomes is allowed, as assumed by Rizzo (1999) and Windmeijer et al. (2005), we conduct a Roy–Zellner pooling test separately for each of the five submarkets. We show the F-values and the corresponding critical values in Table 7.4, which strongly indicate that pooling for each of the five submarkets is not appropriate. Accordingly, pooling over all the

[20]These markets also have been studied by Windmeijer et al. (2005).

categories is not allowed either, and estimates of pooled relations cannot be used to characterize brand behavior.

Table 7.4 Roy–Zellner test per submarket

Submarket	*F*-value	Critical value
Ulcers	611.9	1.16
Hypertension	1070.7	1.08
Cholesterol	578.5	1.20
Depression	896.2	1.17
Asthma	386.5	1.39

Source: Wieringa and Leeflang (2013, p. 3392)

Because pooling is not allowed, we perform separate analyses for established and new brands. We define established brands as those on the market before January 1, 1994, which is the first observation in our sample. New brands are those introduced during the observational period.

We estimated the effects of pharmaceutical marketing expenditures on demand for 64 *established brands*. The effects center close to zero; the mean is negative. For only eight out of 64 established brands promotional expenditures have a positive and significant effect on demand. This finding may emerge because sales of established pharmaceuticals show modest growth, and marketing expenditures tend to decline over time.

In contrast with Rizzo (1999) and Windmeijer et al. (2005), we thus conclude that for established brands, pharmaceutical marketing does not increase demand on average, and therefore, we cannot conclude that marketing expenditures have an effect on either category demand (market making) or substitute brands (market stealing).

With respect to the role of price, we note that pharmaceutical prices are heavily regulated in the Netherlands, so managers of branded pharmaceuticals cannot engage in price competition, even after patent expiration. Nevertheless, considering Rizzo's (1999) and Windmeijer et al.'s (2005) focus on the role of price, we estimate both pre- and post-marketing prices elasticities. The average pre-marketing price elasticity is very close to 0.

Of the brands with negative pre-marketing price elasticities, four reveal a substantially positive interaction effect, and one has a substantially negative interaction effect. These findings do not support the conclusion offered by Rizzo (1999) and Windmeijer et al. (2005) that marketing tends to reduce the absolute value of the price elasticity to 0. Thus, we cannot confirm the presence of persuasive effects of marketing efforts in these markets.

Overall, our results for established brands provide no evidence in support of Rizzo's conclusion that (1) marketing expenditures systematically lower price sensitivity and (2) sales increase with marketing expenditures.

We estimated several models do determine the effects of marketing expenditures on the sales of *new brands*. Diffusion models with marketing variables (see Volume

II) outperform other models. The parameter estimates of these models also confirm the finding of the established brands that the effects of marketing efforts on sales are significant for only a small portion of brands and tend to be moderate in size (Wieringa and Leeflang 2013, p. 3398).

7.3.2.4 Modeling Brand Sales Indirectly

By definition, brand sales, q, equals product class sales, Q, times market share, m.[21] To *model brand sales indirectly,* we closely follow the derivation in Lambin et al. (1975). They observed that for profit maximization the Dorfman–Steiner theorem (derived in the Appendix to this chapter) remains valid independent of whether the market is a monopoly or an oligopoly, making a separate derivation for each case unnecessary.[22] For an oligopoly, however, brand sales elasticities can be decomposed. We illustrate this below for the brand sales advertising elasticity.[23] We formulate the relation between brand sales-, product class sales-, and market share elasticities as follows.

$$q = Qm. \tag{7.14}$$

Differentiating brand sales with respect to advertising (a) gives,

$$\frac{\partial q}{\partial a} = m\frac{\partial Q}{\partial a} + Q\frac{\partial m}{\partial a}. \tag{7.15}$$

Multiplying both sides by a/q, we obtain,

$$\frac{a}{q}\frac{\partial q}{\partial a} = m\frac{a}{q}\frac{\partial Q}{\partial a} + Q\frac{a}{q}\frac{\partial m}{\partial a} \tag{7.16}$$

which can also be written as,

$$\frac{a}{q}\frac{\partial q}{\partial a} = \frac{a}{Q}\frac{\partial Q}{\partial a} + \frac{a}{m}\frac{\partial m}{\partial a} \tag{7.17}$$

or

$$\eta_{q,a} = \eta_{Q,a} + \eta_{m,a}. \tag{7.18}$$

[21] The time and brand index are omitted for notational convenience.

[22] In normative marketing mix studies one generally seeks the optimal policy for one brand assuming particular competitive reaction patterns. This means that one does not derive a simultaneous optimum for all brands in the product class. The latter would call for a game theoretical approach. We discuss game theoretical approaches in Volume II.

[23] For a more formal treatment, extending to other variables as well, see Lambin et al. (1975, pp. 106–115). In that paper the special character of quality as a decision variable is also discussed. A generalization to multiproduct markets is given by Bultez (1975), whereas Plat and Leeflang (1988) extend this model to account for more segments.

i.e., the brand sales elasticity with respect to advertising, $\eta_{q,a}$, is equal to the total product class sales elasticity, $\eta_{Q,a}$, plus the market share elasticity, $\eta_{m,a}$, with respect to the same variable. Thus the brand sales elasticity can be obtained *indirectly* from the sum of the product category sales- and the brand market share elasticities.

Relation (7.18) can be extended to account for competitive reactions. These reactions with the same (advertising) or other marketing instruments (for example, price) are called *indirect* effects. They may influence $\eta_{Q,a}$ and/or $\eta_{m,a}$, as will be discussed in Volume II.

7.3.3 Market Share Models

7.3.3.1 Attraction Models

Market share models can be specified to be logically consistent, in the sense that predicted values satisfy range (being between zero and one) and sum (summing to one across brands) constraints. One class of models that satisfy these constraints are the *attraction models*. The attraction of a brand depends on its marketing mix. Let A_{jt} be the attraction of brand j in period t (note that this symbol is also used for total advertising expenditures). Market share attraction models are defined as:

$$m_{jt} = \frac{A_{jt}}{\sum_{r=1}^{n} A_{rt}} \tag{7.19}$$

where n is the number of brands on the market. If A_{jt} is specified to be nonnegative, the attraction model has the desirable characteristics of both satisfying the range constraint ($0 \leq m_{jt} \leq 1$ for all j), and the sum constraint ($\sum_{r=1}^{n} m_{rt} = 1$). In this section, we focus on this type of model.

Equation (7.19) represents the overall structure. The attraction function itself remains to be specified. We present six different specifications. Two well-known market share specifications are the MCI model and the MNL model. The attraction for brand j in the "*Multiplicative Competitive Interaction*" (MCI) model is specified as[24]:

$$A_{jt} = \exp\left(\alpha_j\right) \prod_{\ell=1}^{L} x_{\ell jt}^{\beta_\ell} \varepsilon_{jt} \tag{7.20}$$

where

$$x_{\ell jt} = \text{the value of the } \ell\text{-th explanatory variable for brand } j, \text{ in period } t,$$

[24]The MCI models have been developed by Nakanishi and Cooper (1974, 1982). These and other market share models are discussed extensively in Cooper and Nakanishi (1988).

ε_{jt} = a disturbance term,

L = the number of marketing instruments.

Throughout it is assumed that L is independent of j.

The attraction for the MultiNomial Logit (MNL) market share model is specified as:

$$A_{jt} = \exp\left(\alpha_j + \sum_{\ell=1}^{L} \beta_\ell x_{\ell jt} + \varepsilon_{jt}\right). \tag{7.21}$$

The structure of this component is similar to the specification of the numerator of the MNL model at the individual level (7.8).

While the attraction specification (7.20) has attractive characteristics, there are also two disadvantages. First, the attraction is zero if one of the explanatory variables (for example, advertising) is zero in period t. This problem does not apply to the MNL model (7.21). Second, the response parameter for instrument ℓ is β_ℓ, which is assumed to be equal for each brand. Marketing executives, in general, find the assumption of equal response parameters across brands unacceptable.

An extension that allows a variable's response to vary across brands is the *extended attraction model* [versus the *simple attraction model* (7.20) and (7.21)]. This model is also known as the *differential effects model*[25]:

MCI-Differential Effects model: (MCI-DE):

$$A_{jt} = \exp\left(\alpha_j\right) \prod_{\ell=1}^{L} x_{\ell jt}^{\beta_{\ell j}} \varepsilon_{jt}. \tag{7.22}$$

MNL-Differential Effects model: (MNL-DE):

$$A_{jt} = \exp\left(\alpha_j + \sum_{\ell=1}^{L} \beta_{\ell j} x_{\ell jt} + \varepsilon_{jt}\right). \tag{7.23}$$

Before we introduce the last two of the six attraction specifications, we present expressions for elasticities.

7.3.3.2 Own-Brand Elasticities

Market share elasticities need to be determined separately for each attraction model. That is, the formula depends on the attraction specification. The *direct* or *own* market share elasticities (e_j^ℓ) for models (7.20)–(7.23) are (ignoring time subscript t):

[25] Cooper and Nakanishi (1988, Chapters 3 and 5).

$$\begin{aligned}
\text{MCI}: e_j^\ell &= \frac{\partial m_j}{\partial x_{\ell j}} \times \frac{x_{\ell j}}{m_j} = \beta_\ell (1 - m_j) \\
\text{MNL}: e_j^\ell &= \beta_\ell (1 - m_j) x_{\ell j} \\
\text{MCI-DE}: e_j^\ell &= \beta_{\ell j} (1 - m_j) \\
\text{MNL-DE}: e_j^\ell &= \beta_{\ell j} (1 - m_j) x_{\ell j}.
\end{aligned} \tag{7.24}$$

Note that the four elasticities differ with regard to the homogeneity/heterogeneity of the marketing variable parameters and in the presence/absence of the marketing variable itself. Specifically, the DE versions have $\beta_{\ell j}$ (versus β_ℓ), indicating that the parameter is brand-specific (heterogeneous). Apart from this distinction, each elasticity expression includes a marketing effort responsiveness parameter and the share of the market not captured by the brand $(1 - m_j)$. Thus, even if the responsiveness parameters are homogeneous, the elasticities differ across brands according to the remaining share. The inclusion of this $(1 - m_j)$ term has the desirable property that the elasticity goes toward zero as the brand's market share goes to one.

The MNL-based elasticities differ from the corresponding MCI-based ones in the inclusion of $x_{\ell j}$, which measures the marketing effort for variable ℓ used by brand j. Holding market share constant, the elasticity expression shows that an increase in marketing effort, for $\beta > 0$, increases the elasticity. However, we know that market share is affected by marketing activities. Also, it is generally accepted that it becomes harder to gain share as the marketing effort increases.[26] The MNL-based elasticity expression implies that if own-brand market share increases proportionally faster than the marketing effort, the own-brand market share elasticity will decrease with increasing $x_{\ell j}$. Cooper and Nakanishi (1988, p. 35) find that MNL-based elasticities increase to a point and then decrease.

7.3.3.3 Cross-Brand Elasticities

We now turn to a discussion about *cross elasticities*, which are defined as:

$$e_{j,r}^\ell = \frac{\partial m_j}{\partial x_{\ell r}} \times \frac{x_{\ell r}}{m_j} \tag{7.25}$$

where $x_{\ell r}$ = value of marketing instrument ℓ of competitor $r, r \neq j$.

[26] We assume $\beta_\ell, \beta_{\ell j} \geq 0$ for all ℓ, j which applies to variables such as distribution, selling effort, advertising, and sales promotions. For variables such as price for which $\beta_\ell, \beta_{\ell j} \leq 0$ an analogous reasoning can be formulated.

The expressions for the cross elasticities of the four attraction models are:

$$\begin{aligned} \text{MCI} &: e^{\ell}_{j,r} = -\beta_{\ell} m_r \\ \text{MNL} &: e^{\ell}_{j,r} = -\beta_{\ell} x_{\ell r} m_r \\ \text{MCI-DE} &: e^{\ell}_{j,r} = -\beta_{\ell r} m_r \\ \text{MNL-DE} &: e^{\ell}_{j,r} = -\beta_{\ell r} x_{\ell r} m_r. \end{aligned} \tag{7.26}$$

The four cross-elasticity expressions have properties that are similar to the own-brand elasticities. The effect of r's activity on brand j's attraction is either homogeneous ($-\beta_{\ell}$) or heterogeneous ($-\beta_{\ell r}$), and MCI and MNL differ with regard to the exclusion (MCI) or inclusion (MNL) of the effort for instrument ℓ by brand r. Also, all expressions include m_r, implying that the cross elasticity is more negative (stronger) as r's market share is larger. The cross elasticity does not depend on the market share of the focal brand j that is affected by the marketing efforts of competitor r. However, the actual change in a brand's share varies, reflecting its current level. The new share of any brand other than r may be simply calculated by: new share of brand j = (1− new share of brand r) × old share of brand j, $j \neq r$.

Because the expressions (7.26) are independent of m_j, the effects of marketing variable $x_{\ell r}$ are distributed among its competitive brands *in proportion* to their market shares. The models (7.20)–(7.23) constrain the competition to being *symmetric*. This symmetry is the result of the *Independence of Irrelevant Alternatives* (IIA) assumption.[27] This assumption implies that the ratio of two (market) shares does not depend on the presence or absence of other choice alternatives. That this assumption holds for the models (7.20)–(7.23) can be demonstrated easily by taking the ratio of the market share attractions of two brands j and r. These ratios are independent of the other brands r', $r' \neq j, r$. The IIA-properties also hold for the individual choice models such as, for example, (7.8).

An equality of cross elasticities, symmetry between brands, does not fit what we tend to observe in the market place. Brands belonging to a given product category are usually not equally substitutable. For example, Blattberg and Wisniewski (1989) found that consumers who normally purchase brands with *low* regular prices (e.g. store brands) are sensitive to temporary price cuts for (national) brands with *high* regular prices. On the other hand, consumers normally purchasing these national brands tend to be insensitive to temporary price cuts for store brands.

There are a number of alternative ways to account for asymmetric competition.[28] Briefly, one possibility is to expand the differential effects models one step further. Relations (7.27) and (7.28) are (the numerators of) attraction models

[27] See Luce (1959); Debreu (1960); Ben-Akiva and Lerman (1985); Sethuraman et al. (1999).

[28] See Foekens (1995); Bronnenberg and Wathieu (1996); Cooper et al. (1996). Examples are the Cluster-Asymmetry Model (Vanden Abeele et al. 1990), the CCHM-model (Carpenter et al. 1988) and hierarchical models (Foekens et al. 1997).

with differential cross-competitive effects, called *Fully Extended Attraction (FEA) models*. We consider two versions of FEA models.

Fully Extended MCI model (FEMCI):

$$A_{jt} = \exp\left(\alpha_j\right) \prod_{\ell=1}^{L} \prod_{r=1}^{n} x_{\ell rt}^{\beta_{\ell jr}} \varepsilon_{jt}. \tag{7.27}$$

Fully Extended MNL model (FEMNL):

$$A_{jt} = \exp\left(\alpha_j + \sum_{\ell=1}^{L} \sum_{r=1}^{n} \beta_{\ell jr} x_{\ell rt} + \varepsilon_{jt} \right). \tag{7.28}$$

For the own- and cross elasticities the following expressions can be derived:

$$\begin{aligned} \text{FEMCI} \ &: e_{j,r}^{\ell} = \beta_{\ell jr} - \sum_{r'=1}^{n} \beta_{\ell jr'} \times m_{r'} \\ \text{FEMNL} &: e_{j,r}^{\ell} = \left(\beta_{\ell jr} - \sum_{r'=1}^{n} \beta_{\ell jr'} \times m_{r'} \right) x_{\ell r}. \end{aligned} \tag{7.29}$$

These formulas apply to both own- and cross elasticities. The own elasticities are obtained from (7.29) by letting $j = r$. The above expressions indicate that the effects of changes in marketing variables differ between the brands. Thus the FEMCI- and FEMNL-models account for asymmetric competition.[29]

In their monograph on market share models and market share analysis, Cooper and Nakanishi (1988, p. 18) raise the question:

> "Why, …, are the MCI and MNL models not used more extensively?"

given that these models are flexible in the parameters, are logically consistent, and can account for asymmetric competition.

> "The answer is that for a time both of these models were considered to be intrinsically nonlinear models, requiring estimation schemes which were expensive in analysts' time and computer resources. This, however, turned out to be a hasty judgment because these models may be changed into a linear model (in the model parameters) by a simple transformation" (Cooper and Nakanishi 1988, p. 28).

These transformations, known as log-centering, are briefly introduced below for the MNL model [Eqs. (7.19) and (7.21)]. First (7.21) is substituted in (7.19). Then taking logarithms on both sides yields:

[29] For a more thorough discussion see Cooper and Nakanishi (1988, pp. 62–65).

$$\log m_{jt} = \alpha_j + \sum_{\ell=1}^{L} \beta_\ell x_{\ell jt} + \varepsilon_{jt} - \log\left(\sum_{r=1}^{n} \exp\left(\alpha_r + \sum_{\ell=1}^{L} \beta_\ell x_{\ell rt} + \varepsilon_{rt}\right)\right). \tag{7.30}$$

Summing (7.30) over j ($j = 1, \ldots, n$) and dividing by n gives:

$$\log \tilde{m}_t = \bar{\alpha} + \bar{\varepsilon}_t + \frac{1}{n} \sum_{r=1}^{n} \sum_{\ell=1}^{L} \beta_\ell x_{\ell rt} - \log\left(\sum_{r=1}^{n} \exp\left(\alpha_r + \sum_{\ell=1}^{L} \beta_\ell x_{\ell rt} + \varepsilon_{rt}\right)\right) \tag{7.31}$$

where

$$\tilde{m}_t = \text{geometric mean of } m_{jt},\ \tilde{m}_t = \left[\prod_{r=1}^{n} m_{rt}\right]^{\frac{1}{n}},$$

$$\bar{\alpha} = \frac{1}{n} \sum_{r=1}^{n} \alpha_r,$$

$$\bar{\varepsilon}_t = \frac{1}{n} \sum_{r=1}^{n} \varepsilon_{rt}.$$

If we subtract (7.31) from (7.30) we obtain the following form, which is linear in the parameters:

$$\begin{aligned} \log m_{jt}/\tilde{m}_t &= \alpha_j - \bar{\alpha} + \sum_{\ell=1}^{L} \beta_\ell x_{\ell jt} - \frac{1}{n} \sum_{r=1}^{n} \sum_{\ell=1}^{L} \beta_\ell x_{\ell rt} + \varepsilon_{jt} - \bar{\varepsilon}_t, \\ &= \alpha_j^* + \sum_{\ell=1}^{L} \beta_\ell \left(x_{\ell jt} - \frac{1}{n} \sum_{r=1}^{n} x_{\ell rt}\right) + \varepsilon_{jt}^*. \end{aligned} \tag{7.32}$$

The intercepts α_j can be estimated up to an arbitrary constant. Hence, if we set α_1 equal to zero, (7.32) can be written as:

$$\log(m_{jt}/\tilde{m}_t) = \sum_{r=2}^{n} \left(d_r - \frac{1}{n}\right) \alpha_r + \sum_{r=1}^{n} \sum_{\ell=1}^{L} \left(d_r - \frac{1}{n}\right) x_{\ell rt} \beta_\ell + \varepsilon_{jt}^* \tag{7.33}$$

where $d_r = 1$ if $r = j$, and 0 otherwise.

Model (7.33) requires transformations of all the variables. There is an equivalent model of (7.33) which yields identical estimates for both intercepts and response parameters but does not require any variable transformation apart from taking logarithms of m_{jt}. This model has the form[30]:

$$\log m_{rt} = \alpha_0 + \sum_{r=2}^{n} d_r \alpha_r + \sum_{t'=2}^{T} D_{t'} \Theta_{t'} + \sum_{\ell=1}^{L} \beta_\ell x_{\ell rt} + \varepsilon_{rt} \tag{7.34}$$

where

$$\alpha_0 = \text{overall intercept,}$$

$$D_{t'} = 1 \text{ if } t' = t \text{ and 0 otherwise.}$$

Equation (7.34) has the same structure as (7.32) where $\Theta_{t'} + \alpha_0$ is interpreted as an estimate for the logarithm of the denominator of (7.19) for period t'. Equation (7.34) involves T additional parameters α_0 and $\Theta_{t'}$, $t' = 2, \ldots, T$.[31,32]

Many books and articles on market share response models and their use have been published, documenting the successful formulation and implementation of this class of models. An example is a study by DeSarbo et al. (2002) who develop a generalization of this class of models to a latent structure framework incorporating within-segment random brand effects.

Competitive reactions to entry using a market share model have been investigated by Fok and Franses (2004) while Klapper and Herwartz (2000) investigate the predictive abilities of market share direction models.[33,34]

7.4 Examples of Normative/Prescriptive Models

7.4.1 Introduction and Illustrations

7.4.1.1 Basic Model

In this section we briefly discuss normative models. Their purpose is to determine a recommended course of action that should improve performance. In other words,

[30] See Nakanishi and Cooper (1982).

[31] This has consequences for the degrees of freedom and the estimated standard errors. See Foekens (1995, p. 169).

[32] We do not discuss the assumptions of the disturbances nor the estimation techniques required to estimate these relations.

[33] See also Brodie and De Kluyver (1984); Naert and Weverbergh (1981, 1985); Leeflang and Reuyl (1984) and Brodie et al. (2001).

[34] Other, more recent application of market share response models are Mukherjee and Kadiyali (2011) and Leeflang and Parreño Selva (2012).

one wants to determine which decision is best for an objective such as profit maximization. We discuss a marketing mix problem to illustrate a normative model for profit maximization.

The objective of this model is to evaluate the profitability of advertising, holding other variables constant. The model is based on demand equation (7.35), which numerically specifies the demand[35] for 24 bottles of Heineken beer (consumer price 6 Euro).

$$\begin{aligned}\hat{q}_t = &-32,733 + 12,423 \log Inc_t + 0.507\, q_{t-1} + 1,777 \log_{10} a_t \\ &- 2.2\, wi_t + 843 \log f_t \end{aligned} \tag{7.35}$$

where

q_t = sales per 1,000 potential consumers,

Inc_t = real private disposable income,

q_{t-1} = lagged sales,

a_t = real advertising expenditures per 1,000 potential consumers,

wi_t = weather index (rainfall), and

f_t = visit frequency to sales outlets by sales representatives,

t = year.

There is no price variable among the explanatory variables in the demand equation. In the original specification, price was included but it was dropped because the estimated effect was statistically insignificant. This is one of the dilemmas often faced in applied econometrics. We know that price affects demand in general. However, if price (or any other variable for that matter) is relatively stable over the period of observation, then it cannot have much explanatory power in the sample. If price is excluded, its effect will be taken up by the constant term, the implication being that, on the basis of historical data, nothing can be said about the impact of price on sales. Thus, if the objective is to determine the optimal marketing mix, the information contained in the historical data would be insufficient. Other means, such as experimentation or subjective judgment may provide useful insights. The lack of sufficient price variation is of no consequence if the model is to be used for a determination of optimal advertising spending, under the assumption that the actual level of advertising does not depend on price. The output is a recommended budget. In this sense, the model is normative. However, from a practical point of view, the word "normative" is perhaps too strong. What one is really interested in is to determine whether current advertising expenditures are too high, too low, or

[35] This model is a modification of a model developed by Lambin (1969).

about right. The model can produce a specific figure but it is useful to take it as a guideline rather than as something absolute. We provide the following reasons why we might not use the word "normative" in an absolute sense:

1. the demand equation is estimated, which implies that there is uncertainty about the true values of the response coefficients;
2. advertising is only one instrument in the firm's marketing mix;
3. the firm faces multiple objectives, while most models assume the existence of a single objective such as profit maximization;
4. the effectiveness of advertising depends on the quality of the copy, the selection of media, and so on. The regression coefficient gives at best an idea of an average effect per Euro advertising expenditures.

We now derive an optimal advertising budget for the profit-maximizing firm whose demand function is given by (7.35). We first show how to optimize advertising expenditures when lagged effects of advertising are *not* taken into account. After that an optimal solution is derived which assumes that the lagged sales variable represents advertising dynamics.

7.4.1.2 Determination of the Short-Term Advertising Budget

Since advertising spending is the object of the study, we assume that other variables (price, visit frequency) have been decided upon, and uncontrollable variables have been predicted. For example, national statistics can be used to predict future levels of disposable income. Substituting the values of these variables, Eq. (7.35) reduces to:

$$\hat{q}_t = -2,231 + 1,777 \log_{10} a_t. \tag{7.36}$$

The objective of the firm is to maximize profit per 1,000 potential consumers, π_t:

$$\hat{\pi}_t = (p_t - c_t)\hat{q}_t - a_t \tag{7.37}$$

where c_t is unit variable cost. In the remainder of this discussion, we assume that c_t is constant, i.e. $c_t = c$ for all t. In other words, variable cost (c) and marginal cost (MC) are equal. Fixed costs can be ignored since these do not affect the optimal level of advertising spending. Given that price is predetermined, we let p_t equal p, so that (7.37) reduces to:

$$\hat{\pi}_t = (p - c)\hat{q}_t - a_t. \tag{7.38}$$

To maximize profit, the following relationship needs to hold[36]:

$$\mu = 1/w \tag{7.39}$$

[36]Following the Dorfman and Steiner (1954) theorem derived in the Appendix to this chapter.

where

$$\mu = p\frac{\partial q}{\partial a} \quad = \text{marginal revenue product of advertising, and}$$

$$w = \frac{p - MC}{p} = \frac{p - c}{p} = \text{percent gross margin.}$$

Currently, average advertising expenditure for 1,000 potential consumers is $\bar{a} = 3{,}440$ Euro, with corresponding sales, $\bar{q} = 4{,}060$.

The price is 6 Euro (average price paid by the retailer to the manufacturer) and marginal cost (assumed constant over the relevant range) is 2.70 Euro. The percent of gross margin for the manufacturer is then:

$$w = (6 - 2.70)/6 \times 100\,\% = 55\,\%. \tag{7.40}$$

From (7.39) we find that at the optimum the marginal revenue product of advertising should satisfy:

$$\mu = 1/0.55 = 1.818. \tag{7.41}$$

Since $\mu = p\dfrac{\partial q}{\partial a}$, at optimality we should have:

$$\partial q/\partial a = (1.818)/6 = 0.303. \tag{7.42}$$

In general, $\partial q/\partial a$ can be written as follows[37]:

$$\frac{\partial q}{\partial a} = \left[\frac{\partial q}{\partial \log_{10} a}\right]\left[\frac{\partial \log_{10} a}{\partial \ln a}\right]\left[\frac{\partial \ln a}{\partial a}\right]$$

which is

$$\frac{\partial q}{\partial a} = \left[\frac{\partial q}{\partial \log_{10} a}\right]\left[0.4343 \times \frac{1}{a}\right]. \tag{7.43}$$

It follows from (7.43) that the optimal advertising spending level a^* should satisfy:

$$a^* = \frac{(\partial q/\partial \log_{10} a)(0.4343)}{\partial q/\partial a}. \tag{7.44}$$

From (7.35) we know that $\partial q/\partial \log_{10} a = 1{,}777$, and from (7.42) $\partial q/\partial a = 0.303$. Thus a^* equals:

$$a^* = (1{,}777)(0.4343)/0.303 = 2{,}545 \text{ Euro}.$$

[37] This rather complex expression results from the fact that in Eq. (7.35) logarithms to the base ten were used.

Compared to actual expenditures of 3,440 Euro, it appears that the firm is overspending on (short-term) advertising. It is, however, instructive to determine how sensitive profit is to changes in advertising spending. We first examine the profit when the firm continues the current advertising spending. Profit is predicted to be:

$$\hat{\pi} = (p - MC)\hat{q} - a = 3.3\hat{q} - a.$$

With $\bar{a} = 3{,}440$ Euro, we found $\bar{q} = \hat{q} = 4{,}060$, so that current profit is predicted to be:

$$\hat{\pi} = 3.3 \times 4{,}060 - 3{,}440 = 9{,}958 \text{ Euro}.$$

At optimality, $a^* = 2{,}545$ Euro, with corresponding sales:

$$\begin{aligned}\hat{q} &= -2{,}213 + 1{,}777 \log_{10} 2{,}545 \\ &= -2{,}213 + 1{,}777 \times 3.4057 = 3{,}839.\end{aligned}$$

Maximum profit is then:

$$\hat{\pi} = 3.3 \times 3{,}839 - 2{,}545 = 10{,}124 \text{ Euro}.$$

We see that, if only short-term effects are considered, actual advertising expenditures exceed the optimal level by about 35 %. However, current profit is only about 1.5 % below its maximum value. This suggests that profit is quite insensitive to changes in advertising expenditures.

7.4.1.3 Determination of the Long-Term Advertising Budget

The effect of advertising in period t also occurs in later periods. Let λ be the retention rate of advertising. An advertising investment of a_t in period t, yields q_t sales in t, λq_t in $t+1$, $\lambda^2 q_t$ in $t+2, \ldots$ However, given that an Euro of profit in the future is less valuable than the same profit today, we adjust for the time value of money when we evaluate an advertising investment. Let the discount rate be i. The present value of the long term (LT) profit stream $\pi(LT)$ generated by an advertising expenditure of a dollars is:

$$\pi(LT) = q(p-c)\left[1 + \frac{\lambda}{1+i} + \frac{\lambda^2}{(1+i)^2} + \frac{\lambda^3}{(1+i)^3} + \ldots\right] - a. \tag{7.45}$$

Since $0 \leq \lambda < 1$, we also have $0 \leq \dfrac{\lambda}{1+i} < 1$, and (7.45) reduces to:

$$\pi(LT) = \frac{q(p-c)}{1 - \lambda/(1+i)} - a.$$

At optimality we should have[38]:

$$\frac{\mu}{1-\lambda/(1+i)} = \frac{1}{w}. \tag{7.46}$$

Recall from Eq. (2.43) that the coefficient of lagged sales q_{t-1}, in Eq. (7.35) is an estimate of λ. Thus $\lambda = 0.507$, and if we further assume a discount rate of $i = 8\,\%$, the long-term optimal value of μ is[39]:

$$\mu = 1.818\left(1 - \frac{0.507}{1+0.08}\right) = 0.965, \text{ and therefore}$$

$$\partial q/\partial a = 0.965/6 = 0.161.$$

If the firm takes a long-term view, the optimal advertising budget is:

$$a_{LT} = (1{,}777)(0.4343)/0.161 = 4{,}801 \text{ Euro}.$$

Corresponding expected sales are:

$$\hat{q}_{LT} = -2{,}213 + 1{,}777 \log 4{,}801 = 4{,}329$$

and expected long-term profit is:

$$\hat{\pi}_{LT} = \frac{(4{,}329)(3.30)}{0.5306} - 4{,}801 = 22{,}125 \text{ Euro}.$$

If advertising expenditures remain at the *current* level, expected long-term profit is:

$$\hat{\pi}_{LT} = \frac{(4{,}060)(3.30)}{0.5306} - 3{,}440 = 21{,}811 \text{ Euro}.$$

Thus, by taking into account the lagged effects of advertising, we find that actual spending is an estimated 40 % *below* the optimal amount. Importantly, with positive lagged effects, the optimal advertising expenditure increases relative to the case when we ignore these effects. As before, if we increase the advertising budget to the optimal level, the expected profit increases by only 1.44 %.

We note that sensitivity analyses frequently demonstrate that large percentual changes in advertising expenditures result in only small percentual changes in profits over a wide range of expenditure levels. This phenomenon is known as the

[38] $\frac{\partial \pi(LT)}{\partial a} = \frac{(p-c)(\partial q/\partial a)}{1-\lambda/(1+i)} - 1 = 0$, or $\frac{p\partial q/\partial a}{1-\lambda/(1+i)} = \frac{p}{p-c} = \frac{p}{p-MC} = \frac{1}{w}$.

[39] The parameter λ has been estimated from annual data. From Sect. 2.8.2 we know that $\hat{\lambda}$ may be biased upward.

flat maximum principle (Hanssens et al. 2001, p. 26). In this respect we refer to Chintagunta (1993b) and note that profit is more sensitive to departures of price from its optimal level than it is to departures from optimal advertising expenditures.

We have to be careful in the application of optimization rules. In the discussion above we implicitly assumed that there are no profitable alternatives to advertising spending. Indeed, the condition $\mu = 1/w$ implies that at the optimum:

$$(p - MC)\frac{\partial q}{\partial a} = 1.$$

Stated in words, this means that the last Euro invested just pays for itself but nothing more. In fact the firm can do better, by investing that last dollar in some other venture where it can earn a return of *re* percent. The return on the best possible alternative investment should be considered as an opportunity cost. With an opportunity cost of *re* percent, the optimality condition becomes:

$$\mu = \frac{1 + re}{w}.$$

With $re = 0.20$, and $\mu = (1 + 0.20)/0.55 = 2.18$, at optimality $\partial q/\partial a$ should equal $\mu/p = 2.18/6 = 0.36364$. Applying (7.44) for example, we find an optimal short-term advertising budget of 2,122 Euro, instead of 2,545 Euro which we obtained when the opportunity cost was (implicitly) assumed to be zero.

7.4.2 Other Normative Models

We now face many models in which profits, customer life time value and firm value are used as performance measures and that models are developed to optimize these measures.[40] Examples are:

- models to compute optimal catalog mailing decisions: Gönül and Shi (1998); Gönül and Ter Hofstede (2006); Van Diepen et al. (2009);
- models to determine the effect of sales promotions on profit: Zhang et al. (2000); Srinivasan et al. (2004); Ailawadi et al. (2005);
- models to manage channel profits: Jeuland and Shugan (2008); Gensler et al. (2012);
- advertising budget models: Feinberg (2001); Buratto et al. (2006); Wang and Zhang (2008);
- (dynamic) pricing models: Zoltners (1981); Narasimhan (1988); Rao (1993); Kopalle et al. (1999);

[40] See Albers (2012) for a recent wake-up call to develop optimization models, see also the discussion in Sect. 10.5.

- models that are used to determine the optimal sales force: Lodish (1971); Lodish et al. (1988); Rangaswamy et al. (1990); Gopalakrishna and Chatterjee (1992);
- models that are constructed to specify salesforce compensation plans: Basu et al. (1985); Lal and Staelin (1986);
- models to determine sales territory design: Zoltners and Sinha (1983, 2005);
- models to optimize retail assortments (Rooderkerk et al. 2013), and
- very specific models to determine optimal fee-based compensation plans for search engine marketing (Nabout et al. 2012).

In the last decade normative models were developed that optimize and predict individual customer profitability. Examples can be found in Chap. 9, Blattberg et al. (2008) and in Rust et al. (2011).

7.4.3 Allocation Models

The marketing literature contains several allocation models, with the following characteristics. Resources are available in limited quantities; for example, an advertising manager has a budget, a sales person can work eight hours a day, potentially supplemented by a few of hours overtime. The purpose of such models is to allocate this quantity to subvariables (media, market segments, sales accounts …) so as to optimize an objective function (profit, sales, …). We consider a few examples in the areas of advertising, selling, and sales promotions and the allocation of shelf space.

Blattberg and Neslin (1990, p. 391) suggest that the promotion planning process consists of three levels of budgeting decisions: the total marketing budget, the allocation of that budget to promotions (versus advertising and other marketing mix elements), and the preparation of individual promotion budgets or "individual events". The allocation of the *total marketing budget* over promotions, advertising and other marketing mix elements can be accomplished by a model with about the same basic structure as the model used in the Dorfman–Steiner theorem (see the Appendix to this chapter).[41] A model which allocates the promotion budget over advertising and trade promotion expenditures was developed by Neslin et al. (1995). This model represents the manufacturer's attempt to maximize profits by advertising directly to consumers and offering periodic discounts to the retailer in the hope that the retailer will in turn "pass through" a promotion to the consumer. The model considers the allocation in a manner that appears to fit the last two levels of budgeting decisions defined by Blattberg and Neslin.

The best-known allocation models are the *media allocation* models. Some of these only consider the allocation of a given advertising budget to a number of alternative media vehicles. Examples are Lee and Burkart (1960); Lee (1962), and

[41] It is also conceivable, of course, to have both budget determination and allocation in one single model. An example is the "integrated model for sales force structuring" developed by Rangaswamy et al. (1990).

Ellis (1966). Others have modeled the timing of the insertion in the various media. Examples are: Lee (1963); Little and Lodish (1969); Srinivasan (1976); Mahajan and Muller (1986); Feinberg (1992); Bronnenberg (1998); Naik et al. (1998). Reddy et al. (1998) developed the SPOT (Scheduling Programs Optimally to Television)-model. This model is used for optimal prime-time TV program scheduling. Because the advertising revenues of TV-networks are linked directly to the size of the audience delivered to the advertiser, this type of scheduling model is also relevant for decision makers in marketing. The issue of advertising schedules, specifically whether advertising should be steady (constant) or turned on and off (pulsed), has also received attention (Freimer and Horsky 2012; Hahn and Hyun 1991). Objective functions in media allocation models vary from the maximization of reach to maximization of a discounted profit stream over a finite time horizon. Danaher and Dagger (2013) develop a budget allocation model that considers ten (10) media. Their model is based on individual loyalty program data.

An advertising budget can be allocated to subvariables other than media vehicles as well. The subvariables could be market segments such as, for example, different geographic regions. Alternatively an advertising budget can be allocated to different products (Doyle and Saunders 1990). Pieters and Wedel (2004) develop methods to allocate the surface of advertising pages over three key ad elements (brand, pictorial, and text).

The allocation of *sales effort* has also been the subject of numerous studies. Zoltners and Sinha (1983, 2005), Skiera and Albers (1998) examined the spatial allocation of a sales force. Brown et al. (1956) studied the optimal frequency of visiting actual and potential buyers. Lodish (1971) developed procedures to optimize a salesman's allocation of time spent on different accounts. Montgomery et al. (1971) present a procedure to help a salesperson in determining how much time to spend on various products to be sold. Dong et al. (2009) allocated detailing visits across individual physicians.

The allocation of the *(sales) promotion budget* to individual events deserves more attention. Relevant to this question is the empirical result that the frequency and magnitude of price discounts have significant effects on the (own)price elasticities.[42] Higher and more frequent discounts lead to less negative price elasticities. Thus the timing and the determination of the size of the discount are important determinants of the managers' profit optimization problem.[43] However, the allocation of the total amount to discounts in specific time periods and the magnitude of each discount remain important optimization questions.

We note that the allocation of *shelf space* requires the development of idiosyncratic models. See, for example, Lim et al. (2004); Rooderkerk et al. (2013). We introduced idiosyncraticy briefly in Sect. 2.2.6 and we will discuss it in more detail in Sect. 10.2.2.2.

Fischer et al. (2011) developed a dynamic marketing budget allocation model for multiproducts, multicountry firms. Their model allocates budgets across coun-

[42] See, for example, Raju (1992); Foekens et al. (1999).

[43] See also Tellis and Zufryden (1995).

tries, products and marketing activities for Bayer, one of the world's largest pharmaceutical and chemical firms. They demonstrate that implementation of this model leads to an improvement of profits of more than 50 %.

Appendix: The Dorfman–Steiner Theorem

Let $q = q(p,a,\tilde{x})$, be demand (q) as a function of price (p), advertising (a), and quality ($\tilde{x}$).

$c = (q,\tilde{x})$, be variable cost per unit, and

$FC =$ fixed cost.

Profit π is:

$$\pi = p\,q(p,a,\tilde{x}) - c(q,\tilde{x})q(p,a,\tilde{x}) - a - FC. \tag{7.47}$$

If the objective is to maximize profit, at optimality we should have[44]:

$$\frac{\partial \pi}{\partial p} = q + p\frac{\partial q}{\partial p} - c\frac{\partial q}{\partial p} - q\frac{\partial c}{\partial q}\frac{\partial q}{\partial p} = 0 \tag{7.48}$$

$$\frac{\partial \pi}{\partial a} = p\frac{\partial q}{\partial a} - c\frac{\partial q}{\partial a} - q\frac{\partial c}{\partial q}\frac{\partial q}{\partial a} - 1 = 0 \tag{7.49}$$

$$\frac{\partial \pi}{\partial \tilde{x}} = p\frac{\partial q}{\partial \tilde{x}} - c\frac{\partial q}{\partial \tilde{x}} - q\frac{\partial c}{\partial q}\frac{\partial q}{\partial \tilde{x}} - q\frac{\partial c}{\partial \tilde{x}} = 0. \tag{7.50}$$

Dividing (7.48) by $(\partial q/\partial p)$ we obtain:

$$\frac{q}{\partial q/\partial p} + p - c - q\frac{\partial c}{\partial q} = 0. \tag{7.51}$$

Total variable production cost equals $c \cdot q$. Marginal cost (MC) is then:

$$MC = \frac{\partial (cq)}{\partial q} = c + q\frac{\partial c}{\partial q}. \tag{7.52}$$

Using (7.52), we can write (7.51) as:

$$\frac{-q}{\partial q/\partial p} = p - MC.$$

[44] We assume that second-order conditions are satisfied.

Dividing both sides by p, and letting:

$$w = \frac{p - MC}{p} = \text{percentage of gross margin}$$

we obtain:

$$-\eta_p = 1/w \tag{7.53}$$

where

$$\eta_p = \frac{\partial q}{\partial p}\frac{p}{q} = \text{price elasticity.}$$

Dividing (7.49) by $(\partial q/\partial a)$,

$$p - c - q\frac{\partial c}{\partial q} - \frac{1}{\partial q/\partial a} = 0$$

or

$$p - MC = \frac{1}{\partial q/\partial a}.$$

After dividing both sides by p, we find:

$$\mu = 1/w \tag{7.54}$$

where

$$\mu = p\frac{\partial q}{\partial a} = \text{marginal revenue of product advertising.}$$

Finally, we divide (7.50) by $\partial q/\partial \tilde{x}$:

$$p - c - q\frac{\partial c}{\partial q} - q\frac{\partial c/\partial \tilde{x}}{\partial q/\partial \tilde{x}} = 0$$

or

$$\frac{p - MC}{p} = \frac{q\,\partial c/\partial \tilde{x}}{p\,\partial q/\partial \tilde{x}}$$

or

$$\eta_{\tilde{x}}\frac{p}{c} = 1/w \tag{7.55}$$

where

$$\eta_{\tilde{x}} = \frac{(\partial q/\partial \tilde{x})/q}{(\partial c/\partial \tilde{x})/c}.$$

At optimality (7.53), (7.54), and (7.55) should hold simultaneously, or:

$$-\eta_p = \mu = \eta_{\tilde{x}} \frac{p}{c} = \frac{1}{w}. \tag{7.56}$$

This result is generally known as the Dorfman and Steiner (1954) theorem. This theorem has been modified and extended in many directions. Examples are the models of Lambin (1970); Lambin et al. (1975); Leeflang and Reuyl (1985b); Plat and Leeflang (1988); Mantrala et al. (2007).

References

Ailawadi, K.L., Kopalle, P.K., Neslin, S.A.: Predicting competitive response to a major policy change: Combining game-theoretic and empirical analyses. Mark. Sci. **24**, 12–24 (2005)

Albers, S.: Optimizable, implementable and generalizable aggregate response modeling for decision support. Int. J. Res. Mark. **29**, 111–122 (2012)

Andrews, R.L., Currim, I.S.: Multi-stage purchase decision models: Accommodating response heterogeneity, common demand shocks, and endogeneity using disaggregate data. Int. J. Res. Mark. **26**, 197–206 (2009)

Andrews, R.L., Currim, I.S., Leeflang, P.S.H.: A comparison of sales response predictions from demand models applied to store-level vs. panel data. J. Bus. Econ. Stat. **29**, 319–326 (2011)

Andrews, R.L., Currim, I.S., Leeflang, P.S.H., Lim, J.: Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? Int. J. Res. Mark. **25**, 22–33 (2008)

Basu, A.K., Lal, R., Srinivasan, V., Staelin, R.: Salesforce compensation plans: An agency theoretic perspective. Mark. Sci. **4**, 267–291 (1985)

Ben-Akiva, M., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. MIT, Cambridge (1985)

Blattberg, R.C., Kim, B.D., Neslin, S.A.: Database Marketing: Analyzing and Managing Customers. Springer Science, New York (2008)

Blattberg, R.C., Neslin, S.A.: Sales Promotions: Concepts, Methods and Strategies. Prentice-Hall, Englewood Cliffs (1990)

Blattberg, R.C., Wisniewski, K.J.: Price-induced patterns of competition. Mark. Sci. **8**, 291–309 (1989)

Brodie, R.J., Danaher, P., Kumar, V., Leeflang, P.S.H.: Econometric models for forecasting market share. In: J.S. Armstrong (ed.) Principles of Forecasting. Kluwer Academic, Hingham/Dordrecht (2001)

Brodie, R.J., De Kluyver, C.A.: Attraction versus linear and multiplicative market share models: An empirical evaluation. J. Mark. Res. **21**, 194–201 (1984)

Bronnenberg, B.J.: Advertising frequency decisions in a discrete Markov process under a budget constraint. J. Mark. Res. **35**, 399–406 (1998)

Bronnenberg, B.J., Wathieu, L.: Asymmetric promotion effects and brand positioning. Mark. Sci. **15**, 379–394 (1996)

Brown, A.A., Hulswit, F.L., Ketelle, J.D.: A study of sales operations. Oper. Res. **4**, 296–308 (1956)

Bultez, A.V.: Competitive strategy for interrelated markets. Unpublished Ph.D. thesis, Louvain University, Louvain (1975)
Buratto, A., Grosset, L., Viscolani, B.: Advertising a new product in a segmented market. Eur. J. Oper. Res. **42**, 1343–1347 (2006)
Campo, K., Gijsbrechts, E., Goossens, T., Verhetsel, A.: The impact of location factors on the attractiveness and optimal space shares of product categories. Int. J. Res. Mark. **17**, 255–279 (2000)
Capella, M., Webster, C., Kinard, B.R.: A review of cigarette advertising. Int. J. Res. Mark. **28**, 269–279 (2011)
Carpenter, G.S., Cooper, L.G., Hanssens, D.M., Midgley, D.F.: Modelling asymmetric competition. Mark. Sci. **7**, 393–412 (1988)
Chintagunta, P.K.: Heterogeneity in nested logit models: An estimation approach and empirical results. Int. J. Res. Mark. **9**, 161–175 (1992)
Chintagunta, P.K.: Investigating purchase incidence, brand choice and purchase quantity decisions of households. Mark. Sci. **12**, 184–208 (1993a)
Chintagunta, P.K.: Investigating the sensitivity of equilibrium profits to advertising dynamics and competitive effects. Manag. Sci. **39**, 1146–1162 (1993b)
Chintagunta, P.K., Desiraju, R.: Strategic pricing and detailing behavior in international markets. Mark. Sci. **24**, 67–80 (2005)
Chintagunta, P.K., Jain, D.C., Vilcassim, N.J.: Investigating heterogeneity in brand preferences in logit models for panel data. J. Mark. Res. **28**, 417–428 (1991)
Christen, M., Gupta, S., Porter, J.C., Staelin, R., Wittink, D.R.: Using market-level data to understand promotion effects in a nonlinear model. J. Mark. Res. **34**, 322–334 (1997)
Cooper, L.G., Klapper, D., Inoue, A.: Competitive component analysis: A new approach to calibrating asymmetric market-share models. J. Mark. Res. **33**, 224–238 (1996)
Cooper, L.G., Nakanishi, M.: Market-Share Analysis: Evaluating Competitive Marketing Effectiveness. Kluwer Academic, Boston (1988)
Danaher, P.J., Dagger, T.S.: Comparing the relative effectiveness of advertising channels: A case study of a multimedia blitz campaign. J. Mark. Res. **50**, 517–534 (2013)
De Laat, E., Windmeijer, F., Douven, R.: How Does Pharmaceutical Marketing Influence Doctors' Prescription Behaviour? CPB, Bureau for Economic Policy Analysis, The Hague (2002)
Debreu, G.: Review of R.D. Luce's individual choice behavior: A theoretical analysis. Am. Econ. Rev. **50**, 186–188 (1960)
Dekimpe, M.G., François, P., Gopalakrishna, S., Lilien, G.L., Van den Bulte, C.: Generalizing about trade show effectiveness: A cross-national comparison. J. Mark. **61**(3), 55–64 (1997)
DeSarbo, W., Degeratu, A., Ahearne, M., Saxton, M.: Disaggregate market share response models. Int. J. Res. Mark. **19**, 253–266 (2002)
Ding, M., Eliashberg, J., Stremersch, S. (eds.): Innovation and Marketing in the Pharmaceutical Industry: Emerging Practices, Research and Policies. Springer, New York (2014)
Dong, X., Manchanda, P., Chintagunta, P.K.: Quantifying the benefits of individual-level targeting in the presence of firm strategic behavior. J. Mark. Res. **46**, 207–221 (2009)
Dorfman, R., Steiner, P.O.: Optimal advertising and optimal quality. Am. Econ. Rev. **44**, 826–836 (1954)
Doyle, P., Saunders, J.: Multiproduct advertising budgeting. Mark. Sci. **9**, 97–113 (1990)
Duesenberry, J.S.: Income, Saving and the Theory of Consumer Behavior. Harvard University Press, Cambridge (1949)
Duffy, M.: Econometric studies of advertising, advertising restrictions and cigarette demand: A ssurvey. Int. J. Advert. **15**, 1–23 (1996)
Ellis, D.M.: Building up a sequence of optimum media schedules. Oper. Res. Q. **17**, 413–424 (1966)
Feinberg, F.M.: Pulsing policies for aggregate advertising models. Mark. Sci. **11**, 221–234 (1992)
Feinberg, F.M.: On continuous-time optimal advertising under S-shaped response. Manag. Sci. **47**, 1476–1487 (2001)

Fischer, M., Albers, S.: Patient- or physician-oriented marketing: What drives primary demand for prescription drugs? J. Mark. Res. **47**, 103–121 (2010)
Fischer, M., Albers, S., Wagner, N., Frie, M.: Dynamic marketing budget allocation across countries, products, and marketing activities. Mark. Sci. **30**, 568–585 (2011)
Foekens, E.W.: Scanner data based marketing modelling: Empirical applications. Unpublished Ph.D. thesis, University of Groningen, Groningen (1995)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation. Int. J. Forecast. **10**, 245–261 (1994)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: Hierarchical versus other market share models for markets with many items. Int. J. Res. Mark. **14**, 359–378 (1997)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: Varying parameter models to accommodate dynamic promotion effects. J. Econom. **89**, 249–268 (1999)
Fok, D., Franses, P.H.: Analyzing the effects of a brand introduction on competitive structure using a market share attraction model. Int. J. Res. Mark. **21**, 159–177 (2004)
Franses, P.H.: Primary demand for beer in the Netherlands: An application of ARMAX model specification. J. Mark. Res. **28**, 240–245 (1991)
Freimer, M., Horsky, D.: Periodic advertising pulsing in a competitive market. Mark. Sci. **31**, 637–648 (2012)
Gensler, S., Leeflang, P.S.H., Skiera, B.: Impact of online channel use on customer revenues and costs to serve: Considering product portfolios and self-selection. Int. J. Res. Mark. **29**, 192–201 (2012)
Gönül, F.F., Shi, M.Z.: Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. Manag. Sci. **44**, 1249–1262 (1998)
Gönül, F.F., Srinivasan, K.: Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. Mark. Sci. **12**, 213–229 (1993)
Gönül, F.F., Ter Hofstede, F.: How to compute optimal catalog mailing decisions. Mark. Sci. **25**, 65–74 (2006)
Gopalakrishna, S., Chatterjee, R.: A communications response model for a mature industrial product: Applications and implications. J. Mark. Res. **29**, 189–200 (1992)
Guadagni, P.M., Little, J.D.C.: A logit model of brand choice calibrated on scanner data. Mark. Sci. **2**, 203–238 (1983)
Gupta, S., Chintagunta, P.K.: On using demographic variables to determine segment membership in logit mixture models. J. Mark. Res. **31**, 128–136 (1994)
Gupta, S., Chintagunta, P.K., Kaul, A., Wittink, D.R.: Do household scanner data provide representative inferences from brand choices: A comparison with store data. J. Mark. Res. **33**, 383–398 (1996)
Hahn, M., Hyun, J.S.: Advertising cost interactions and the optimality of pulsing. Manag. Sci. **37**, 157–169 (1991)
Hanssens, D.M., Parsons, L.J., Schultz, R.L.: Market Response Models: Econometric and Time Series Analysis. Kluwer Academic, Boston (2001)
Jeuland, A.P., Shugan, S.M.: Managing channel profits. Mark. Sci. **27**, 52–69 (2008)
Kamakura, W.A., Russell, G.J.: A probabilistic choice model for market segmentation and elasticity structure. J. Mark. Res. **26**, 379–390 (1989)
Kapteyn, A., van de Geer, S., van de Stadt, H., Wansbeek, T.J.: Interdependent preferences: An econometric analysis. J. Appl. Econom. **12**, 665–686 (1997)
Kapteyn, A., Wansbeek, T.J., Buyze, J.: The dynamics of preference formation. J. Econ. Behav. Org. **1**, 123–157 (1980)
Klapper, D., Herwartz, H.: Forecasting market share using predicted values of competitive behavior: Further empirical results. Int. J. Forecast. **16**, 399–421 (2000)
Klein, L.R., Lansing, J.B.: Decisions to purchase consumer durable goods. J. Mark. **20**, 109–132 (1955)
Kopalle, P.K., Mela, C.F., Marsh, L.: The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications. Mark. Sci. **18**, 317–332 (1999)

Kremer, S.T.M., Bijmolt, T.H.A., Leeflang, P.S.H., Wieringa, J.E.: Generalizations on the effectiveness of pharmaceutical promotional expenditures. Int. J. Res. Mark. **25**, 234–246 (2008)
Lal, R., Staelin, R.: Salesforce compensation plans in environments with asymmetric information. Mark. Sci. **5**, 179–198 (1986)
Lambin, J.J.: Measuring the profitability of advertising: An empirical study. J. Ind. Econ. **17**, 86–103 (1969)
Lambin, J.J.: Modèles et Programmes de Marketing. Presses Universitaires de France, Paris (1970)
Lambin, J.J., Naert, P.A., Bultez, A.V.: Optimal marketing behavior in oligopoly. Eur. Econ. Rev. **6**, 105–128 (1975)
Lancaster, K.M.: Brand advertising competition and industry demand. J. Advert. **13**(4), 19–24 (1984)
Lee, A.M.: Decision rules for media scheduling: Static campaigns. Oper. Res. Q. **13**, 229–241 (1962)
Lee, A.M.: Decision rules for media scheduling: Dynamic campaigns. Oper. Res. Q. **14**, 113–122 (1963)
Lee, A.M., Burkart, A.J.: Some optimization problems in advertising media. Oper. Res. Q. **11**, 113–122 (1960)
Leeflang, P.S.H., Parreño Selva, J.: Cross-category demand effects of price promotions. J. Acad. Mark. Sci. **40**, 572–586 (2012)
Leeflang, P.S.H., Parreño Selva, J., van Dijk, A., Wittink, D.R.: Decomposing the sales promotion bump accounting for cross-category effects. Int. J. Res. Mark. **25**, 201–214 (2008)
Leeflang, P.S.H., Reuyl, J.C.: On the predictive power of market share attraction models. J. Mark. Res. **11**, 211–215 (1984)
Leeflang, P.S.H., Reuyl, J.C.: Advertising and industry sales: An empirical study of the West German cigarette market. J. Mark. **49**(4), 92–98 (1985a)
Leeflang, P.S.H., Reuyl, J.C.: Competitive analysis using market response functions. In: R. Lusch et al. (ed.) Educators' Proceedings. American Marketing Association, Chicago (1985b)
Leeflang, P.S.H., Reuyl, J.C.: Effects of tobacco advertising on tobacco consumption. Int. Bus. Rev. **4**, 39–54 (1995)
Leeflang, P.S.H., Wieringa, J.E.: Modeling the effects of pharmaceutical marketing. Mark. Lett. **21**, 121–133 (2010)
Leone, R.P., Schultz, R.L.: A study of marketing generalizations. J. Mark. **44**(1), 10–18 (1980)
Lim, A., Rodrigues, B., Zhang, X.: Metaheuristics with local search techniques for retail shelf-space optimization. Manag. Sci. **50**, 117–131 (2004)
Little, J.D.C., Lodish, L.M.: A media planning calculus. Oper. Res. **17**, 1–35 (1969)
Lodish, L.M.: CALLPLAN: An interactive salesman's call planning system. Manag. Sci. **18**, B25–B40 (1971)
Lodish, L.M., E. Curtis, M.N., Simpson, M.K.: Sales force sizing and deployment using a decision calculus model at Syntex Laboratories. Interfaces **18**, 5–20 (1988)
Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
Luik, J.C., Waterson (eds.), M.J.: Advertising & Markets. NTC Publications, Oxfordshire (1996)
Ma, Y., Ailawadi, K.L., Gauri, D.K., Grewal, D.: An empirical investigation of the impact of gasonline prices on grocery shopping behavior. J. Mark. **75**(2), 18–35 (2011)
Mahajan, V., Muller, E.: Advertising pulsing policies for generating awareness for new products. Mark. Sci. **5**, 89–106 (1986)
Manchanda, P., Wittink, D.R., Ching, A., Cleanthous, P., Ding, M., Dong, X., Leeflang, P.S.H., Misra, S., Mizik, N., Narayanan, S., Steenburgh, T., Wieringa, J.E., Wosínska, M., Xie, Y.: Understanding firm, physician and consumer choice behavior in the pharmaceutical industry. Mark. Lett. **16**, 293–308 (2005)
Mantrala, M.K., Naik, P.A., Sridhar, S., Thorson, E.: Uphill or downhill? Locating the firm on a profit function. J. Mark. **71**(2), 26–44 (2007)
McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: P. Zarembarka (ed.) Frontiers in Econometrics. Academic, New York (1974)

McFadden, D., Reid, F.: Aggregate travel demand forecasting from disaggregated behavioral models. Transp. Res. Rec. **534**, 24–37 (1975)
Montgomery, D.B., Silk, A.J., Zaragoza, C.E.: A multiple-product sales force allocation model. Manag. Sci. **18**(Part II), P3–P24 (1971)
Mukherjee, A., Kadiyali, V.: Modeling multichannel home video demand in the U.S. motion picture industry. J. Mark. Res. **48**, 985–995 (2011)
Nabout, N.A., Skiera, B., Stepanchuk, T., Gerstmeier, E.C.: An analysis of the profitability of fee-based compensation plans for search engine marketing. Int. J. Res. Mark. **29**(1), 68–80 (2012)
Naert, P.A., Weverbergh, M.: On the prediction power of market share attraction models. J. Mark. Res. **18**, 146–153 (1981)
Naert, P.A., Weverbergh, M.: Market share specification, estimation and validation: Toward reconciling seemingly divergent views. J. Mark. Res. **22**, 453–461 (1985)
Naik, P.A., Mantrala, M.K., Sawyer, A.G.: Planning media schedules in the presence of dynamic advertising quality. Mark. Sci. **17**, 214–235 (1998)
Nakanishi, M., Cooper, L.G.: Parameter estimation for a multiplicative competitive interaction model - least squares approach. J. Mark. Res. **11**, 303–311 (1974)
Nakanishi, M., Cooper, L.G.: Simplified estimation procedures for MCI models. Mark. Sci. **1**, 314–322 (1982)
Narasimhan, C.: Competitive promotional strategies. J. Bus. **61**, 427–449 (1988)
Narayanan, S., Desiraju, R., Chintagunta, P.K.: Return on investment implications for pharmaceutical promotional expenditures: The role of marketing-mix interactions. J. Mark. **68**(4), 90–105 (2004)
Nelson, J.P.: Cigarette advertising regulations: A meta-analysis. Int. Rev. Law Econ. **26**, 195–226 (2006)
Neslin, S.A., Powell, S.G., Schneider Stone, L.G.: The effects of retailer and consumer response on optimal manufacturer advertising and trade promotion strategies. Manag. Sci. **41**, 749–766 (1995)
Nijs, V.R., Dekimpe, M.G., Steenkamp, J.B.E.M., Hanssens, D.M.: The category-demand effects of price promotions. Mark. Sci. **20**, 1–22 (2001)
Pieters, F.G.M., Wedel, M.: Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. J. Mark. **68**(2), 36–50 (2004)
Plat, F.W., Leeflang, P.S.H.: Decomposing sales elasticities on segmented markets. Int. J. Res. Mark. **5**, 303–315 (1988)
Raju, J.S.: The effect of price promotions on variability in product category sales. Mark. Sci. **11**, 207–220 (1992)
Rangaswamy, A., Sinha, P., Zoltners, A.A.: An integrated model-based approach to sales force structuring. Mark. Sci. **9**, 279–298 (1990)
Rao, V.R.: Pricing models in marketing. In: J. Eliashberg, G.L. Lilien (eds.) Handbooks in Operations Research and Management Science Marketing, vol. 5. North-Holland, Amsterdam (1993)
Reddy, S.K., Aronson, J.E., Stam, A.: SPOT: Scheduling programs optimally for television. Manag. Sci. **44**, 83–102 (1998)
Rizzo, J.A.: Advertising and competition in the ethical pharmaceutical industry: The case of antihypertensive drugs. J. Law Econ. **42**, 89–116 (1999)
Rooderkerk, R.P., Van Heerde, H.J., Bijmolt, T.H.A.: Optimizing retail assortments. Mark. Sci. **32**, 699–715 (2013)
Rossi, P.E., Allenby, G.M.: A Bayesian approach to estimating household parameters. J. Mark. Res. **30**, 171–182 (1994)
Rust, R.T., Kumar, V., Venkatesan, R.: Will the frog change into a prince? Predicting future customer profitability. Int. J. Res. Mark. **28**, 281–294 (2011)
Sethuraman, R., Srinivasan, V., Kim, D.: Asymmetric and neighborhood cross-price effects: Some emprirical generalizations. Mark. Sci. **18**, 23–42 (1999)
Skiera, B., Albers, S.: COSTA: Contribution optimizing sales territory alignment. Mark. Sci. **17**, 196–213 (1998)

Srinivasan, S., Pauwels, K., Hanssens, D.M., Dekimpe, M.G.: Do promotions benefit manufacturers, retailers, or both? Manag. Sci. **50**, 617–629 (2004)
Srinivasan, V.: Decomposition of a multi-period media scheduling model in terms of single period equivalents. Manag. Sci. **23**, 349–360 (1976)
Tellis, G.J., Zufryden, F.S.: Tackling the retailer decision maze: Which brands to discount, how much, when and why? Mark. Sci. **14**, 271–299 (1995)
Theil, H.: A multinomial extension of the linear logit model. Int. Econ. Rev. **10**, 251–259 (1969)
Van Diepen, M., Donkers, B., Franses, P.H.: Dynamic and competitive effects of direct mailings: A charitable giving application. J. Mark. Res. **46**, 120–133 (2009)
Van Heerde, H.J., Gupta, S., Wittink, D.R.: Is 75% of the sales promotion bump due to brand switching? No, only 33% is. J. Mark. Res. **40**, 481–491 (2003)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: The estimation of pre- and postpromotion dips with store-level scanner data. J. Mark. Res. **37**, 383–395 (2000)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Semiparametric analysis to estimate the deal effect curve. J. Mark. Res. **38**, 197–215 (2001)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: How promotions work: SCAN*PRO-based evolutionary model building. Schmalenbach Bus. Rev. **54**, 198–220 (2002)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Decomposing the sales promotion bump with store data. Mark. Sci. **23**, 317–334 (2004)
Vanden Abeele, P., Gijsbrechts, E., Vanhuele, M.: Specification and empirical evaluation of a cluster-asymmetry market share model. Int. J. Res. Mark. **7**, 223–247 (1990)
Venkataraman, S., Stremersch, S.: The debate on influencing doctors' decisions: Are drug characteristics the missing link? Manag. Sci. **53**, 1688–1701 (2007)
Wang, F., Zhang, X.P.: Reasons for market evolution and budgeting implications. J. Mark. **72**(5), 15–30 (2008)
Wedel, M., Kamakura, W.A.: Market Segmentation: Conceptual and Methodological Foundations. Kluwer Academic, Boston (1998)
Wieringa, J.E., Leeflang, P.S.H.: Modelling the effects of promotion expenditures on sales of pharmaceuticals. Appl. Econ. **45**(24), 3389–3399 (2013)
Wieringa, J.E., Osinga, E.C., Ruiz Conde, E., Leeflang, P.S.H., Stern, P.: Modeling the effects of promotional efforts on aggregate pharmaceutical demand: What we know and challenges for the future. In: M. Ding, J. Eliashberg, S. Stremersch (eds.) Innovation and Marketing in the Pharmaceutical Industry: Emerging Practices, Research and Policies. Springer, New York (2014)
Windmeijer, F., de Laat, E., Douven, R., Mot, E.: Pharmaceutical promotion and GP prescription behaviour. Health Econ. **15**, 5–18 (2005)
Wittink, D.R., Addona, M.J., Hawkes, W.J., Porter, J.C.: SCAN*PRO: The estimation, validation and use of promotional effects based on scanner data. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in Honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)
Yang, S., Allenby, G.M.: A model for observation, structural, and household heterogeneity in panel data. Mark. Lett. **11**, 137–150 (2000)
Zhang, Z.J., Krishna, A., Dhar, S.K.: The optimal choice of promotional vehicles: Front-loaded or rear-loaded incentives? Manag. Sci. **46**, 348–62 (2000)
Zoltners, A.A.: Normative marketing models. In: R.L. Schultz, A.A. Zoltners (eds.) Marketing Decision Models. North-Holland, New York (1981)
Zoltners, A.A., Sinha, P.: Sales territory alignment: A review and model. Manag. Sci. **29**, 1237–1256 (1983)
Zoltners, A.A., Sinha, P.: Sales territory design: Thirty years of modeling and implementation. Mark. Sci. **24**, 313–331 (2005)

Chapter 8
Individual Demand Models

8.1 Introduction

According to Wyner (2013) some of the biggest potential growth areas in marketing research include:

- more data: measurement of more consumer behaviors;
- more control: predictability of consumers' response to marketing initiatives;
- more depth: understanding how consumers' minds work, especially the non-rational side.

As we discussed in Sect. 3.5.6, Big Data obtained through web search, digital media, e-commerce, mobile and social media have become important for understanding consumers' behavior. Studying and modeling individual behavior has become more and more the focus in marketing research. Individual demand constitutes an important part of individual behavior, but we are now also able to study word-of-mouth (WOM-behavior), online-browsing, co-creation, etc. (see Chap. 9).

In this chapter we concentrate on individual *demand* models. Consumers choose to buy or not to buy. If they decide to buy they choose a certain product category and within that category a certain brand. Next they decide to buy 1, 2, 3, … units of the chosen brand. Finally consumers have to time their purchase. Hence, we have four decisions: "whether to buy", "what to buy", "how much to buy" and "when to buy". We start to discuss choice models that are developed for the first two decisions in Sect. 8.2. We discuss models that explain how much people buy in Sect. 8.3. These models are called purchase quantity models. The timing of the purchase ("when to buy") is the focus of the purchase timing/duration models. These models are discussed in Sect. 8.4. Section 8.5 discusses models that are used to describe and predict multiple decisions at the individual level. Here we spend attention to the so-called Tobit models.

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_8

8.2 Choice Models

8.2.1 Introduction

Choice modeling has grown to be a very substantial area in marketing research over the past few decades.[1,2] We distinguish binary choices from multinomial choices. Many marketing problems are binary. Examples are: what is the probability that a consumer i:

- wants to spend money (spending versus saving money)?
- wants to buy a product in a certain product category (and not in a competing product category, i.e. booking a vacation instead of buying clothes)?[3]
- buys a certain brand in that category and not a competing brand?
- terminates her contract with a certain company?
- responds to a direct mail campaign?
- will not pay her bills?
- will adopt a new product?, etc.

In all these cases we have two outcomes: 0 or 1, or Yes or No, with probabilities $P(Y_i = 0)$, $P(Y_i = 1)$. Hence, the dependent variable takes two values and is called binomial. Because there are no other outcomes than these two we have:

$$P(Y_i = 1) + P(Y_i = 0) = 1. \tag{8.1}$$

Let us consider a firm that targets its customers with a direct mail campaign. The question is whether the probability that a consumer i responds $P(Y_i = 1) = \pi_i$ or not responds $P(Y_i = 0) = 1 - \pi_i$ to the direct mail campaign depends on x_{1i}, the amount that customer i spent earlier on products of that firm. To this end, we specify the following model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i, \quad i = 1, \ldots, N \tag{8.2}$$

where

Y_i = the dependent variable that can take the values 0 or 1,

β_0 = an unknown intercept,

β_1 = an unknown slope parameter,

x_{1i} = the amount that customer i spent earlier with the firm (continuous variable),

[1] See, for example, the Special Issues on Choice Models of *Marketings Letters*, vol 8(3), 1997; vol 10(3),1999; and Chandukala et al. (2007).

[2] The authors like to thank Hans Risselada who provided important information for this chapter.

[3] Du and Kamakura (2008).

ε_i = the (unobserved) value of the disturbance term,

N = the number of consumers in a sample.

In our discussions of the General Linear Model we have so far assumed that the disturbance term is normally distributed (see Sect. 4.2.3). However, if we assume that the ε_i are normally distributed in Eq. (8.2) Y_i is also, by definition, normally distributed. This, however, is a heroic assumption given that Y_i can only take on two values: 0 or 1. As a consequence we have to make other assumptions about the disturbance terms which will lead to other models than the General Linear Model. Figure 8.1 illustrates what goes wrong if we make the assumption that the ε_i are normally distributed. Figure 8.1 shows the fit of an OLS regression of Y_i on an intercept and the amount that customer i spent earlier on products of that firm. The graph in Fig. 8.1 clearly demonstrates that the assumption of normally distributed disturbance terms is unlikely to be useful. Similar arguments hold when we consider the situation that the dependent variable has more than two choice options; in such cases the dependent is said to be multinomial.

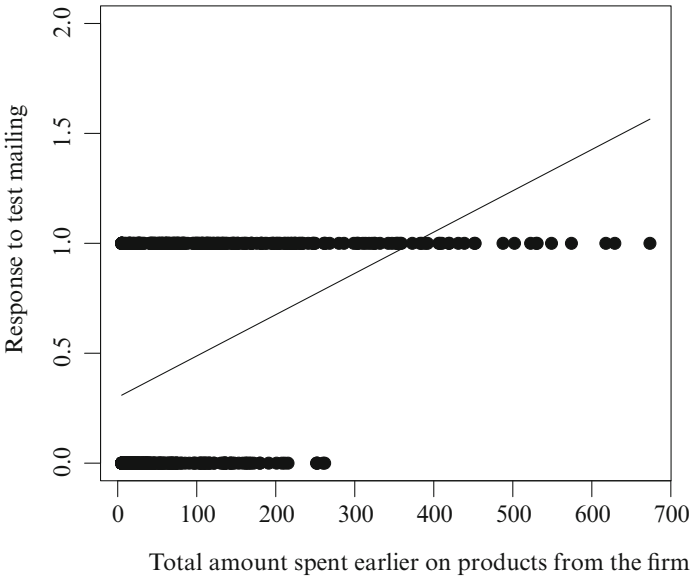

Fig. 8.1 OLS regression of Y_i on amount spent with the firm and an intercept

In what follows we discuss the following (brand) choice models:

- 8.2.2. Binary choice models;
- 8.2.3. Multinomial choice models;
- 8.2.4. Markov models.

8.2.2 Binary Choice Models Specification

8.2.2.1 Basic Model

Consider the linear model[4]

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, N \tag{8.3}$$

where

$$Y_i = \begin{cases} 1 \text{ if consumer } i \text{ buys brand } A, \\ 0 \text{ if } i \text{ buys brand } B, \end{cases}$$

x_i = the price difference between A and B for consumer i,

ε_i = the (unobserved) value of a random disturbance term.

We define $P(Y_i = 1) = \pi_i$ as the *probability* that brand A is chosen and by definition $P(Y_i = 0) = 1 - \pi_i$. These probabilities depend on the intercept, the slope parameter and x_i. Hence, instead of considering the precise value of Y_i one may focus on the *probability* that, for example, $Y_i = 1$ given β_0, β_1, x_i and ε_i:

$$\pi_i = P(Y_i = 1 \,|\, \beta_0, \beta_1, x_i, \varepsilon_i). \tag{8.4}$$

Hence, the probability that brand A is chosen is conditional on the β parameters, x_i and ε_i. In binomial (and multinomial) models these probabilities are not observed; we only observe the binary (or multinomial) values of the final choices of the individuals. The idea is that high probabilities (π_i) "predict" that $Y_i = 1$. In choice modeling these probabilities are estimated (e.g. by Maximum Likelihood) and it is determined how these probabilities are affected by the independent variable(s).

Franses and Paap (2001, p. 52) discuss two interpretations for the fact that in binomial models the focus is on modeling *probabilities* instead of on *observed values*. The first interpretation is that the probability that $Y_i = 1$ (choice of brand A) depends on the value of a so-called latent variable y_i^*. A *latent* or *unobservable* variable may be distinguished from *observable variables* or *indicators*. An explanatory observable indicators x_i may be related to y_i^* in, for example, the following way:

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{8.5}$$

where we leave the distribution of ε_i unspecified. In (8.5) the latent variable y_i^* can be interpreted as the difference between *unobserved* preferences between the brands A and B for each individual i. The relation between y_i^* and Y_i , the *observed dependent* variable can be specified as:

[4]The text of this subsection is based, a.o., on Franses and Paap (2001, Chapter 4) and Wooldridge (2012, Chapter 17).

$$Y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \leq 0. \end{cases} \tag{8.6}$$

A second interpretation towards modeling probabilities is based on microeconomic theory. The choice of either brand A or brand B is based on utilities U_{A_i} and U_{B_i} which are assigned to brands A and B respectively by consumer i:

$$\begin{aligned} U_{A_i} &= \beta_{0A} + \beta_{1A} x_i + \varepsilon_{A_i} \\ U_{B_i} &= \beta_{0B} + \beta_{1B} x_i + \varepsilon_{B_i}. \end{aligned} \tag{8.7}$$

One may define that individual i buys A if the utility of A exceeds that of B:

$$\begin{aligned} &P(Y_i = 1 \mid \beta_{0A}, \beta_{0B}, \beta_{1A}, \beta_{1B}, x_i, \varepsilon_{A_i}, \varepsilon_{B_i}) \\ &\quad = P(Y_i = 1 \mid Z_i) \\ &\quad = P(U_{A_i} > U_{B_i} \mid Z_i) \\ &\quad = P(\beta_{0A} - \beta_{0B} + (\beta_{1A} - \beta_{1B}) x_i > \varepsilon_{B_i} - \varepsilon_{A_i} \mid Z_i) \\ &\quad = P(\varepsilon_i \leq \beta_0 + \beta_i x_i \mid Z_i) \end{aligned} \tag{8.8}$$

where $Z_i = \left\{\beta_{0A}, \beta_{0B}, \beta_{1A}, \beta_{1B}, x_i, \varepsilon_{A_i}, \varepsilon_{B_i}\right\}$ and

$$\begin{aligned} \varepsilon_i &= \varepsilon_{B_i} - \varepsilon_{A_i}, \\ \beta_0 &= \beta_{0A} - \beta_{0B}, \text{ and} \\ \beta_1 &= \beta_{1A} - \beta_{1B}. \end{aligned} \tag{8.9}$$

Expressions (8.9) demonstrate that one cannot identify the individual parameters in (8.7). The specification of the distribution of ε_i has to make sure that the probabilities are between zero and one. The last line of (8.8) states that the probability of observing $Y_i = 1$ given X_i is equal to the cumulative distribution function of ε_i evaluated at $\beta_0 + \beta_1 x_i$. We can specify this as:

$$\pi_i = P(Y_i = 1 | Z_i) = F(\beta_0 + \beta_1 x_i) \tag{8.10}$$

where $F(\beta_0 + \beta_1 x_i)$ denotes the *cumulative* distribution function of ε_i. The two most considered distributions are either the logistic distribution function or the normal distribution function.

In the first case, we have:

$$F(\beta_0 + \beta_1 x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.11}$$

If this distribution is chosen the *logit model* will result, and logistic regression can be used to estimate the relation. Equation (8.11) can be generalized if there are more explanatory variables:

$$F(X_i'\beta) = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}, \tag{8.12}$$

where X_i the matrix of observations of the independent variables for consumer i, and β is a vector of parameters.

The second commonly considered cumulative distribution that satisfies the condition that the π_i's are between zero and one is the standard normal cumulative distribution. Now $F(X_i'\beta)$ is expressed as an integral:

$$F(X_i'\beta) = \int_{-\infty}^{X_i'\beta} \phi(u)\mathrm{d}u \tag{8.13}$$

where $\phi(\cdot)$ is the standard normal density function. This assumption about the cumulative distribution leads to the *probit model*.

Both cumulative distribution functions are quite similar. The specification of a logit model is often preferred because of mathematical convenience. Also the interpretation of the parameter estimates is (somewhat) easier for the logit model as compared to the probit model.[5]

Equation (8.3) can be modified in several ways to account for unobserved *heterogeneity*. The individual parameters could be consumer-specific (see, e.g. Wedel et al. (1999), who use latent classes of consumers), and/or consumer-specific variables such as e.g. income, age or gender could be included in the model (Franses and Paap 2001, p. 72). We return to the issue of heterogeneity in binomial and multinomial models at the end of Sect. 8.2.3.

8.2.2.2 Estimation

Logit and probit models are estimated by Maximum Likelihood estimation methods. The models are then written in terms of their joint density distribution $P(y|X,\beta)$ where y are the observed variables, X is a matrix of observations of the independent variables and β a vector of model parameters. The likelihood functions for observation i can be written as:

$$L_i(\beta) = P(y_i|X,\beta) \tag{8.14}$$

[5]Economists, however, tend to favor the normality assumption for ε_i which is why the probit model is more popular than logit in econometrics (Wooldridge 2012, p. 562).

and the logarithmic likelihood function (log-likelihood function) as:

$$LL_i(\beta) = l_i(\beta) = \ln(L_i(\beta)). \tag{8.15}$$

The log-likelihood function for a sample size N is obtained by summing $l_i(\beta)$ over all observations $l(\beta) = \sum_{i=1}^{N} l_i(\beta)$ [compare Eq. (6.41)]. The maximum likelihood estimator of β, denoted by $\hat{\beta}$, maximizes this log-likelihood.

The maximization of (8.15) has to be done using a numerical optimization algorithm (Franses and Paap 2001, p. 59). Because of the nonlinear nature of the maximization problem we are not able to write formulas for the logit or probit maximum likelihood estimates (Wooldridge 2012, p. 564).

8.2.2.3 Numerical Examples

We consider Eq. (8.3) and assume that the parameter estimates $(\hat{\beta}_0, \hat{\beta}_1)$ are $(4, -2)$. The price differences between brand A and brand B is 1.44 (Euro). Then: $\hat{U}_{A_i} = 4 + (-2 \times 1.44) = 1.12$. Using (8.11) we find that $F(1.12) = \exp(1.12)/(1 + \exp(1.12)) = 0.75$. This is the *probability* that individual i buys brand A given the parameter estimates and $x_i = 1.44$.

We can simply extend our analysis including several independent variables. Let us, for example, assume that the utility that someone buys at Hennes and Mauritz depends on income (per month, x_1) and gender ($x_2, 0 =$ male, $1 =$ female). The utility of a consumer i of buying at Hennes and Mauritz can be specified as:

$$U_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i. \tag{8.16}$$

We suppose that $\hat{\beta}_0 = 0.5$, $\hat{\beta}_1 = -0.01$ and $\hat{\beta}_2 = 1.1$. From these estimates it is clear that when customer i has a high income the utility is lower than when i has a low income. The utility to buy at Hennes and Mauritz is higher for women than for men. The probability that a female student with a monthly income of 400 Euro will buy at Hennes and Mauritz is:

$$\begin{aligned} \pi_i &= \exp(0.5 - 0.01 \times 400 + 1.1 \times 1)/(1 + \exp(0.5 - 0.01 \times 400 + 1.1 \times 1)) \\ &= \exp(-2.4)/(1 + \exp(-2.4)) \\ &= 0.08. \end{aligned}$$

The probability that a male student with the same income (400) will buy at Hennes and Mauritz is 0.03 (3 %).

8.2.2.4 Validation

Logit and probit models are validated using several criteria. Statistical programmes provide estimates of the standard error of the parameters. The expressions for the standard error are quite complicated. Once we have the standard errors, we can construct (asymptotic) t-tests and confidence intervals just as with the General Linear Model. Also p-values are often used to determine the significance of the individual parameters. An alternative test to determine the significance of individual parameters is the Wald test (see Sect. 6.4.4).

An important criterion to interpret the parameters of the logit model is the *odds ratio*: the probability of $Y_i = 1$ divided by the probability of $Y_i = 0$:

$$\text{Odds ratio} = \frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} = \frac{\pi_i}{1 - \pi_i}. \tag{8.17}$$

An odds ratio of 3 means that the probability of $Y_i = 1$ is three times larger than the probability of $Y_i = 0$. For the logit model with one variable, it is easy to see using (8.11) that:

$$\text{Odds ratio} = \exp(\beta_0 + \beta_1 x_i). \tag{8.18}$$

It is common practice to consider the log odds ratio:

$$\log \text{odds ratio} = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i. \tag{8.19}$$

When $\beta_1 = 0$, the log odds ratio equals β_0.[6] If β_0 is also equal to zero, then the log odds ratio is equal to zero which means that the probabilities of observing $Y_i = 0$ and $Y_i = 1$ are the same and equal 0.5.

In the output of most statistical packages that are used to estimate a logit model we find expressions for e^{β_k}. This indicates the change in odds ratio if the k-th variable changes one unit. So, for example, if the variable x_k represents gender (0 = male, 1 = female), a parameter estimate of $\beta_k = 0.693$ means that a one-unit increase in the explanatory variable x_k results in a change in the odds ratio of $\exp(\beta_k) = 2$. Hence, the odds ratio for females is twice the odds ratio for males.

The *probit* model is based on the cumulative distribution function of a standard normal distribution (with mean zero and variance 1). The coefficients of the probit model are the effects of the explanatory variables on a cumulative standard normal function of the probabilities that the response variable $Y_i = 1$. The coefficients are the changes in the z-scores of that distribution [see (8.13)]. If the coefficients are zero, then there is no relation, a z-score larger (smaller) than zero indicates a positive (negative) relation.

[6] We closely follow Franses and Paap (2001, p. 57).

To assess the fit of a choice model, we cannot directly compute an R^2 as in the linear model (see Sect. 4.3.1). To facilitate interpretation, several so-called pseudo R-squared statistics have been developed that have a similar interpretation as the R^2 of a linear model. These pseudo R-squares are all based on comparing the log-likelihood of a model with only an intercept (LL_0) with the log-likelihood of a model with K explanatory variables (LL_K). Popular pseudo R squares are:

$$\text{McFadden } R^2 = 1 - (LL_k / LL_0) \tag{8.20}$$

$$\text{Cox and Snell } R^2 = 1 - \left(\frac{-2LL_0}{-2LL_K}\right)^{2/N} \tag{8.21}$$

$$\text{Nagelkerke } R^2 = \left(1 - \left(\frac{-2LL_0}{-2LL_K}\right)^{2/N}\right) \Big/ \left(1 - (-2LL_0)^{2/N}\right) \tag{8.22}$$

where the last two measures correct for the number of observations (N). If the covariates x_k have no explanatory power, the parameters will be estimated as zero. This results in $LL_K = LL_0$ and the pseudo R-squares are zero, just as the usual R-squared is zero in a linear regression.

The choice between models can be performed using the Hit Rate, the likelihood ratio test, and information criteria (see Sect. 5.6). For determining the Hit Rate, the observations are classified into two groups. The first group consists of the observations for which the model predicts that $Y_i = 1$, based on the estimated π_i. The second group contains the observations for which the model predicts that $Y_i = 0$. The Hit Rate measures the percentage of correctly classified observations. A Hit Rate of 80 % indicates that 80 % of the values $Y_i = 1$ or $Y_i = 0$ are correctly predicted by the model. Typically, observation i is classified in the $Y_i = 1$ group if $\hat{\pi}_i \geq 0.5$, and to the $Y_1 = 0$ group otherwise. Other classification rules than using a threshold value of 0.5 may be more appropriate. An example of such a threshold is to use the fraction of successes in the sample, i.e. the percentage of customer for which $Y_i = 1$. If this fraction is 8 %, one may use a threshold of 0.08 (Wooldridge 2012, p. 566).

The above diagnostic checks implicitly consider the adequacy of the functional form. There are, however, no clear guidelines as to how one should choose between a logit and probit model (Franses and Paap 2001, p. 63).

8.2.2.5 Empirical Example[7]

Wieringa and Verhoef (2007) studied the switching behavior of customers who preferred another electricity supplier. They estimate a logit model [see Eq. (8.12)] with the probability that a consumer i prefers a different electricity supplier as the dependent variable and the following explanatory variables:

[7] We closely follow Wieringa and Verhoef (2007).

$PC1_i$ = relationship quality (high to low),

$PC2_i$ = switching cost (high to low),

$PC3_i$ = attractiveness of switching (low to high),

$Usage_i$ = a measure of the demand for electricity by consumer i,

$NumContr_i$ = the number of contracts (e.g. electricity, gas maintenance, cable) respondent i holds with the local company.

It is assumed in this study that the error term follows the standardized logit distribution with mean zero. The estimation results of this model are shown in Table 8.1. The variables in Table 8.1 significantly affect customer switching ($\chi^2(6) = 2.015, p < 0.01$) and the model fit is reasonable with a McFadden R^2 of 0.255.

Table 8.1 Logistic regression results of Wieringa and Verhoef (2007)

Variable	Parameter estimate	Standard error
Constant	−0.566*	0.098
PC1: Relationship quality (high to low)	1.391*	0.043
PC2: Switching costs (high to low)	0.652*	0.035
PC3: Attractiveness of switching (low to high)	0.343*	0.034
Usage	−0.008*	0.002
NumContr	−0.291*	0.036
(McFadden) $R^2 = 0.255$		
% correctly classified (Hit Rate) = 82.6 %		

*Significant at the 0.01 level (two-tailed)
Source: Wieringa and Verhoef (2007, p. 180)

The percentage of correct classifications by the model is 82.6 %, which compares favorably to the benchmark accuracy of 76.5 and 69.7 % that is achieved by using the so-called maximum chance criterion or proportional chance criterion, respectively, as defined by Morrison (1969).

8.2.3 Multinomial Choice Models

8.2.3.1 Structure

In a multinomial choice model an individual chooses between n alternatives. The choices are either unordered or ordered, but always discrete. An ordered multinomial variable differs from an unordered multinomial variable by the fact that the values that the variable can attain are *ranked*. Such a variable is, for example, obtained when a Likert scale is used with items "very bad", "bad", …, "good", "very good", "excellent". There are not many empirical marketing models that

are specified as ordered multinomial models. Examples are Verhoef et al. (2001); Sridhar and Srinivasan (2012). We concentrate on the unordered multinomial models and refer for the ordered multinomial models to Franses and Paap (2001, Chapter 6).

Unordered choice models considers the case an individual chooses between n alternatives. The binomial model is a special case for which $n = 2$. This means that many topics that are discussed in Sect. 8.2.2 also apply for the multinomial case. Let us first return to the case where $n = 2$. If the random variable Y takes one of only two values (representing two brands, or a purchase and non-purchase situation) $Y = 1$, with probability π, and $Y = 0$, with probability $1 - \pi$ we obtain the so-called Bernouilli model:

$$P(Y = y) = \pi^y (1 - \pi)^{1-y}. \tag{8.23}$$

The expectation of the Bernouilli distribution is $\mathrm{E}(Y) = \pi$, and its variance is $\mathrm{Var}(Y) = \pi(1 - \pi)$. This is a distribution for an individual consumer. Let T be the number of purchase occasions for each consumer in the population. All have the same probability of purchasing a brand and $1 - \pi$ is the non-purchase probability. Hence, we assume homogeneity of all consumers. Then $\pi \times T = \eta$ is the expected number of purchases by a consumer. The probability that y purchases takes place is represented by a Binomial distribution:

$$P(Y = y) = \binom{T}{y} \pi^y (1 - \pi)^{T-y}, \quad y = 0, 1, 2, \ldots, T. \tag{8.24}$$

In a multi-brand market, brands are denoted by $j = 1, \ldots, n$. Next, we assume that each consumer $i = 1, \ldots, I$ may purchase a brand j with probability π_j, $j = 1, \ldots, n$. We do not consider the situation of a non-purchase for convenience, and assume T purchase occasions. We observe y_j, the *number of times brand j is purchased*, for each j. Let $T = \sum_{j=1}^{n} y_j$ denote the total number of purchases observed. Now the multinomial model applies:

$$P(Y_1, \ldots, Y_n = y_1, \ldots, y_n) = \left(\frac{T!}{y_1!, \ldots, y_n!} \right) \prod_{j=1}^{n} \pi_j^{y_j}. \tag{8.25}$$

The multinomial model arises from T Bernouilli trials (8.23). The expectation of the multinomial random variable is: $\mathrm{E}(Y_j) = \pi_j T$, its variance is: $\mathrm{Var}(Y_j) = \pi_j (1 - \pi_j) T$, and the covariance is $\mathrm{Cov}(Y_j, Y_k) = -\pi_j \pi_k T$. The binomial distribution results if $n = 2$. Thus, in Bernouilli, binomial and multinomial models the market is described by a set of (stationary) probabilities, describing the purchase probabilities (or alternatively the market shares) of the brands.

Similarly to the assumption for binary choice models (see Sect. 8.2.2), we assume for a multinomial choice model that a consumer will choose the alternative that gives him maximal utility (see, for example, Luce 1959). The utilities are assumed to have

a fixed component and a random component. Hence for each individual i, we can specify U_i, an $(n \times 1)$ vector that contains the (unobserved) random utilities that individual i derives from the n alternatives. The jth element of this vector is denoted by U_{ij} and is the utility that individual i associates with the jth alternative. U_{ij} can be represented as:

$$U_{ij} = x'_{ij}\beta + \varepsilon_{ij} \tag{8.26}$$

where x'_{ij} is a $(1 \times L)$ row-vector of marketing variables related to the j-th choice for the ith individual, β is a $(L \times 1)$ vector of unknown parameters, and ε_{ij} is the error term or the random part of utility. Each individual i chooses the brand with the maximal utility. Thus the observed choice variable Y_{ij} is defined as:

$$Y_{ij} = \begin{cases} 1 \text{ if } U_{ij} > U_{ir} & \text{for all } r \neq j, \quad r = 1, \ldots, n, \\ 0 \text{ otherwise.} \end{cases} \tag{8.27}$$

Define $\pi_{ij} = P(Y_{ij} = 1)$. If the ε_{ij} are independently and identically distributed with Weibull density functions, the multinomial logit model applies. The choice probability of an individual i for alternative j, given a multinomial logit model, can be expressed as:

$$\pi_{ij} = \frac{\exp\left(x'_{ij}\beta\right)}{\sum_{r=1}^{n} \exp\left(x'_{ir}\beta\right)}. \tag{8.28}$$

This leads to a multinomial distribution of the choice probabilities as in (8.25). There are several opportunities to specify the nominator of (8.28). Assuming a single variable w_{ij} we may specify:

$$\exp\left(\beta_0 + \beta_1 w_{ij}\right) \qquad \text{(homogeneous parameters)} \tag{8.29}$$

$$\exp\left(\beta_{0j} + \beta_1 w_{ij}\right) \qquad \text{(fixed effect)} \tag{8.30}$$

$$\exp\left(\beta_{0j} + \beta_{1j} w_{ij}\right) \qquad \text{(heterogeneous parameters)} \tag{8.31}$$

or accounting for multiple explanatory variables:

$$\exp\left(\beta_{0j} + \beta_1 w_{ij} + \beta_{2j} v_i\right) \tag{8.32}$$

$$\exp\left(\beta_{0j} + \beta_1 w_{ij} + \beta_{2j} v_i + \beta_3 z_j\right). \tag{8.33}$$

In (8.32) the choice probabilities depend an explanatory variable w_{ij} which has a common impact β_1 on the probabilities, and on an individual-specific variable

v_i whose effect on the probabilities differs per brand. In (8.32) w_{ij} could be the number of direct mailings that individual i received from brand j, whereas v_i may be gender. This model is called the *Conditional Logit model* (McFadden 1974). Expression (8.33) includes an additional explanatory variable z_j that is different across categories/brands, but the same for each individual. The z_j often represent marketing instruments: price, advertising expenditures, etc. of the category/brand j. This expression is called the *General Logit specification*.

Because the probabilities π_i sum to 1, a base category has to be assigned. We assume that the "last brand" n is picked as the base category. It is quite common that this base refers to a "no buy"-situation or an "outside (category) option" (Chib et al. 2004). All utilities are defined relative to n and the utility of alternative n is zero. We are now able to specify the probabilities of (8.28) in more detail where we opt to specification (8.32):

$$\pi_{ij} = \frac{\exp\left(\beta_{0j} + \beta_1 w_{ij} + \beta_{2j} v_i\right)}{1 + \sum_{r=1}^{n-1} \exp\left(\beta_{0r} + \beta_1 w_{ir} + \beta_{2r} v_i\right)} \quad \text{for } j = 1, \ldots, n-1 \tag{8.34}$$

and

$$\pi_{in} = \frac{1}{1 + \sum_{r=1}^{n-1} \exp\left(\beta_{0r} + \beta_1 w_{ir} + \beta_{2r} v_i\right)}. \tag{8.35}$$

From (8.34) and (8.35) it is clear that $\sum_{r=1}^{n} \pi_{ir} = 1$. Note that when $n = 2$ (8.34) and (8.35) reduce to the binomial model.

As a small numerical example of the foregoing, consider the case where $\exp(U_1) = 4, \exp(U_2) = 3, \exp(U_3) = 1$. If $n = 3$ is the base, then $\pi_{i1} = 0.5, \pi_{i2} = 0.375$ and $\pi_{i3} = 0.125$.

Estimation and validation of the multinomial model is similar to that of the binomial model. Odds ratios, relative to the base case, may be used to interpret the estimated parameters. The multinomial logit model is perhaps the most frequently used choice model in marketing (as well as in other disciplines).[8] Important application areas include the analysis of household-level scanner data (Guadagni and Little 1983) and conjoint choice experiments (Louvière and Woodworth 1983).

[8]For marketing applications see, for example, Punj and Staelin (1978); Guadagni and Little (1983); Louvière and Hensher (1983); Carpenter and Lehmann (1985); Kamakura and Russell (1989); Chintagunta et al. (1991); Erdem (1996); Ainslie and Rossi (1998); Seetharaman (2004); Gilbride and Allenby (2006); Chandukala et al. (2007). An overview of issues arising in logit model applications in marketing is provided by Malhotra (1984); McFadden (1986); Franses and Paap (2001, Chapter 5); Hruschka et al. (2004).

The multinomial logit model, however, suffers from the *Independence of Irrelevant Alternatives* (IIA) property, which states that the odds of choosing one alternative over another is constant regardless of whichever other alternatives are present. Formally if C and $D \subset C$ denote two sets of alternatives, then the IIA-assumption for two alternatives j and r (Luce 1959) is:

$$\frac{P(j \mid C)}{P(r \mid C)} = \frac{P(j \mid D)}{P(r \mid D)}. \tag{8.36}$$

In the multinomial logit model this property arises directly from the independence assumption of the error terms. It may not be realistic in many marketing applications, especially if some of the alternatives are close substitutes. On the positive side, if the IIA-assumption holds, future demand can simply be predicted with the closed-form expression (8.28) and the estimated values of the parameters (Urban and Hauser 1980, Chapter 11). However, if similarities across alternatives are incorrectly assumed away, the estimated effects of marketing variables are incorrect.

McFadden (1986) shows how one can deal with problems that arise from the IIA-assumption, including statistical tests of IIA. If IIA does not hold, other models can be used, often at the cost of computational complexity. A number of these models have been discussed in Sect. 7.3.3. The structure of these models, that are specified at the aggregate demand level, can also be used to model asymmetric individual choice behavior. Another example is Rooderkerk et al. (2011) who proposed a choice model that explicitly accounts for the fact that consumer choice behavior is affected by the composition of the choice set. In particular, their model accounts for three context effects, which are a function of the composition of the choice set. First, the compromise effect refers to the phenomenon that the middle option of a choice set tends to get a relatively large choice share. Second, the attraction effect causes that a relatively large choice share for items that are superior to otherwise similar items. Finally, the similarity effect refers to the phenomenon that items are affected more by similar than by dissimilar items. By decomposing the utility of an item into a context-free part and a context-dependent part, the model by Rooderkerk et al. (2011) accommodates these context effects and the model is no longer affected by the IIA assumption.

Other models that are developed to alleviate the IIA assumption are the Nested MultiNomial Logit (NMNL) model and the MultiNomial Probit (MNP) model.

In the NMNL model (McFadden 1981), consumer choice may follow a hierarchy of differentiating characteristics. We consider an example from Foekens et al. (1997) in which consumers consider products in the detergent market according to two main characteristics: the brand name and the package size. Following structure 1 in Fig. 8.2 a consumer first chooses a brand with probabilities π_r, $r = 1, \ldots, 7$, and then conditional upon the choice of the brand, a certain package size j with conditional probability $\pi_{j|r}$.[9] We assume that there are five package sizes: $j = 1, \ldots, 5$.

[9] See also Roberts and Lilien (1993).

If a consumer first chooses a package size and then a brand we have the conditional probabilities $\pi_{r|j}$: see structure 2 in Fig. 8.2.

We continue with structure 1 and we assume that a consumer's utility can be separated into two components, one attributable to brand (r) and the other to the package size (j)[10]:

$$U_{jr} = U_r + U_{j|r} = V_r + V_{j|r} + \varepsilon_r + \varepsilon_{j|r} \tag{8.37}$$

where V_r is the deterministic part of the utility associated with the highest level of the hierarchy (the brand), and $V_{j|r}$ is the systematic part of the utility that is associated with the second level in the hierarchy (the package size, given the brand). Analogously to the derivation of (8.28), we specify the choice probability at the *lowest* level of the hierarchy (package size j given brand choice r) as:

$$\pi_{j|r} = \frac{\exp\left(V_{j|r}\right)}{\sum_{k=1}^{n} \exp\left(V_{k|r}\right)}. \tag{8.38}$$

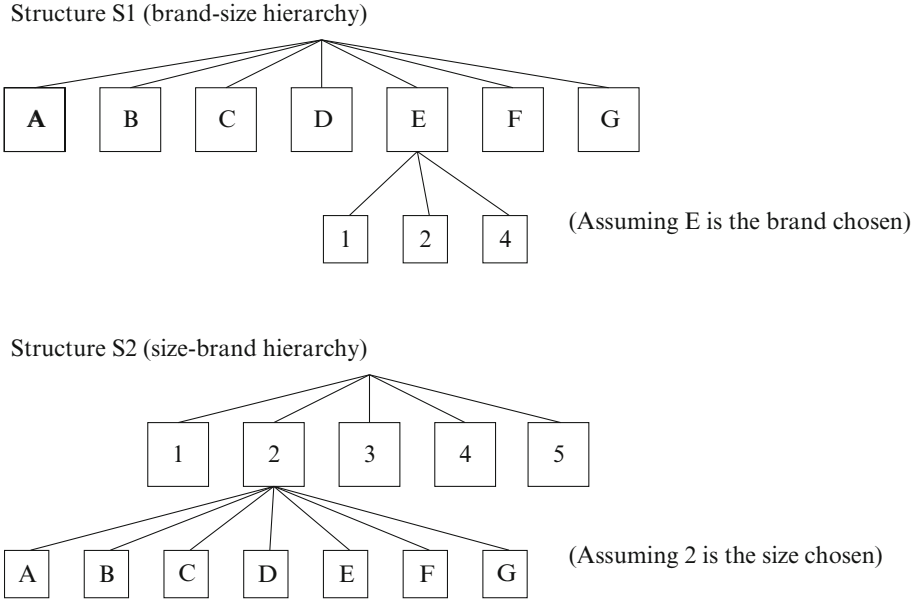

Fig. 8.2 Two alternative hierarchical structures with respect to the choice of brands (A, …, G) and package sizes (1, …, 5)

[10]We assume that the hierarchical structure is consumer specific. See also Vanden Abeele and Gijsbrechts (1991) and Siddarth et al. (1995).

The choice at the highest level of the hierarchy [brand choice, (r)] can be derived from utility maximization:

$$\pi_r = P\left(\max_j U_{jr} > \max_j U_{jr'} \text{ for all } r', \quad r' \neq r\right) \tag{8.39}$$

$$= P\left(U_r + \max_j U_{j|r} > U_{r'} + \max_j U_{j|r'} \text{ for all } r' = 1,\ldots,n', \quad r' \neq r\right). \tag{8.40}$$

An expression for $\max_j U_{j|r}$ can be obtained from the properties of the double exponential distribution, since the maximum of a set of double exponentially distributed variables (with unit variance) also follows a double exponential distribution with expectation:

$$\mathrm{E}\left(\max_j U_{j|r}\right) = \ln\left(\sum_{j'=1}^{n} e^{V_{j'|r}}\right). \tag{8.41}$$

Expression (8.41) is called the "inclusive value" of the utility for the brand name which is included in the utility for the form as shown in (8.42). From (8.39) and (8.41) the choice probabilities at the highest level of the hierarchy can be shown to be:

$$\pi_r = \frac{\exp\left(V_r + \ln\left(\sum_{j'=1}^{n} e^{V_{j'|r}}\right)\right)}{\sum_{r'=1}^{n'} \exp\left(V_{r'} + \ln\left(\sum_{j'=1}^{n} e^{V_{j'|r'}}\right)\right)}. \tag{8.42}$$

The unconditional choice probability of any alternative, jr, is simply $\pi_{jr} = \pi_{j|r} \times \pi_r$. In this model the brand utilities at the lowest (brand name) level of the hierarchy affect the utilities at the highest (form) level through the inclusive values. For a comprehensive treatment of the NMNL model see Ben-Akiva and Lerman (1985, Chapter 10), who include extensions to higher-order nestings and implications for the elasticity structure. Other models which are based on a hierarchical structure such as the Fully Extended Multinomial Logit model and the Extended Multinomial Logit model also accommodate for asymmetry. For a specification of these nested models at the aggregate level, see Foekens et al. (1997).

In the MultiNomial Probit (MNP) model the n disturbances of the random utilities for individual i in (8.26) are assumed to follow an n-dimensional multivariate normal distribution:

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in} \end{pmatrix} \sim N_n(0, \Omega). \tag{8.43}$$

This distribution allows the utilities of alternatives to be correlated, so that the IIA-assumption can be relaxed. However, a closed-form expression for the probability that individual i chooses alternative j cannot be derived, because it involves a multidimensional integral. Probabilities can be obtained by numerical methods if the number of choice alternatives is limited to 3 or 4. Early applications include those by Currim (1982) and Kamakura and Srivastava (1984, 1986). A comprehensive treatment of the MNP model is given by Daganzo (1979) and Franses and Paap (2001, pp. 86–88).

8.2.3.2 Heterogeneity

The choice models discussed so far in this section assume that the effect parameters are the same across consumers, the so-called homogeneity assumption. *Heterogeneity* can be taken into account in the binomial model by allowing π to follow a beta distribution across the population of consumers. The beta distribution is a flexible distribution that can take a variety of shapes:

$$f(\pi \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \tag{8.44}$$

with α_1 and α_2 as parameters and $\Gamma(\cdot)$ are gamma distributions. From (8.24) and (8.44) the number of purchases (y) can be shown to follow a Beta-Binomial (BB) distribution:

$$f(y \mid \alpha_1, \alpha_2) = \binom{T}{y} \frac{\Gamma(\alpha_1 + \alpha_2)\Gamma(y + \alpha_1)\Gamma(T - y + \alpha_2)}{\Gamma(T + \alpha_1 + \alpha_2)\Gamma(\alpha_1)\Gamma(\alpha_2)} \tag{8.45}$$

with mean: $E(Y) = \alpha_1 T/(\alpha_1 + \alpha_2)$.

In a similar way, heterogeneity can be accounted for in multinomial models describing multi-brand markets. Goodhardt et al. (1984) proposed a Dirichlet distribution for the choice probabilities. The Dirichlet can be seen as a multivariate extension of the beta distribution. The Dirichlet distribution is defined as:

$$f(\pi_1,\ldots,\pi_n \mid \alpha_1,\ldots,\alpha_n) = \frac{\Gamma\left(\sum_{j=1}^{n}\alpha_j\right)\prod_{j=1}^{n}\pi_j^{\alpha_j-1}}{\prod_{j=1}^{n}\Gamma(\alpha_j)}. \tag{8.46}$$

By compounding the multinomial and the Dirichlet the DirichletMultinomial (DM) is obtained:

$$f(y_1,\ldots,y_n \mid \alpha_1,\ldots,\alpha_n) = \frac{T!\Gamma\left(\sum_{j=1}^{n}\alpha_j\right)\prod_{j=1}^{n}\Gamma(y_j+\alpha_j)}{\Gamma\left(\sum_{j=1}^{n}\alpha_j+T\right)\prod_{j=1}^{n}\Gamma(\alpha_j)y_j!}. \tag{8.47}$$

The mean of Y_j equals $E(Y_j) = \alpha_j T/\left(\sum_{r=1}^{n}\alpha_r\right)$. The Beta-Binomial in Eq. (8.45) arises as a special case for a two-brand market ($n = 2$).

Strong empirical support for the DM is provided by Ehrenberg and coauthors in their work over the past 30 years (see for example Ehrenberg 1988; Uncles et al. 1995). They discuss many applications where the DM is a useful description of brand purchase behavior. Regularities based on the DM have been found in product markets for food and drink products, personal care products, gasoline, aviation fuel, and motor cars, OTC medicines, as well as in TV program and channel choice and shopping behavior. In many markets, the DM model does a good job of explaining observed regularities in purchase behavior, such as the percentage of consumers buying in a certain period, the number of purchases per buyer, the repeat purchases, the percentage of loyals, etc.

8.2.4 Markov Models

8.2.4.1 Markov Models

Markov brand choice models are based on Markov chains, in which one considers probabilities such as[11]:

$$P(Y_t = j \mid Y_{t-1} = i, Y_{t-2} = k,\ldots) \tag{8.48}$$

i.e. the probability that brand j is purchased at time t, given that brand i was purchased at $t-1$, brand k at $t-2$,.... These probabilities are called transition

[11] We only consider so-called discrete-time Markov chains.

probabilities. A simple specification of the transition probability in (8.48) is obtained by assuming that the conditional probability at t depends only on the purchase at $t-1$. This is the first-order Markov model. The Markov assumption implies that:

$$P(Y_t = j \mid Y_{t-1} = i, Y_{t-2} = k, \ldots) = P(Y_t = j \mid Y_{t-1} = i) = p_{ijt}. \tag{8.49}$$

If Y_t satisfies (8.49), it is said to follow a first-order Markov chain. In general, a stochastic process is said to be t'-order Markov if:

$$\begin{aligned} &P(Y_t = j \mid Y_{t-1} = i, Y_{t-2} = k, \ldots, Y_{t-t'} = l, \ldots) \\ &= P(Y_t = j \mid Y_{t-1} = i, Y_{t-2} = k, \ldots, Y_{t-t'} = l). \end{aligned} \tag{8.50}$$

In this section we only consider first-order ($t' = 1$) and zero-order ($t' = 0$) Markov models.

One refers to a zero-order (Markov) model, if the probability of purchasing a particular brand at t does not depend on purchasing behavior at $t-1$, $t-2$, etc. In other words, a zero-order model applies when current and future purchasing behavior does not depend on past purchase history.[12] Thus (8.49) reduces to:

$$P(Y_t = j \mid Y_{t-1} = i, Y_{t-2} = k, \ldots) = P(Y_t = j). \tag{8.51}$$

If the random variable takes two values, we obtain the Bernoulli model (8.23). If the random variable takes more than two values, we obtain the multinomial model (8.28).

Because the p_{ijt} are (conditional) probabilities, they must have the following properties:

$$0 \le p_{ijt} \le 1, \quad \text{for all } i, j = 1, \ldots, n, \quad t = 1, \ldots, T \tag{8.52}$$

$$\sum_{j=1}^{n} p_{ijt} = 1, \quad \text{for all } i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{8.53}$$

where n is the number of brands. We assume consumer homogeneity, i.e. consumers have the same p_{ijt} (compare Sect. 7.1). If one makes the additional, simplifying assumption that $p_{ijt} = p_{ij}$, for all i and j, i.e. the transition probabilities are independent of time, the resulting Markov chain is said to be stationary. The transition probabilities p_{ij} can be represented in a matrix. This transition probability matrix TP is represented as:

[12] Massy et al. (1970, Chapter 3).

$$TP = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}. \tag{8.54}$$

The diagonal elements $(p_{11}, p_{22}, \ldots, p_{nn})$ are the repeat purchase probabilities. The off-diagonal elements are the brand-switching probabilities.

We illustrate the use of Markov models by using examples and variables that are defined at the aggregate level. Hence, the probabilities are now interpreted as market shares and the transition probabilities as percentages of consumers who remain loyal (p_{ii}) or switch $(p_{ij}, i \neq j)$.

If the market shares in period t are in the (row) vector $m_t = (m_{1t}, m_{2t}, \ldots, m_{nt})$ one can use the transition probability matrix, TP, to predict market shares in future periods through the relation between the market shares at $t+1$, $m_{t+1} = (m_{1,t+1}, \ldots, m_{n,t+1})$ and market shares in period t, m_t, written in matrix formulation as:

$$m_{t+1} = m_t TP. \tag{8.55}$$

We illustrate this with an example of two brands. Consider the following transition probability matrix:

$$TP = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}$$

and assume that the current (period 0) market shares are $m_{10} = 0.50$, $m_{20} = 0.50$. In period $t = 1$, the predicted market shares are computed as the matrix product $m_t TP$:

$$\begin{aligned} \hat{m}_{11} &= m_{10} p_{11} + m_{20} p_{21} \\ &= 0.5 \times 0.8 + 0.5 \times 0.3 = 0.55 \end{aligned}$$

and

$$\begin{aligned} \hat{m}_{21} &= m_{10} p_{12} + m_{20} p_{22} \\ &= 0.5 \times 0.2 + 0.5 \times 0.7 = 0.45. \end{aligned}$$

Predicting further into the future, we use the relation:

$$m_t = m_0 TP^t \tag{8.56}$$

where $TP^2 = TP \times TP$, and $TP^3 = TP \times TP \times TP$, and so on. Using this relation, one obtains the market shares shown in Table 8.2. In equilibrium (steady state), the predicted market shares are respectively 0.60 and 0.40. These steady state market shares collected in the vector m, are independent of time, and satisfy:

Table 8.2 Market shares predicted with a first-order Markov Model

Time period	Market share brand 1	Market share brand 2
0	0.5	0.5
1	0.55	0.45
2	0.575	0.425
3	0.5975	0.4025
⋮	⋮	⋮
∞	0.600	0.400

$$m = m \times TP \leftrightarrow m(I - TP) = 0 \tag{8.57}$$

which shows that m is a (left) eigenvector of TP corresponding to the eigenvalue 1. In the particular example above with two states:

$$m = \left(\frac{p_{21}}{p_{21} + p_{12}}, \frac{p_{12}}{p_{21} + p_{12}} \right). \tag{8.58}$$

The stationary market shares in Table 8.2 can be directly computed from Eq. (8.58) as: $m_1 = 0.3/0.5$, and $m_2 = 0.2/0.5$. It can be shown that under suitable conditions the Markov chain reaches an equilibrium situation, in which $m_0 \rightarrow m$ as $t \rightarrow \infty$, regardless of the initial market shares m_0.

We illustrate the use of a Markov model through a study by Gensler et al. (2007).[13] Gensler et al. (2007) examined customers' behavioral channel loyalty and inter-channel switching behavior for a large European home-shopping company than runs two direct sales channels: a call center and an Internet channel. The cost per order is substantially lower when customers use the Internet channel. Hence, management has a strong preference to migrate customers to this channel. Gensler et al. (2007) proposed the Colombo and Morisson (1989)-model to determine

- the customers who are intrinsically loyal and stay with the same sales channel: the hard-core loyal;
- the customers who potentially switch from sales channel to another on every purchase occasion: potential switchers.

In the Colombo–Morrison-model (stationary) probabilities (p_{ii}, p_{ij}) are linked to choice probabilities (π_i) through:

$$p_{ii} = \alpha_i + (1 - \alpha_i)\pi_i \quad \text{for } i = 1,2,3 \tag{8.59}$$

$$p_{ij} = (1 - \alpha_i)\pi_j \quad \text{for } j = 1,2,3, \quad (i \neq j), \tag{8.60}$$

[13] We closely follow Gensler et al. (2007).

where

$p_{ii}(p_{ij})$ = conditional probability that a consumer who last used channel i will next use channel $i(j)$,

α_i = fraction of channel i's current customers who are completely loyal to that channel (hard-core loyals),

$\pi_i(\pi_j)$ = fraction of potential switchers who will next use channel $i(j)$.

In the Colombo–Morrison-model the diagonal elements of the matrix of transition probabilities is decomposed into repeat purchases from hard-core loyal (α_i) and soft-loyal customers $(1-\alpha_i)\pi_i$.

Call center customers are indexed by $j = 1$, Internet customers by the index $j = 2$. To accommodate customers who did not use any sales channel within the observation period, a "no purchase option" ($j = 3$) is introduced.

The estimation of the parameters is based on Maximum Likelihood. Table 8.3 shows the aggregate matrix of transition probabilities based on data on about 1.5 million customers for 15 consecutive months of observations. We may deduce from

Table 8.3 Aggregate matrix of transition probabilities

	Call center t	Internet t	No option t
Call center $t-1$	94.7	0.7	4.6
Internet $t-1$	22.6	67.9	9.5
No option $t-1$	7.0	0.5	92.5

Source: Gensler et al. (2007, p. 19)

Table 8.3 that 94.7 % of the current users of the call center and 67.9 % of the internet channel users would be considered loyal to these channels. A decomposition of these fractions shows that $\widehat{\alpha_1} = 84.3$ and $\widehat{\alpha_2} = 66.4$. This means that for the call center a substantial fraction (94.7 % – 84.3 % = 10.4 %) of the customer have considered a potential switch, which is substantially larger than for the Internet channel (1.5 % = 67.9 % – 66.4 %). Gensler et al. (2007) constructed switching matrices for different time periods. They found that the fraction of hard-core loyal call-center users experienced a significant drop off over time which was in line with management's aspiration to see more channel migration towards the more profitable Internet channel.

A specific class at Markov models is the Markov response models in which the transition probabilities p_{ijt} are related to customer characteristics and/or decision variables. Early examples are found in Leeflang (1974, Chapter 7); Horsky (1977) and Givon and Horsky (1990). A more recent example of a Markov response model is given by Freimer and Horsky (2012) who study the effect of advertising (pulsing) on buying or not buying a brand.

A general framework for including marketing decision variables and other variables into Markov models is provided by Zufryden (1981, 1982, 1986) and Jones and Landwehr (1988). They demonstrated that the multinomial logit model framework (8.28) can be used to estimate Markov response with explanatory variables. Zufryden introduces a last-state specification vector z_i, for individual i, where $z = (z_{1i}, z_{2i}, \ldots, z_{ni})$ is a vector of zeros and ones indicating the state an individual was last in. For example, $z_{1i} = \cdots = z_{n-1,i} = 0$, $z_{ni} = 1$ indicates that brand n was purchased last by individual i. This last-state specification vector is included among the explanatory variables x_{ij} in a logit specification analogous to (8.28), resulting in:

$$\pi_{ij|z_i} = \frac{\exp\left(z_i'\gamma + x_{ij}'\beta\right)}{\sum_{r=1}^{n} \exp\left(z_i'\gamma + x_{ir}'\beta\right)}. \tag{8.61}$$

Thus, this model includes past purchases of individual i as a predictor in a logit model with a parameter vector γ indicating the effect of a previous brand purchase on the probability to choose j, $\pi_{ij|z_i}$. Including $z_i'\gamma$ into the logit model (8.61) results in a specification of the conditional probability of choosing j, given a previous purchase by individual i, indicated by z_i.

The zero-order multinomial model (8.28) is a restricted version of the model (8.61). The first-order Markov hypothesis is maintained if we do not reject the null hypothesis $\gamma = 0$. If the evidence favors (8.61), the implication is that if the values of the explanatory variables x_{ij} change, the first-order Markov transition probabilities also change. We note that a more general formulation can be obtained if the impact of the explanatory variables is allowed to depend on the last brand purchased, which amounts to including an interaction of x_{ij} and z_i in Eq. (8.61).

8.2.4.2 Hidden Markov Models

In the past decade the attention for Markov models has been intensified through the development of Hidden Markov Models (HMMs). In our discussion of Markov models so far, we mainly used the Markov property to capture the "transitions" of customers between brands over time. More generally, we can define "states" to describe the behavior of customers. For example, in brand choice Markov models, a customer is in state j at time t, if brand j was purchased in t by that customer. In such cases, the "states" are observed. In HMMs however, the states are not observed (or latent), but the state that customer i is in at time t influences her observed behavior at time t.

We illustrate these ideas with a study by Netzer et al. (2008)[14] in which they capture the dynamics of customer relationships using a HMM. They consider a university-alumni customer relation data set, and identify three states: dormant, occasional, and active donors. These states are unobserved, but influence the observed donations of the alumni. Alumni can move from one state to another over time, and these transitions are determined by several university-alumni (time-varying) interactions, also called relationship encounters, such as reunions and possibilities to volunteer for a university role. These variables fulfill a similar role as the decision variables in (8.61).

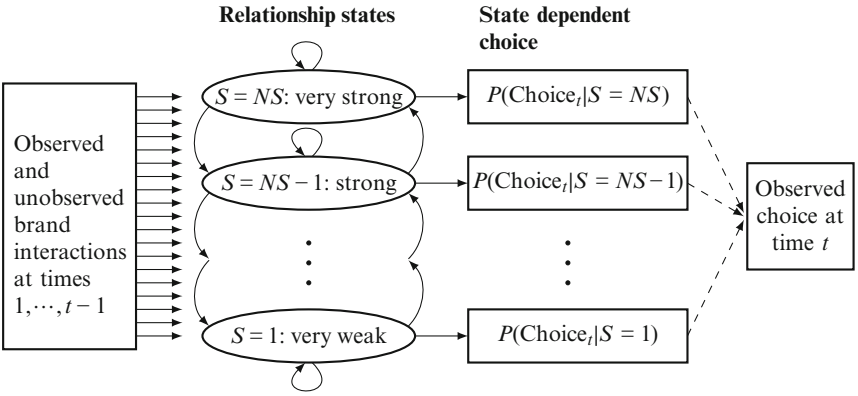

Fig. 8.3 Structure of a Hidden Markov Model of Customer Relationships. *Source*: Netzer et al. (2008, p. 189)

Figure 8.3 graphically illustrates the structure of their Hidden Markov Model for customer relationships. From left to right we see that the customer-"brand" interactions until time $t-1$ affect the state a customer is in at time t. The states S range from $S = 1$, indicating a very weak relationship to $S = NS$, which means that the customers perceives a very strong relationship with the brand. These states are not observed, but they influence the probability that the brand is chosen, which in turn influences the observed brand choice at time t.

The relationship state of customers in Fig. 8.3 can change over time, and the transitions between the states are modeled as a Markov process. However, since the state of a customer at time t also depends on customer-brand interactions until time $t-1$, the transition probabilities are not constant over time, which means that the associated Markov chain is not stationary. The customer- and time-specific transition probability matrix is given in Eq. (8.62):

[14]Other examples of HMM's in marketing are Liechty et al. (2003); Montgomery et al. (2004); Moon et al. (2007); Paas et al. (2007); Ebbes et al. (2010); Kumar et al. (2011); Schwartz et al. (2014); Zhang et al. (2014).

$$TP_{i,t} = \begin{array}{c} \\ \text{State at time } t-1 \\ \\ 1 \\ 2 \\ \vdots \\ NS \end{array} \begin{array}{c} \begin{array}{cccccc} \multicolumn{6}{c}{\text{State at time } t} \\ 1 & 2 & 3 & \ldots & NS-1 & NS \end{array} \\ \begin{pmatrix} p_{i,t,1,1} & p_{i,t,1,2} & 0 & \cdots & 0 & 0 \\ p_{i,t,2,1} & p_{i,t,2,2} & p_{i,t,2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{i,t,NS,NS-1} & p_{i,t,NS,NS} \end{pmatrix} \end{array}. \quad (8.62)$$

In Eq. (8.62) $p_{i,t,s,s'}$ is the probability that customer i moves from relationship state s to relationship state s' at time t. HMMs are usually estimated using a Markov chain Monte Carlo (MCMC) Hierarchical Bayes procedure (see Sect. 6.8 and Rossi and Allenby 2003).

8.3 Purchase Quantity Models

8.3.1 General Structure

Purchase quantity models describe the total number of units bought of a particular brand (brand sales) or product category (industry sales).[15] The dependent variable of interest in these "count models" is a count variable which can take on nonnegative integer values: $\{0, 1, 2, \ldots\}$. We are interested in cases where the dependent variable takes on relatively few values, including zero. For the same reason discussed for binary and multinomial responses the General Linear Model might not provide the best fit given that the distribution of the disturbances and hence the distribution of the dependent variable, in general, will be very different from normal. Instead the distribution for count data is the Poisson distribution (Wooldridge 2012, p. 580).

Early purchase quantity models were applied in marketing by Ehrenberg (1959, 1972, 1988) and Chatfield et al. (1966). These models are based on the Poisson process, which has the property that the distribution of the number of units purchased in any interval depends only on the length of that interval. The random variable (Y_{it}), denoting the number of units purchased by consumer i in a certain time period t, then follows a Poisson distribution with parameter λ:

$$P(Y_{it} = y_{it}) = f(y_{it} \mid \lambda, t) = \frac{e^{-\lambda t}(\lambda t)^{y_{it}}}{y_{it}!}, \quad i = 1, \ldots, I, \quad t = 1, \ldots, T.$$

[15] Purchase quantity models at the product category level have often been used to explain the composition of shopping baskets. See, for example, Manchanda et al. (1999); Seetharaman et al. (2005); Blattberg et al. (2008, Chapter 13); Chen and Steckel (2012).

The Poisson process has expectation: $E(Y_{it}) = \lambda t$, which shows that λ can be interpreted as the rate of the process. Its variance is equal to the mean. The probability of at least one purchase in the interval t, the penetration, which is of primary interest in purchase incidence models is:

$$P(Y_{it} > 0) = 1 - e^{-\lambda t}. \tag{8.63}$$

The interpurchase times in the Poisson process follow an exponential distribution with mean $1/\lambda$:

$$f(t \mid \lambda) = \lambda e^{-\lambda t}. \tag{8.64}$$

An estimator for λ in the Poisson process is simply the mean of the observed purchase frequencies: $\hat{\lambda} = \sum_{i=1}^{I} y_{it}/It$, where I is the total number of consumers.

8.3.2 Heterogeneity in Count Models

The assumptions underlying the Poisson process are quite restrictive in many marketing applications. For example, the assumption that all consumers have the same value of λ is unrealistic. Heterogeneity has been accommodated in several ways, most frequently by assuming that λ is a random variable that follows a gamma distribution across individuals:

$$f(\lambda \mid \beta, \alpha) = \frac{\alpha^{\beta} \lambda^{\beta-1} e^{-\alpha\lambda}}{\Gamma(\beta)} \tag{8.65}$$

with α and β being parameters of the gamma distribution and $\Gamma(\cdot)$ the gamma function. The gamma distribution is a very flexible distribution that can take on a variety of shapes. We note that if $\beta = 1, 2, 3, \ldots$, takes on integer values, then an Erlang distribution arises. From (8.63) and (8.65) the number of purchases for a (randomly selected) individual can be shown to follow a Negative Binomial Distribution (Ehrenberg 1959; Morrison and Schmittlein 1988; East and Hammond 1996):

$$f(y_{it} \mid \beta, \alpha) = \binom{y_{it} + \beta - 1}{y_{it}} \left(\frac{\alpha}{\alpha + t}\right)^{\beta} \left(\frac{t}{\alpha + t}\right)^{y_{it}}. \tag{8.66}$$

The NBD has expectation: $\mathrm{E}(Y_{it}) = \beta t/\alpha$, and its variance is: $\mathrm{Var}(Y_{it}) = \beta t/\alpha + \beta t^2/\alpha^2$. Thus, the variance of the NBD exceeds that of the Poisson distribution, which equals the mean $\beta t/\alpha$. The probability of at least one purchase in the interval t, the penetration, is:

$$P(Y_{it} > 0) = 1 - \left(\frac{\alpha}{\alpha + t}\right)^{\beta}. \tag{8.67}$$

Estimators for α and β can be derived for example from the estimated mean $\bar{y}$ and variance σ^2 of the NBD: $\hat{\alpha} = (\bar{y}t)/(\hat{\sigma}^2 - \bar{y})$ and $\hat{\beta} = (\hat{\sigma}\bar{y})/t$. Morrison and Schmittlein (1988) derive conditions under which the NBD at the brand level leads to a NBD at the product class level. In empirical applications, the NBD seems to fit either well at both levels or at neither level.

In the past decades many other solutions than the gamma distribution are suggested to accommodate heterogeneity in the purchase rate.[16] A specific category of models arises when the heterogeneity distribution is assumed to be discrete: the so-called finite mixture models.

A problem of the Poisson and NBD purchase incidence models is that they do not accommodate individuals who never buy. Both models predict that every individual will eventually buy the product, as t increases. Since for most products and categories there is a group of individuals who never buy, these purchase quantity models tend to underestimate the percentage of zero purchases. One solution (Morrison and Schmittlein 1988) is to add a component to the model that allows for an additional spike at zero, due to the class of non-buyers, with proportion π_0. The Zero-Inflated Poisson (ZIP) model (also called the "Morrison-model") is:

$$P(Y_{it} = y_{it}) = \pi_0 + (1 - \pi_0)\frac{e^{-\lambda t}(\lambda t)^{y_{it}}}{y_{it}!}. \tag{8.68}$$

The ZIP model has a mean: $E(Y_{it}) = \pi_0 + (1 - \pi_0)\lambda t$, and a variance: $Var(Y_{it}) = \lambda t(1 - \pi_0)(1 - \pi_0\lambda t)$. The penetration is:

$$P(Y_{it} > 0) = (1 - \pi_0)(1 - e^{-\lambda t}). \tag{8.69}$$

Estimates of π_0 and λ can be obtained from the equations for the mean $\bar{y} = (1 - \hat{\pi}_0)\hat{\lambda}t$, and the proportion of zeros: $P_0 = \hat{\pi}_0 + (1 - \hat{\pi}_0)e^{-\hat{\lambda}t}$. These equations need to be solved by iteration.

A straightforward extension accounts for both added zeros and heterogeneity, giving rise to the Zero-Inflated Negative Binomial Distribution (ZINBD). See Schmittlein et al. (1993) for an application.

Much research in purchase quantity modeling has focused on finding the appropriate distribution for purchase frequencies, and estimating important quantities from the data, including penetration, lost buyer percentages and so on. More recently, successful attempts have been made to *include marketing decision variables* as predictors into these models. The primary extension is that one parameterizes the mean of the distribution as a function of predictors. In such response types of purchase quantity models, the effects of decision variables

[16] See Sichel (1982) who has proposed the family of generalized inverse Gaussian distributions; Sikkel and Hoogendoorn (1995) who consider the inverse Gaussian, the lognormal and the Weibull distributions. Abe (2009) extends the Poisson purchase incidence model in several ways to account for drop outs of customers of a customer-base. See also: Fader et al. (2005).

are measured by changes in the shapes and/or the parameters of the probability distributions.

We consider the simplest case, a Poisson distribution for the purchase quantity of a particular product. Assume there are marketing decision variables x_ℓ, $\ell = 1,\ldots,L$, including for example the product's price, frequency of promotions, etc., but also consumer characteristics $z_{i\ell'}, \ell' = 1,\ldots,L'$, such as demographic and socio-economic variables. The idea is that the expected number of purchases by individual i in the period under consideration (μ_i), is related to those explanatory variables:

$$\mu_i = \exp\left(\sum_{\ell=1}^{L}\beta_\ell x_\ell + \sum_{\ell'=1}^{L'}\beta_{\ell'} z_{i\ell'}\right). \tag{8.70}$$

Model (8.70) accounts for variation in the number of purchases across the sample as a function of the explanatory variables. The mean of the purchase quantities is μ_i. The exponent in (8.70) guarantees that the predicted purchase amount is positive. With this formulation we can assess the effects of marketing variables or consumer characteristics. For example, if "Gender" has been included as an explanatory variable (0 = male; 1 = female) and the corresponding parameter is estimated as $\hat{\beta}_{\ell'} = 0.693$, then $\exp(\hat{\beta}) = 2$, and the expected purchase quantity is 2 times higher for females than for males. In Sect. 6.4.5 we already discussed an application of a Poisson model. We return to this application in Sect. 8.5.2.

8.4 Purchase Timing: Duration Models

8.4.1 Introduction

Duration models deal with duration or timing variables. Examples of duration variables are: interpurchase time, the time between sending a mailing and the response to that mailing, the time between the introduction of a new product and the adoption of the new product by customer i, etc. Duration models do *not* answer questions such as which brand/category is chosen and how many units are purchased. The questions that are answered by duration models deal with "when" will a category/brand be purchased. Duration variables which measure the time between two events are often censored. In many cases we do not know events *prior* to the observation period (left censoring). In other cases the event (e.g. a purchase, adoption, customer churn) did not happen yet at the end of the observation period (right censoring). If the event can only occur during the observation period, the duration variable is fully observed and hence *uncensored*.

8.4.2 Hazard Models

Researchers have used hazard models to account for censored duration variables.[17] The major advantage of this approach is that it accounts for so-called right-censoring. Right censoring occurs if a sample of consumers or households is observed for a time period of fixed length only, causing longer interpurchase times to have a larger probability of falling (partially) outside the observation period. If one does not account for right-censoring, the estimates are biased.

In hazard models the probability of a purchase during a certain time interval, say t to $t+\Delta t$, given that is has not occurred before t, is formulated as:

$$P(t \leq T \leq t+\Delta t \mid T \geq t) \tag{8.71}$$

where T is the random interpurchase time variable.[18] Parametric methods for interpurchase times involve assumptions about their distribution. Two distinct classes of hazard models arise according to whether a discrete or a continuous distribution of the interpurchase times is assumed.

In *discrete-time* hazard models, the probability of a purchase in Eq. (8.71) is specified directly for given values of Δt. In the *continuous-time* approach, Δt approaches zero in (8.71), to yield a continuous hazard rate $\lambda(t)$:

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T \leq t+\Delta t \mid T \geq t)}{\Delta t}. \tag{8.72}$$

The hazard rate can be interpreted as the instantaneous rate of purchasing at time t, given that purchasing has not occurred until t. The distribution function of interpurchase times can then be derived:

$$f(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T \leq t+\Delta t)}{\Delta t} = \lambda(t) P(T > t). \tag{8.73}$$

Alternatively, (8.73) can be written as $f(t) = \lambda(t) S(t)$. Here, $S(t) = P(T > t)$ is the probability (or survival) that a purchase has not yet occurred at t where $S(t)$ is the so-called survivor function. The advantage of this formulation becomes apparent in case of censoring, because $f(t)$ represents the density of any *uncensored observation* such as a completely observed interpurchase time. If, due to censoring, an interpurchase time is not completely observed, $S(t)$ provides the probability that the purchase has not yet occurred, since we only know that the interpurchase time is larger than t, the end of the observation period. The continuous-time approach

[17] See, for example, Gupta (1991); Jain and Vilcassim (1991); Gönül and Srinivasan (1993); Helsen and Schmittlein (1993); Wedel et al. (1995); Dekimpe et al. (2000); Prins and Verhoef (2007); Schweidel et al. (2008); Risselada et al. (2014).

[18] An extension of the formulation to multiple purchases is straightforward.

appears to be the most commonly used approach in the marketing literature. If the interpurchase times follow an exponential distribution (and the purchase incidence is Poisson), then the hazard rate and survival functions are:

$$\lambda(t) = \lambda \tag{8.74}$$

and

$$S(t) = e^{-\lambda t}. \tag{8.75}$$

Thus, for the exponential distribution the hazard of a purchase is constant and independent of time. Also, purchases occur at random time periods, independent of past purchases. For other distributions the hazard rate and survivor functions can be formulated as well.

Flinn and Heckman (1983) proposed a very flexible (Box–Cox) formulation for hazard functions. This is applied to brand switching problems by Vilcassim and Jain (1991). It includes many of the frequently used distribution functions as special cases, and the hazard rate is formulated as:

$$\lambda(t) = \exp\left(\gamma_0 + \sum_{\ell=1}^{L} \gamma_\ell \frac{t^{\nu_\ell} - 1}{\nu_\ell}\right). \tag{8.76}$$

This formulation includes a variety of interpurchase time distributions as special cases. For example, if $\gamma_\ell = 0$ for all ℓ, then $\lambda(t) = \exp(\gamma_0)$ is constant, and the exponential distribution arises. If $\gamma_\ell = 0$ for $\ell > 2$, and $\nu_\ell = 1$, then $\lambda(t) = \exp[\gamma_0 - \gamma_1 + \gamma_1 t]$ which is the hazard rate of a Gompertz distribution. For all interpurchase time distributions except for the exponential, the hazard rate varies with time, so that those distributions account for nonstationarity of purchase timing and incidence (i.e. their distribution is allowed to change over time). A decline in the event rate as a function of the elapsed time is often called inertia and is observed for interpurchase times (Helsen and Schmittlein 1993). A hazard rate that increases with time is often called a snowballing phenomenon, observed in new-product adoption processes (Helsen and Schmittlein 1993). Another useful way to accommodate nonstationarity is through the so-called piecewise exponential formulation. The piecewise exponential formulation assumes that the total observation period can be decomposed into, say, R shorter periods, in which the hazard is (approximately) constant. Thus, one considers, for example, R weekly or monthly purchase intervals, denoted by t_r, $r = 1, \ldots, R$, and $t_{r+1} = t_r + \Delta t$ where:

$$P(t_r \leq T \leq t_{r+1}) = \lambda_r. \tag{8.77}$$

Note that the hazard rate is constant within each period, but varies from period to period, thereby accounting for nonstationarity. This approach is quite flexible, because across time periods the hazard may take an arbitrary form. Since the hazard

between t_r and t_{r+1} is constant for all r, the model assumes random purchase timing within each period. Several hazard models have been compared empirically.[19] One of the main conclusions is that the simple exponential model does not seem to describe interpurchase timing processes well. Each of the models used in the studies referred above, however, is more complex than the ones described here in that they accommodate heterogeneity of the hazard rate across subjects, and/or include marketing decision variables. We discuss these two topics below.

8.4.3 Heterogeneity in Duration Models

The hazard modeling approach discussed in the previous subsection assumes that consumers are homogeneous. Gupta (1991) handled heterogeneity in a way that is comparable to the approach described before in purchase incidence models. He assumed a parametric distribution of interpurchase times, specifically exponential or Erlang-2, and let heterogeneity be captured by a gamma distribution for the scale parameter β. The advantages of this approach are that it provides a natural extension of the NBD models discussed in Sect. 8.3, and that β can be directly interpreted as a measure of heterogeneity.

Heterogeneity can also be accommodated through the inclusion of consumer characteristics and other explanatory variables. The basic idea for models that allow marketing decisions to influence purchase timing is to reformulate the hazard function. An important case is presented by so-called *Proportional hazard models* (Cox 1975), for which:

$$\lambda(t \mid x(t)) = \lambda_0(t) \exp\left(x(t)'\beta\right) \tag{8.78}$$

where

$$\begin{aligned} \lambda_0(t) &= \text{the baseline hazard,} \\ x(t) &= \text{a vector of explanatory variables at time } t, \\ \beta &= \text{vector of parameters.} \end{aligned}$$

Model (8.78) has two multiplicative components. The first term, $\lambda_0(t)$ is called the baseline hazard, and captures the changes of the hazard over time. The baseline hazard is the same for all customers, and is often modeled as a polynomial function in t. As an example, consider the case when customers become more likely to adopt an innovation over time. The upward trend in the baseline hazard can be modeled with a linear function in t. In *proportional hazard models* the baseline hazard can take any of the discrete or continuous time specifications outlined above [compare (8.71)].

[19] See, for example, Jain and Vilcassim (1991); Bayus and Mehta (1995); Wedel et al. (1995); Chang et al. (1999); Schweidel et al. (2008).

The second term in Eq. (8.78) shifts the hazard up or down *proportionally* dependent upon the effects of marketing decision variables and/or customer characteristics. The hazard is shifted upward if the explanatory variables and the corresponding are such that $\exp(x(t)'\beta) > 1$, and downward if $\exp(x(t)'\beta) < 1$.

Consider the case where we are modeling response time to adoption of a new product and that we like to know whether there are differences in the adoption role between men and women. We assume that x (in (8.78)) is a dummy variable indicating gender: 1 =female, 0 =male. The hazard rate for female consumers is $\lambda_0(t)\exp(\beta_{gender})$ and for male consumers: $\lambda_0(t)\exp(0) = \lambda_0(t)$. It is found that $\beta_{gender} = \ln 2$. Using the so-called *hazard ratio* (HR) is it found that:

$$\text{HR} = \frac{\lambda_0(t)\mathrm{e}^{\ln 2}}{\lambda_0(t)} = 2, \tag{8.79}$$

which means that the adoption rate of women is twice that of men, at any time t. If the parameter which corresponds to a variable x is β then the hazard increases by $\{100 \times (\exp(\beta) - 1)\}\%$ for a unit increase of x. This example demonstrates why (8.78) is called a proportional hazard model.[20]

We now specify a hazard model developed by Sinha and Chandrashekaran (1992) which is used to model the diffusion of an innovation, or rather the timing of the adoption of an innovation. They explicitly account for the fact that a proportion of the subjects will never adopt the product. If the proportion of adopters is denoted by π_0, then the probability of observing a certain adoption time, given that the adoption occurs in the time period under study ($y = 1$) is:

$$f(t \mid y = 1) = \pi_0 \lambda(t) S(t) \tag{8.80}$$

which is the probability of an adoption times the rate of purchase at time t multiplied by the probability of no purchase at time t. The probability of not adopting during the sample period ($y = 0$) is:

$$f(t \mid y = 0) = (1 - \pi_0) + \pi_0 S(t) \tag{8.81}$$

which is the probability of never adopting plus the probability of eventual adoption after the sample period (i.e. censored in the particular sample). If the distribution of the adoption time is assumed to be exponential, so that the hazard rate and survival function are provided by (8.74) and (8.75) respectively, this model is equivalent to the ZIP model for purchase incidence described in Sect. 8.3.

[20] An alternative model is the so-called accelerated lifetime (hazard) model. In this model one may scale (or accelerate) t by a function of explanatory variables (Franses and Paap 2001, pp. 165–166).

8.4.4 Estimation and Validation of Duration Models

The estimation of duration models is usually done via ML-methods that we discuss next. The model fit is determined using the estimated values of the likelihood (8.14), the log-likelihood (8.15), the pseudo R^2s introduced in Sect. 8.2, etc. Model selection is performed using the likelihood ratio test (nested models) and the use of the familiar AIC and BIC as discussed in Sect. 5.6.3. An illustration is presented in Table 8.4 in which we show the result of a study by Prins and Verhoef (2007) that we discuss next.

Prins and Verhoef (2007) specified and estimated a hazard model in which they account for a group of customers that will probably never adopt a new service of a telecommunications operator. The probability of adoption of these customers is zero. They applied the so-called *split hazard approach* developed by Schmidt and Witte (1989). In this approach both the adoption probability and the adoption timing of the new service of existing customers are modeled. Adoption timing is modeled as a hazard function of both time-varying marketing communication effects and time-invariant covariates such as relationship characteristics and customer characteristics. The hazard part of the Prins–Verhoef-model is specified as:

$$\begin{aligned}\lambda_i(t) = 1 - \exp\Big[-\exp\big(&\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 DMC_{it} + \beta_4 SA_t + \beta_5 SA_{t-1} + \beta_6 BA_t \\ &+ \beta_7 BA_{t-1} + \beta_8 CSA_t + \beta_9 CSA_{t-1} + \beta_{10} CBA_t + \beta_{11} CBA_{t-1} + \beta_{12} RA_i + \beta_{13} RA_i^2 \\ &+ \beta_{14} SU_i + \beta_{15} SU_i^2 + \beta_{16} Age_i + \beta_{17} Age_i^2 + \beta_{18} Gend_i + \beta_{19} Innov_i\big)\Big] \qquad (8.82)\end{aligned}$$

where

DMC_{it} = dummy for individual offer by telephone by customer i,

SA_t = advertising expenditures in millions of euros in months related to the new service in t,

BA_t = advertising expenditures in millions of euros in month t, not related to the new service

CSA_t = competitive advertising expenditures in t related to the service,

CBA_t = competitive advertising expenditures in t not related to a similar service,

RA_i = number of years customer i has been with the provider,

SU_i = average monthly amount spent by i before $t = 1$,

$Gend_i$ = gender (male = 0, female = 1),

Age_i = age at $t = 1$, and

$Innov_i$ = dummy for usage of prior generation of service by i.

Prins and Verhoef (2007) model the unobserved probability of eventual adoption as a logit function of time-invariant customer characteristics and relationship characteristics:

$$\pi_i = 1/\Big[1+\exp\Big(-\big\{\gamma_0+\gamma_1 RA_i+\gamma_2 RA_i^2+\gamma_3 SU_i+\gamma_4 SU_i^2 + \gamma_5 Age_i+\gamma_6 Age_i^2+\gamma_7 Gend_i+\gamma_8 Innov_i\big\}\Big)\Big] \tag{8.83}$$

where π_i is the probability of eventual adoption.

The log-likelihood function of the total model is:

$$LL = \sum_{i=1}^{N} d_i \times \ln\left[\pi_i \times \lambda_i(t) \times S_{i,t-1}\right] + (1-d_i) \times \ln\left[(1-\pi_i)+\pi_i \times S_{i,t}\right] \tag{8.84}$$

where

N = number of customers,

d_i = 1 if adoption is observed and 0 if the observation is censored, and

$S_{i,t}$ = the survival rate.

From (8.84) we observe that the contribution to the likelihood function by customer i at time t is the probability that he or she will eventually adopt, as given by π_i, multiplied by both the conditional probability of adoption at t, as given by the hazard rate $\lambda_i(t)$, and the probability that he or she has not adopted before t as given by the survival rate $S_{i,t-1}$. Prins and Verhoef estimate this model by maximum likelihood. Table 8.4 shows the estimation results of the adoption model (π_i) and the hazard part.[21] From Table 8.4 we conclude from the logit part of the model [Eq. (8.83)] that relationship age has a positive (although diminishing) effect on adoption and service usage rate has a negative effect. Direct marketing communication, service and brand advertising have a positive effect on adoption timing. For example, the parameter for direct marketing communication (dummy for whether or not an individual offer by telephone has been made) is estimated as 1.865, which means that a phone call leads to an exp(1.865) = 6.457 times higher hazard rate, and a shorter expected time to adoption. Competitive service advertising has a positive effect on the adoption of the new service which is due to "market marking" effects. Competitive brand advertising has a significant negative effect which implies that it lengthens the time to adoption.

[21] We do not show the results that have been obtained through the inclusion of interaction effects: see Table 3 in Prins and Verhoef (2007).

8.5 Integrated Models

8.5.1 Integrate Incidence, Timing and Choice

In the previous sections we described models for purchase incidence, brand choice, purchase quantity and purchase timing separately. Several authors have integrated such descriptions of separate consumer behavior processes into one single framework. Some of these models have also included components of market segmentation. The purpose of those models is to provide managers with insights about the possible sources of gains and losses of sales, about the consumer characteristics and marketing variables that affect sales, and about the causes of brand loyalty and brand switching. Many of the approaches compound the (Poisson- or NBD-)

Table 8.4 Estimation results of Eqs. (8.83) and (8.84)

	Logit part: p(Adoption)		Hazard part: time to adoption	
Variables	Coefficient	z-value	Coefficient	z-value
t	NA		0.1618	1149***
t^2	NA		−0.0048	−9.82***
DMC_{it}	NA		1.8652	43.20***
SA_t	NA		0.2243	7.22***
SA_{t-1}	NA		0.0012	0.05
BA_t	NA		0.0521	2.25**
BA_{t-1}	NA		0.0432	1.88*
CSA_t	NA		0.0969	4.85***
CSA_{t-1}	NA		−0.0065	−0.29
CBA_t	NA		−0.0863	−10.26***
CBA_{t-1}	NA		−0.0078	−0.85
RA_t	0.8493	1.75*	0.0521	1.00
RA_t^2	−0.1296	−1.86*	−0.0033	−0.42
SU_t	−0.2164	−3.00***	0.0580	4.45***
SU_t^2	0.0032	2.35**	−0.0009	−2.45**
Age_i	−1.6270	−0.88	−0.2123	−1.55
Age_i^2	0.2413	0.97	−0.0075	−0.46
$Gend_i$	0.4401	0.62	−0.1486	−2.44***
$Innov_i$	−0.2685	−0.44	0.3166	2.48***
Constant	5.5730	1.67*	−3.6189	−11.98***
Log-likelihood	−13,586.947			
Likelihood ratio test	$\chi^2(28) = 1935.99$***			
AIC statistic	27,229.89			
BIC statistic	27,401.83			

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$ (two-sided)
Note: *NA* not applicable
Source: Prins and Verhoef (2007, p. 178)

distributions for purchase frequency or distributions of purchase timing (exponential or Erlang) with (multinomial, Dirichlet-Multinomial, or Markov) models of brand choice. Early examples of this approach are Jeuland et al. (1980) and Goodhardt et al. (1984). However the importance of this stream of research was recognized through the seminal work of Gupta (1988). The basic setup of these approaches is as follows:

1. Assume a Poisson distribution for purchase frequency, y: $\mathrm{P}(y \mid \lambda)$.
2. Assume a gamma heterogeneity distribution for the purchase rate: $\mathrm{G}(\lambda)$.
3. Obtain the unconditional distribution of y by integrating out the heterogeneity distribution. This leads to a NBD (see Sect. 8.3):

$$\mathrm{NBD}(y) = \int \mathrm{P}(y \mid \lambda)\mathrm{G}(\lambda)d\lambda.$$

4. Assume a multinomial distribution for choice, x: $\mathrm{M}(x \mid p)$.
5. Assume a Dirichlet heterogeneity distribution for the choice probabilities: $\mathrm{D}(p)$.
6. Obtain the unconditional distribution of x by integrating out the heterogeneity distribution. This leads to a Dirichlet-Multinomial (DM) distribution:

$$\mathrm{DM}(x) = \int \mathrm{M}(x \mid p)\mathrm{D}(p)dp.$$

7. In the final step, the joint distribution of purchase frequency and choice is obtained, assuming independence: $\mathrm{NBDM}(y, x) = \mathrm{NBD}(y) \times \mathrm{DM}(x)$.

From this framework, important quantities such as market share, penetration, duplication, brand switching and repeat buying can be obtained. Most approaches in this field follow a similar format, but there are many variations.[22]

8.5.2 Tobit Models

8.5.2.1 Introduction

Tobit models constitute a specific subset of "integrated" models. Tobit models are used when the dependent response variable (usually demand or response to a mailing, a promotion, etc.) is a limited continuous variable. The demand and/or response are only observed for those customers who purchase/respond. For a customer that does not purchase/respond, the variable equals zero. Given that we usually do not have information about customers who do not buy/do not respond

[22] Some examples of integrated models are Böckenholt (1993a,b); Bucklin et al. (1998); Song and Chintagunta (2007); Andrews and Currim (2009); Vroegrijk et al. (2013). An aggregate level version of the integrated model by Gupta (1988) was proposed by Pauwels et al. (2002), who show that the long-term brand sales effects of price promotions are mostly due to lifts in category incidence, not brand choice.

within the given observation period, the continuous demand variable is *censored*. In Tobit models one specifies two equations: one to explain the decision to respond and one to explain the demand (in units or amount spent). Hence, in these models one integrates two decisions: which factors determine whether one responds or not and, if there is a response, which factors determine the size of the response. In Tobit models one accounts for values of the dependent variable which are zero or below another threshold. One may, for example, consider only these expenditures of a customer on a shopping trip that exceed $ 10. Then the threshold is $ 10. In this way one "truncates" the values of the (continuous) dependent variable.[23] If one does not account for the truncated observations the parameter estimates are biased.[24]

To describe the fluctuations of a censored dependent variable one may apply the Censored Regression models such as the Type-1 Tobit and Type-2 Tobit models.

8.5.2.2 Type-1 Tobit Model

In Sect. 8.2.2 we specify the relation between a latent variable y_i^* and Y_i, the observed dependent variable: compare Eqs. (8.5) and (8.6). In the Tobit models indicators are also related to the latent variable y_i^*, like in (8.5). The specification of a Type-1 Tobit model is:

$$Y_i = 0 \quad \text{if} \quad y_i^* = x_i'\beta + \varepsilon_i \leq 0 \tag{8.85}$$

$$Y_i = x_i'\beta + \varepsilon_i \quad \text{if} \quad y_i^* = x_i'\beta + \varepsilon_i > 0 \tag{8.86}$$

where

Y_i = a variable for individual i which takes the value 0 if a latent/unobserved variable y_i^* is smaller than or equal to zero, and $x_i'\beta + \varepsilon_i$ if this latent variable is positive,

x_i = a vector of explanatory variables including an intercept,

β = a vector of unknown parameters,

ε_i = a normally distributed disturbance term: $\varepsilon_i \sim N(0, \sigma^2)$.

[23] One way to deal with these truncations is to estimate a model using Truncated Regression. See, for example, Wooldridge (2012, pp. 589–591).

[24] This bias can be measured by the so-called inverse Mills ratio; see, for example, Franses and Paap (2001, p. 138).

The equations (8.85) and (8.86) may represent the online-demand for a product by customer i. The equations jointly model the decision whether or not to purchase, and if so, how much is purchased. Note that in the Type-1 Tobit, model, the same set of explanatory variables and effect parameters are used for (1) explaining whether or not a purchase will be made by individual i ($y_i^* \leq 0$ versus $y_i^* > 0$) and (2) how much will be purchased if a purchase does occur Y_i. Hence, the same drivers and effects are assumed for both outcomes. That might be a heroic assumption in many applications. The Type-2 Tobit model relaxes this assumption.

8.5.2.3 Type-2 Tobit Model

The Type-2 Tobit model can be formulated as:

$$Y_i = 0 \quad \text{if} \quad y_i^* = x_i'\alpha + \varepsilon_{1i} \leq 0 \tag{8.87}$$

$$Y_i = x_i'\beta + \varepsilon_{2i} \quad \text{if} \quad y_i^* = x_i'\alpha + \varepsilon_{1i} > 0 \tag{8.88}$$

where, at least in principle, $\alpha \neq \beta$. The disturbances ε_{1i} are $N(0,1)$-distributed[25] and $\varepsilon_{21} \sim N(0, \sigma_2^2)$.[26] Both error terms may be correlated: $E(\varepsilon_{1i}, \varepsilon_{2i}) = \sigma_{12}$. In this model we may account for the fact that age income and experience may have other effects on ordering online than the amount spent if people order. So, for example, age and income have a negative effect on the probability that one orders. These variables may, however, have a positive effect on the amount spent online. Note that Type-2 Tobit model in Eqs. (8.87) and (8.87) accommodates cases where the decision to order is driven by other explanatory variables than the decision how much to order (by setting appropriate elements of α and β to zero)

The Tobit models can be estimated by ML-methods. A simpler method to obtain estimates is known as the Heckman (1976, 1979) two-step procedure. This procedure is also used in cases where the variables in the right-hand part (also called the participation part) are different from the variables in the left-hand part (also called the intensity equation) (Greene 2012, Chapter 19). The model which is specified in this way is also known as an hurdle model.[27]

Example To conclude this chapter we give an example of the use of a Type-2 Tobit model. To this end, we return to the study by Van Nierop et al. (2011) that we discussed in Sect. 6.4.5.[28] Van Nierop and colleagues studied the effect of the

[25] They correspond with the Probit part.

[26] This is the standard regression model for the positive values of y_i^*.

[27] First the parameters of the first equation (the probit part) are estimated using ML. Correcting for the bias using the inverse Mills ratio, the parameters of the second equation can be estimated using a regression model.

[28] Other examples of the application of Type-2 Tobit models in marketing are found in, for example, Bucklin and Sismeiro (2003); Fox et al. (2004); Prins et al. (2009); Danaher and Dagger (2013).

introduction of an informational website on the number of store visits of a large retailer. As discussed in Sect. 6.4.5, they find negative effects of the number of website visits on the number of store visits. Using a Type-2 Tobit model they also investigate how the variables that are discussed in Sect. 6.4.5 affect:

1. the *purchase incidence* of a consumer to buy in a category c in month t, and
2. the *amount spent* by individual i in category c in month t.[29]

Following Eqs. (8.87) and (8.88) a Type-2 Tobit model for Y_{itc}, the amount that individual i spends in category c in month t, can be written as:

$$Y_{itc} = 0 \quad \text{if} \quad y^*_{itc} = z'_{itc}\alpha + \varepsilon_{1itc} \leq 0 \tag{8.89}$$

$$Y_{itc} = z'_{itc}\beta + \varepsilon_{2itc} \quad \text{if} \quad y^*_{itc} = z'_{itc}\alpha + \varepsilon_{1itc} > 0 \tag{8.90}$$

where z_{itc} is a $K-$vector of explanatory variables for customer i, in month t, for category c. The explanatory variables in Eqs. (8.89) and (8.90) are the same as those that are discussed in Sect. 6.4.5, with the exception of the number of website visits. For the Type-2 Tobit model this variable is measured per category, which explains the addition of an index c. Furthermore, α and β allow for individual-specific effects, so that variables that only vary across customers drop out. For this reason, the age variable is not included in this analysis.

The elements in α describe the effect of the explanatory variables on purchase incidence, whereas the elements in β indicate which explanatory variables influence the amount spent by customer i in category c in month t. Van Nierop et al. (2011) specify multivariate error distributions for ε_{1itc} and ε_{2itc} to accommodate contemporaneous correlation of the disturbances across product categories, which might occur because excess expenditures in one category may result in either less spending in substitute categories, or in additional spending in complementary categories.

Based on pooling test, Van Nierop and colleagues estimate separate Type-2 Tobit models for six product categories: Ladies' Fashion, Men's Fashion, Children's products, Accessories, Living and Sport. Table 8.5 contains the outcomes that are obtained using Markov Chain Monte Carlo (MCMC) methodology (see Sect. 6.8) for two categories: Ladies' Fashion and Men's Fashion.

[29] The following text is taken from Van Nierop et al. (2011).

Table 8.5 Parameter estimates for the multivariate Type-2 Tobit model [Eqs. (8.89) and (8.90), N = 4,572]

	Ladies' fashion		Men's fashion	
Variable	Purch. Inc. (α)	Amount (β)	Purch. Inc.(α)	Amount (β)
Intercept	−0.241*	−0.238**	−0.973***	−0.917***
Number of category pages visited[a]	−0.441***	−0.105**	−0.550***	−0.078
Dummy user/non-user[b]	0.171*	0.222***	0.099	0.139
Holiday shopping season dummy	−0.022	−0.027	−0.023	−0.019
Fashion promotion dummy	0.183***	0.166***	0.230***	0.233***
General promotion dummy	0.147**	0.142**	0.251***	0.249***
Competitor 1 starts webstore	−0.195***	−0.199***	−0.157**	−0.156**
Competitor 1 major TV advertisement	−0.194	−0.188	−0.363**	−0.351**
Competitor 1 advertisement in 2001	−0.091**	−0.086**	−0.002	−0.002
Competitor 2 introduces magazine	−0.010	0.008	0.003	−0.003
(Log) months since introduction	0.077	0.065	0.013	0.007
Distance to closest store in miles	−0.017*	−0.011**	−0.009	−0.003
Gender (0 = male, 1 = female)	0.084	0.130**	−0.313***	−0.245***
Higher education (0 = no, 1 = yes)	−0.080	−0.057	0.072	0.058

[a]Because this is a category-specific model, a category-specific number of website visits is included that replaces the overall number of website visits in Table 6.8

[b] This variable equals 0 before the introduction of the international website: after the introduction it equals 1 for non-users and 0 for registered users (compare Table 6.8)

*Zero is not contained in the 90 % HDP (highest posterior density) interval; **zero is not contained in the 95 % HDP interval; ***zero is not contained in the 99 % HDP interval

Source: Van Nierop et al. (2011, p. 161)

The performance of the Purchase Incidence stage of the Type-2 Tobit model is computed by a Hit Rate that indicates the percentage of observations for which the model correctly classifies purchase incidence (see Sect. 8.2.2). In the estimation sample a Hit Rate of 72 % is achieved, for a hold-out sample this value equals 66 %. Across (six) categories a pseudo R^2 of 0.77 is found. From Table 8.5 two main conclusions can be drawn:

1. Website visits significantly decrease purchase incidence for the two categories. The partial relation between website visits and the average amount of money spent indicates a negative sign for ladies' fashion.
2. The parameters associated with the dummy for non-users after the introduction of the website does not change and even increases in the category ladies' fashion.

For other conclusions we refer to Van Nierop et al. (2011). With the multivariate Type-2 Tobit model it is also possible to investigate which product categories are correlated. The coefficients that indicate contemporary correlation suggest cross-category or co-occurrence effects (Manchanda et al. 1999). Van Nierop and colleagues encountered co-occurrence effects for the Ladies' Fashion category with, among others, Men's Fashion. These categories are complementary.

References

Abe, M.: "Counting your customers" one by one: A hierarchical Bayes extension to the Pareto/NBD model. Mark. Sci. **28**, 541–553 (2009)

Ainslie, A., Rossi, P.E.: Similarities in choice behavior across product categories. Mark. Sci. **17**, 91–106 (1998)

Andrews, R.L., Currim, I.S.: Multi-stage purchase decision models: Accommodating response heterogeneity, common demand shocks, and endogeneity using disaggregate data. Int. J. Res. Mark. **26**, 197–206 (2009)

Bayus, B.L., Mehta, R.: A segmentation model for the targeted marketing of consumer durables. J. Mark. Res. **32**, 463–469 (1995)

Ben-Akiva, M., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. MIT, Cambridge (1985)

Blattberg, R.C., Kim, B.D., Neslin, S.A.: Database Marketing: Analyzing and Managing Customers. Springer Science, New York (2008)

Böckenholt, U.: Estimating latent distributions in recurrent choice data. Psychometrica **58**, 489–509 (1993a)

Böckenholt, U.: A latent class regression approach for the analysis of recurrent choice data. Br. J. Math. Stat. Psychol. **46**, 95–118 (1993b)

Bucklin, R.E., Gupta, S., Siddarth, S.: Determining segmentation in sales response across consumer purchase behaviors. J. Mark. Res. **35**, 189–197 (1998)

Bucklin, R.E., Sismeiro, C.: A model of web site browsing behavior estimated on clickstream data. J. Mark. Res. **40**, 249–267 (2003)

Carpenter, G.S., Lehmann, D.R.: A model of marketing mix, brand switching and competition. J. Mark. Res. **22**, 318–329 (1985)

Chandukala, S.R., Kim, J., Otter, T., Rossi, P.E., Allenby, G.M.: Choice models in marketing: Economic assumptions, challenges and trends. Found. Trends Mark. **2**(2), 97–184 (2007)

Chang, K., Siddarth, S., Weinberg, C.B.: The impact of heterogeneity in purchase timing and price responsiveness estimates of sticker shock effects. Mark. Sci. **18**, 178–192 (1999)

Chatfield, C., Ehrenberg, A.S.C., Goodhardt, G.J.: Progress on a simplified model of stationary purchasing behavior. J. R. Stat. Soc. **79**, 317–367 (1966)

Chen, Y., Steckel, J.H.: Modeling credit card share of wallet: Solving the incomplete information problem. J. Mark. Res. **49**, 655–669 (2012)

Chib, S., Seetharaman, P.B., Strijnev, A.: Model of brand choice with a no-purchase option calibrated to scanner-panel data. J. Mark. Res. **41**, 184–196 (2004)

Chintagunta, P.K., Jain, D.C., Vilcassim, N.J.: Investigating heterogeneity in brand preferences in logit models for panel data. J. Mark. Res. **28**, 417–428 (1991)

Colombo, R.A., Morisson, D.G.: A brand switching model with implications for marketing strategies. Mark. Sci. **8**, 89–99 (1989)

Cox, D.R.: Partial likelihood. Biometrika **62**, 269–276 (1975)

Currim, I.S.: Predictive testing of consumer choice models not subject to independence of irrelevant alternatives. J. Mark. Res. **19**, 208–222 (1982)

Daganzo, C.: Multinomial Probit. Academic, New York (1979)

Danaher, P.J., Dagger, T.S.: Comparing the relative effectiveness of advertising channels: A case study of a multimedia blitz campaign. J. Mark. Res. **50**, 517–534 (2013)

Dekimpe, M.G., Parker, P.M., Sarvary, M.: Global diffusion of technological innovations: A coupled-hazard approach. J. Mark. Res. **37**, 47–59 (2000)

Du, R.Y., Kamakura, W.A.: Where did all that money go? Understanding how consumers allocate their consumption budget. J. Mark. **72**(6), 109–131 (2008)

East, R., Hammond, K.: The erosion of repeat-purchase loyalty. Mark. Lett. **7**, 163–171 (1996)

Ebbes, P., Grewal, R., DeSarbo, W.S.: Modeling strategic group dynamics: A hidden Markov approach. Quant. Mark. Econ. **8**(2), 241–274 (2010)

Ehrenberg, A.S.C.: The pattern of consumer purchases. Appl. Stat. **8**, 26–41 (1959)

Ehrenberg, A.S.C.: Repeat Buying, Theory and Applications. North-Holland, Amsterdam (1972)
Ehrenberg, A.S.C.: Repeat-Buying: Facts, Theory and Applications, 2nd edn. Oxford University Press, New York (1988)
Erdem, T.: A dynamic analysis of market structure based on panel data. Mark. Sci. **15**, 359–378 (1996)
Fader, P.S., Hardie, B.G.S., Lok Lee, K.: "Counting your customers" the easy way: An alternative to the Pareto/NBD model. Mark. Sci. **24**, 275–284 (2005)
Flinn, C., Heckman, J.: The likelihood function for the multi-state multi-episode model. In: Advances in Econometrics, vol. 2. JAI-Press, Greenwich (1983)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: Hierarchical versus other market share models for markets with many items. Int. J. Res. Mark. **14**, 359–378 (1997)
Fox, E.J., Montgomery, A.L., Lodish, L.M.: Consumer shopping and spending across retail formats. J. Bus. **77**(2), S25–S60 (2004)
Franses, P.H., Paap, R.: Quantitative Models in Marketing Research. Cambridge University Press, Cambridge (2001)
Freimer, M., Horsky, D.: Periodic advertising pulsing in a competitive market. Mark. Sci. **31**, 637–648 (2012)
Gensler, S., Dekimpe, M.G., Skiera, B.: Evaluating channel performance in multi-channel environments. J. Retail. Cons. Serv. **14**, 17–23 (2007)
Gilbride, T.J., Allenby, G.M.: Estimating heterogeneous EBA and economic screening rule choice models. Mark. Sci. **25**, 494–509 (2006)
Givon, M., Horsky, D.: Untangling the effects of purchase reinforcement and advertising carryover. Mark. Sci. **9**, 171–187 (1990)
Gönül, F.F., Srinivasan, K.: Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. Mark. Sci. **12**, 213–229 (1993)
Goodhardt, G.J., Ehrenberg, A.S., Chatfield, C.: The Dirichlet: A comprehensive model of buying behavior. J. R. Stat. Soc. A **147**, 621–655 (1984)
Greene, W.H.: Econometric Analysis, 7th international edn. Prentice-Hall, Upper Saddle River (2012)
Guadagni, P.M., Little, J.D.C.: A logit model of brand choice calibrated on scanner data. Mark. Sci. **2**, 203–238 (1983)
Gupta, S.: Impact of sales promotions on when, what and how much to buy. J. Mark. Res. **25**, 342–355 (1988)
Gupta, S.: Stochastic models of interpurchase time with time dependent covariates. J. Mark. Res. **28**, 1–15 (1991)
Heckman, J.J.: The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann. Econ. Soc. Meas. **5**(4), 475–492 (1976)
Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**(1), 153–161 (1979)
Helsen, K., Schmittlein, D.C.: Analysing duration times in marketing: Evidence for the effectiveness of hazard models. Mark. Sci. **11**, 395–414 (1993)
Horsky, D.: Market share response to advertising: An example of theory testing. J. Mark. Res. **14**, 10–21 (1977)
Hruschka, H., Fettes, W., Probst, M.: An empirical comparison of the validity of a neural net based multinomial logit choice model to alternative model specifications. Eur. J. Oper. Res. **159**(1), 166–180 (2004)
Jain, D.C., Vilcassim, N.J.: Investigating household purchase timing decisions: A conditional hazard function approach. Mark. Sci. **10**, 1–23 (1991)
Jeuland, A.P., Bass, F.M., Wright, G.P.: A multibrand stochastic model compounding heterogeneous Erlang timing and multinomial choice process. Oper. Res. **28**, 255–277 (1980)
Jones, J.M., Landwehr, J.T.: Removing heterogeneity bias from logit model estimation. Mark. Sci. **7**, 41–59 (1988)
Kamakura, W.A., Russell, G.J.: A probabilistic choice model for market segmentation and elasticity structure. J. Mark. Res. **26**, 379–390 (1989)

Kamakura, W.A., Srivastava, R.K.: Predicting choice shares under conditions of brand interdependence. J. Mark. Res. **21**, 420–434 (1984)
Kamakura, W.A., Srivastava, R.K.: An ideal point probabilistic choice model for heterogeneous preferences. Mark. Sci. **5**, 199–218 (1986)
Kumar, V., Sunder, S., Ramaseshan, B.: Analyzing the diffusion of global customer relationship management: A cross-regional modeling framework. J. Int. Mark. **19**, 23–39 (2011)
Leeflang, P.S.H.: Mathematical Models in Marketing, a Survey, the Stage of Development, Some Extensions and Applications. H.E. Stenfert Kroese, Leiden (1974)
Liechty, J., Pieters, F.G.M., Wedel, M.: Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. Psychometrika **68**(4), 519–541 (2003)
Louvière, J.J., Hensher, D.A.: Forecasting consumer demand for a unique cultural event: An approach based on an integration of probabilistic discrete choice models and experimental design data. J. Cons. Res. **10**, 348–361 (1983)
Louvière, J.J., Woodworth, G.: Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. J. Mark. Res. **20**, 350–367 (1983)
Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
Malhotra, N.K.: The use of linear logit models in marketing research. J. Mark. Res. **21**, 20–31 (1984)
Manchanda, P., Ansari, A., Gupta, S.: The 'shopping basket': A model for multicategory purchase incidence decisions. Mark. Sci. **18**, 95–114 (1999)
Massy, W.F., Montgomery, D.B., Morrison, D.G.: Stochastic Models of Buying Behavior. MIT, Cambridge (1970)
McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: P. Zarembarka (ed.) Frontiers in Econometrics. Academic, New York (1974)
McFadden, D.: Econometric models of probabilistic choice. In: C.. Manski, D. McFadden (eds.) Structural Analysis of Discrete Data with Econometric Applications. MIT, Cambridge (1981)
McFadden, D.: The choice theory approach to market research. Mark. Sci. **5**, 275–297 (1986)
Montgomery, A.L., Li, S., Srinivasan, K., Liechty, J.C.: Modeling online browsing and path analysis using clickstream data. Mark. Sci. **23**, 579–595 (2004)
Moon, S., Kamakura, W.A., Ledolter, J.: Estimating promotion response when competitive promotions are unobservable. J. Mark. Res. **44**, 503–515 (2007)
Morrison, D.G.: On the interpretation of discriminant analysis. J. Mark. Res. **6**, 156–163 (1969)
Morrison, D.G., Schmittlein, D.C.: Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? J. Bus. Econ. Stat. **6**, 145–159 (1988)
Netzer, O., Lattin, J., Srinivasan, V.: A hidden Markov model of customer relationship dynamics. Mark. Sci. **27**, 185–204 (2008)
Paas, L.J., Vermunt, J.K., Bijmolt, T.H.A.: Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. J. R. Stat. Soc. Ser. A **170**(4), 955–974 (2007)
Pauwels, K., Hanssens, D.M., Siddarth, S.: The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. J. Mark. Res. **39**, 421–439 (2002)
Prins, R., Verhoef, P.C.: Marketing communication drivers of adoption timing of a new e-service among existing customers. J. Mark. **71**(2), 169–183 (2007)
Prins, R., Verhoef, P.C., Franses, P.H.: The impact of adoption timing on new service usage and early disadoption. Int. J. Res. Mark. **26**(4), 304–313 (2009)
Punj, G.H., Staelin, R.: The choice process for graduate business schools. J. Mark. Res. **15**, 588–598 (1978)
Risselada, H., Verhoef, P.C., Bijmolt, T.H.A.: Dynamic effects of social influence and direct marketing on the adoption of high-technology products. J. Mark. **78**(2), 52–68 (2014)
Roberts, J.H., Lilien, G.L.: Explanatory and predictive models of consumer behaviour. In: J. Eliashberg, G.L. Lilien (eds.) Handbooks in Operations Research and Management Science Marketing, vol. 5. North-Holland, Amsterdam (1993)
Rooderkerk, R.P., Van Heerde, H.J., Bijmolt, T.H.A.: Incorporating context effects into a choice model. J. Mark. Res. **48**, 767–780 (2011)

Rossi, P.E., Allenby, G.M.: Bayesian statistics and marketing. Mark. Sci. **22**, 304–328 (2003)
Schmidt, P., Witte, A.D.: Predicting criminal recidivism using 'split population' survival time models. J. Econom. **40**(1), 141–159 (1989)
Schmittlein, D.C., Cooper, L.C., Morrison, D.G.: Truth in concentration in the land of (80/70) laws. Mark. Sci. **12**, 167–183 (1993)
Schwartz, E.M., Bradlow, E.T., Fader, P.S.: Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. Mark. Sci. **33**, 188–205 (2014)
Schweidel, D.A., Fader, P.S., Bradlow, E.T.: A bivariate timing model of customer acquisition and retention. Mark. Sci. **27**, 829–843 (2008)
Seetharaman, P.B.: Modeling multiple sources of state dependence in random utility models: A distributed lag approach. Mark. Sci. **23**(2), 263–271 (2004)
Seetharaman, P.B., Chib, S., Ainslie, A., Boatwright, P., Chan, T., Gupta, S., Mehta, N., Rao, V., Strijnev, A.: Models of multi-category choice behavior. Mark. Lett. **16**, 239–254 (2005)
Sichel, H.S.: Repeat buying and the generalized inverse Gaussian-Poisson distribution. Appl. Stat. **31**, 193–204 (1982)
Siddarth, S., Bucklin, R.E., Morrison, D.G.: Making the cut: Modeling and analyzing choice set restriction in scanner panel data. J. Mark. Res. **32**, 255–266 (1995)
Sikkel, D., Hoogendoorn, A.W.: Models for monthly penetrations with incomplete panel data. Statistica Neerlandica **49**, 378–391 (1995)
Sinha, R.K., Chandrashekaran, M.: A split hazard model for analysing the diffusion of innovations. J. Mark. Res. **24**, 116–127 (1992)
Song, I., Chintagunta, P.K.: A discrete-continuous model for multicategory purchase behavior of households. J. Mark. Res. **44**, 595–612 (2007)
Sridhar, S., Srinivasan, R.: Social influence effects in online product ratings. J. Mark. **76**(5), 70–88 (2012)
Uncles, M., Ehrenberg, A.S.C., Hammond, K.: Patterns of buyer behavior: Regularities, models and extensions. Mark. Sci. **14**, G71–G78 (1995)
Urban, G.L., Hauser, J.R.: Design and Marketing of New Products. Prentice-Hall, Englewood Cliffs (1980)
Van Nierop, J.E.M., Leeflang, P.S.H., Teerling, M.L., Huizingh, K.R.E.: The impact of the introduction and use of an informational website on offline customer buying behavior. Int. J. Res. Mark. **28**, 155–165 (2011)
Vanden Abeele, P., Gijsbrechts, E.: Modeling aggregate outcomes of heterogeneous non-IIA choice. In: EMAC 1991 Annual Conference Proceedings, Dublin (1991)
Verhoef, P.C., Franses, P.H., Hoekstra, J.C.: The impact of satisfaction and payment equity on cross-buying: A dynamic model for a multi-service provider. J. Retail. **77**, 359–378 (2001)
Vilcassim, N.J., Jain, D.C.: Modeling purchase-timing and brand-switching behavior incorporating explanatory variables and unobserved heterogeneity. J. Mark. Res. **28**, 29–41 (1991)
Vroegrijk, M., Gijsbrechts, E., Campo, K.: Close encounter with the hard discounter: A multiple-store shopping perspective on the impact of local hard-discounter entry. J. Mark. Res. **50**, 606–626 (2013)
Wedel, M., Kamakura, W.A., Arora, N., Bemmaor, A.C., Chiang, J., Elrod, T., Johnson, R., Lenk, P., Neslin, S.A., Poulsen, C.S.: Discrete and continuous representations of unobserved heterogeneity in choice modeling. Mark. Lett. **10**, 219–232 (1999)
Wedel, M., Kamakura, W.A., DeSarbo, W.S., Ter Hofstede, F.: Implications for asymmetry, nonproportionality, and heterogeneity in brand switching from piece-wise exponential mixture hazard models. J. Mark. Res. **32**, 457–463 (1995)
Wieringa, J.E., Verhoef, P.C.: Understanding customer switching behavior in a liberalizing service market: An exploratory study. J. Serv. Res. **10**, 174–186 (2007)
Wooldridge, J.M.: Introductory Econometrics: A Modern Approach, 5th edn. Cengage Learning, Mason (2012)
Wyner, G.: The practitioner's dilemma. Mark. News **47**(6), 14 (2013)
Zhang, J.Z., Netzer, O., Ansari, A.: Dynamic targeted pricing in B2B relationships. Mark. Sci. **33**, 317–337 (2014)

Zufryden, F.S.: A Logit-Markovian model of consumer purchase behaviour based on explanatory variables: Empirical evaluation and implications for decision making. Decis. Sci. **12**, 645–660 (1981)

Zufryden, F.S.: A general model for assessing new product marketing decisions and market performance. TIMS/Stud. Manag. Sci. **18**, 63–82 (1982)

Zufryden, F.S.: Multibrand transition probabilities as a function of explanatory variables: Estimation by a least squares approach. J. Mark. Res. **23**, 177–183 (1986)

Chapter 9
Examples of Database Marketing Models

9.1 Introduction

Database marketing and list management are more vital than ever, as marketers have troves of consumer information at their fingertips (*Marketing News*, October 2013, p. 54). Data are available in large quantities and originate from many different sources, as we already discussed in Chap. 3. Database management systems, also called, Customer Relationship Management (CRM) systems organize, at least in principle, all of this information.

Following Blattberg et al. (2008, p. 3) we define database marketing as:

> "The use of customer databases to enhance marketing productivity through more effective acquisition, retention, and development of customers".

Database marketing is the main driving force behind customer management. Customer management is a view and a strategy where the marketing efforts of the firm are organized through the clustering of customers into groups or "portfolios". Each "portfolio" is managed by a customer manager. The customer manager's goal is to increase, or even, to maximize value of the customers in her portfolio. This is called customer value management (Verhoef et al. 2007). Until recently, firms did not have the data management systems nor the statistical tools required to pursue customer management activities such as:

- selecting and acquiring customers;
- customer development including up- and cross-selling;
- retention/loyalty management, and
- chain management.[1]

[1] Blattberg et al. (2008, p. 55).

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_9

The classic view[2] is that the customer is exogenous to the firm and is the passive recipient of the firm's active value creation efforts, and values created in "the factory" (Deshpandé 1983). A different perspective is now emerging, namely, that customers can co-create value, co-create competitive strategy, collaborate in the firm's innovation process, and become endogenous to the firm. Central in this new view is the concept of *customer engagement*, defined as the behavioral manifestation from a customer toward a brand or a firm which goes beyond purchase behavior (Van Doorn et al. 2010). This behavioral manifestation may affect the brand or firm and its constituents in ways other than purchase such as Word-Of-Mouth (WOM), co-creation, referrals, participation in the firm's activities, suggestions for service improvements, customer voice, participation in brand communities, or revenge activities. As a consequence, the relation between a company and its customers becomes closer, more selective and may become so familiar that even the term intimacy is used (Treacy and Wiersema 1993).

Customer engagement is connected to customer value management through its objective, namely, to maximize the value of a firm's customer base. However, in customer value management, the value of a customer is generally linked to *direct* customer outcomes such as its current and future *transactions* with the firm. In contrast, customer engagement (additionally) includes behavioral manifestations of a customer with a rather *indirect* impact on firm performance.

We start to discuss some data issues which are inherent to database marketing in Sect. 9.2. We then discuss models that are used to determine the customer lifetime value (CLV) in Sect. 9.3. CLV plays a pivotal role in the models that support database marketing. Then we discuss:

- models for customer selection and acquisition (Sect. 9.4);
- models for customer development (Sect. 9.5);
- models for customer retention (Sect. 9.6);
- models for customer engagement (Sect. 9.7).

9.2 Data for Database Marketing

In this monograph we discuss data issues in Chaps. 3 and 10 and in this section. Here we discuss the data that are used for database marketing.

There are three fundamental reasons for companies to practice database marketing. According to Blattberg et al. (2008, p. 45) these are:

- enhancing marketing productivity;
- enabling the enhancement of customer relationships;
- establishing a sustainable competitive advantage.

[2]The following text is based on Bijmolt et al. (2010).

Development in information technology, the growth of the Internet, and the availability of large volumes of very detailed customer-level data (see the discussion on Big Data in Sect. 3.5.6), are reasons why many companies now invest in database marketing and database marketing systems/customer relationship marketing (CRM) systems. We now observe that many companies such as banks, insurance companies, telecom companies (O2), hotels (NH), credit card companies, retailers (Tesco in the UK), airlines, etc. are now active in database marketing.

Effective management of an organization's data that is coming from different sources is the foundation of the platform that will drive data-driven marketing. We distinguish at least three types of variables in a database[3]:

- demographics: name, address, phone number, age, family composition, income, cultural background;
- performance measures: sales, recency of sales (interpurchase time), frequency, amount of money spent, profit contribution, way of payment, channel used, length of the relationship, satisfaction measures, meanings and opinions, where all measures are defined of the individual consumer/firm-level;
- marketing efforts: calls/mailings etc. to the customer, price paid, discounts offered, rewards, customer contacts, use of loyalty card, etc.

The database may also contain other performance measures (see Table 10.4) and metrics (Table 10.5). Developments in data collection and data storage technologies mean that marketing databases have proliferated and grown in both size and complexity and new sources of data have emerged.

Large data sets create other challenges for the marketing scientist. The past 15 years have seen Bayesian statistical methods evolve from being a topic of intellectual curiosity to an essential component of any marketing scientist's toolkit (Rossi et al. 2005). Central to this shift was the development of Markov Chain Monte Carlo (MCMC) methods that involve thousands, if not million, of passes through the data.[4]

Next to the increased size of databases, an important issue is that organizations think more about operating in "real time". Hence, another data "reality" is that of processing streams of data in real time. While such real-time analysis is vital in telecommunications, financial services, and online settings, it may not be necessary for other customer analytics settings. There is a need to think about the "need for speed" in various analytical activities supporting management of customer engagement.

The same information technology developments that have lead to the massive growth in customer data have also lowered the costs of some traditional data sources while facilitating new data sources. For example, online surveys tools make it extremely easy for firms to survey their own customers. As a consequence, merging attitudinal data collected via surveys with transaction data (e.g., Kamakura et al. 2003) becomes even more important.

[3] See also Sect. 3.5.3.

[4] See Sect. 6.8.

Finally, a substantial part of customer behavior occurs in an online setting, resulting in new sources of data for studying customer engagement. For example, the emergence of social media forms an important development for the customer-firm relationship (Kaplan and Haenlein 2010).[5]

9.3 Modeling Customer Life Time Value

CLV is one of the cornerstones of database marketing. This metric quantifies the customer's long-term value to the firm. The CLV for a single customer is[6]:

$$CLV = \sum_{t=0}^{T} \frac{m_t \times r_t}{(1+i)^t} - AC \tag{9.1}$$

where

m_t = margin received from customer purchases at t,

i = discount rate/cost of capital for the firm,

r_t = probability of customer repeat buying or being "alive" at time t,

T = time horizon for estimating CLV, and

AC = acquisition cost of this "single customer".

The net contribution of the customer is determined by:

- the (expected) relationship duration, which is determined by T and r_t;
- the (expected) revenues generated by the customers in t (m_t);
- the (expected) acquisition costs (AC);
- the discount rate (i).

If $m_t = m$, $r_t = r$ and if we use an infinite time horizon, then (9.1) simplifies to (following the derivation of Gupta and Lehmann 2005, Appendix A):

$$\begin{aligned} CLV &= \sum_{t=0}^{\infty} \frac{m \times r^t}{(1+i)^t} - AC \\ &= \frac{m(1+i)}{(1+i-r)} - AC. \end{aligned} \tag{9.2}$$

[5]For a more in-depth discussion about the way data is organized in firms and more specifically the problems that inherent in collecting and storing data, we refer to Blattberg et al. (2008, Chapter 3); Bijmolt et al. (2010); Rust and Huang (2014).

[6]See Reinartz and Kumar (2003); Gupta et al. (2004); Blattberg et al. (2009).

Alternatively one might assume that

$$m_t = (1+g)^t m \tag{9.3}$$

where g is the growth rate, and

$$CLV = \sum_{t=0}^{\infty} \frac{(1+g)^t m \times r^t}{(1+i)^t} - AC. \tag{9.4}$$

Finally, we give an example where we determine CLV given that the firm has multiple products and accounts for the probability that the other products are sold to the "single customer". Then we express CLV (assuming a zero growth rate) as:

$$CLV = \sum_{t=0}^{T} \frac{(r)^t \left[m_t + cs_t mcs_t\right]}{(1+i)^t} \tag{9.5}$$

where

cs_t = the probability that there is cross-selling at t, and

mcs_t = the margin of the product that is cross-sold.

Models have been developed in which the components of CLV are explained by marketing efforts, demographics, purchase histories, etc.[7]:

- hazard, Markov, NBD, logit and probit models are used to model retention (Venkatesan and Kumar 2004);
- regression, logit and probit models have been developed to explain margins (Venkatesan and Kumar 2004);
- cross-selling has been modeled using logit and probit models (Verhoef et al. 2001).[8]

Also models have been developed that predict customer probability in future periods and where all relevant components are modeled in an advanced manner (Rust et al. 2011).

The CLV-metric must be aggregated at a higher level to be useful for managers. For this purpose one may focus on customer equity (CE) which is defined as the lifetime value of current and future customers.[9] To estimate the lifetime value of the

[7] Based on Gupta (2009).

[8] There are a number of surveys of studies to estimate CLV and its components. See: Donkers et al. (2007); Gupta and Lehmann (2008); Reinartz and Venkatesan (2008); Blattberg et al. (2009).

[9] Rust et al. (2004); Gupta and Lehmann (2005); Gupta and Zeithaml (2006).

entire customer base we consider cohorts of customers.[10] Each cohort of customers goes through the defection and profit pattern shown in Table 9.1. The firm acquires n_0 customers at time $t = 0$ at an acquisition cost of c_0 per customer. Over time, customers defect such that the firm is left with $n_0 r$ customers at the end of period 1, $n_0 r^2$ at the end of period 2, and so on.

Table 9.1 Number of customers and margins for each cohort

	Cohort 0		Cohort 1		Cohort 2	
Time	Customers	Margin	Customers	Margin	Customers	Margin
0	n_0	m_0				
1	$n_0 r$	m_1	n_1	m_0		
2	$n_0 r^2$	m_2	$n_1 r$	m_1	n_2	m_0
3	$n_0 r^3$	m_3	$n_1 r^2$	m_2	$n_2 r$	m_1
.	.	.	$n_1 r^3$	m_3	$n_2 r^2$	m_2
.	.	.	.	.	$n_2 r^3$	m_3
.	.	.	.	.	.	.

Source: Gupta and Lehmann (2008, p. 275)

The lifetime value of cohort 0 at time $t = 0$ is:

$$CLV_0 = n_0 \sum_{t=0}^{\infty} \frac{m_t r^t}{(1+i)^t} - n_0 c_0. \tag{9.6}$$

A similar expression can be derived for cohort 1. The lifetime value of cohort 1 at $t = 1$ is given by:

$$CLV_1 = n_1 \sum_{t=1}^{\infty} \frac{m_{t-1} r^{t-1}}{(1+i)^{t-1}} - n_1 c_1. \tag{9.7}$$

The present value of Eq. (9.7) at $t = 0$ equals:

$$CLV_1 = \frac{n_1}{(1+i)} \sum_{t=1}^{\infty} \frac{m_{t-1} r^{t-1}}{(1+i)^{t-1}} - \frac{n_1 c_1}{(1+i)}. \tag{9.8}$$

The customer equity value (CE) is then the sum of the lifetime value of all cohorts[11]:

$$CE = \sum_{k=0}^{\infty} \frac{n_k}{(1+i)^k} \sum_{t=k}^{\infty} \frac{m_{t-k} r^{t-k}}{(1+i)^{t-k}} - \sum_{k=0}^{\infty} \frac{n_k c_k}{(1+i)^k}. \tag{9.9}$$

[10] We closely follow Gupta and Lehmann (2008).

[11] See Gupta et al. (2004) for an empirical application.

CE can be determined, at least in principle, when the key inputs to this model (n_k, m_{t-k}, r, c_k and i) are estimated. Equation (9.9) can be modified taking time and cohort dependent retention rates into account. One may also account for heterogeneity (for all customers or the customers within a cohort).[12] Migration models acknowledge that customers might migrate in and out of being a customer during the course of their lifetime.[13] Here Markov and Hidden Markov models are used (see Chap. 8).

It has been demonstrated that there are close links between CE and firms' market value which emphasizes the importance of this concept. A discussion about these links is, however, beyond the scope of this monograph on "modeling markets".[14]

9.4 Models for Customer Selection and Acquisition

9.4.1 Models for Customer Selection

The initial goal of customer acquisition is to select the "right" prospects for the acquisition campaign.[15] Depending on the objective function a "right" prospect can be someone with maximum response likelihood, maximum purchase probabilities/levels or, as most in line with the customer relationship management (CRM) principles, maximum expected CLV.

Historically, the most frequently used selection technique has been the Recency, Frequency, and Monetary value (RFM) model. The core concept of the RFM model is based on empirical evidence and experiences in practice, particularly in Direct Marketing. Experience learned that customers that responded well in the past are likely to respond in the future. The three most important variables that summarize past consumer behavior are recency (R), frequency (F) and monetary amount (M). Building on the assumption that the "right" customer in the future looks a lot like the "right" customer in the past, the traditional RFM modeling approach creates groups of customers based on their RFM characteristics of prior purchases and then assigns probabilities or "scores" to each group in accordance with its differential response behavior. Marketing programs such as mailing campaigns are then prioritized based on the scores of different RFM groups (Gupta et al. 2006). Extensions of the RFM scoring approach define the customer groups using other behavioral (non-RFM) or sociodemographic variables.

Statistical models that are frequently used include Automatic Interaction Detection (AID) and Chi-square Automatic Interaction Detection (CHAID) selection

[12] See, for example, Blattberg et al. (2008, Chapters 5 and 6); Chan et al. (2011); Rust et al. (2011).

[13] Pfeifer and Carraway (2000).

[14] See Gupta and Zeithaml (2006); Wiesel et al. (2008); Blattberg et al. (2009); Schulze et al. (2012).

[15] We closely follow Bijmolt et al. (2010).

techniques (David Shepard Associates 1999), CART (classification and regression trees), parametric regression-based (scoring) models (Malthouse and Blattberg 2005), discriminant analysis, and log-linear models (LLM).

Columbo and Jiang (1999) developed a *stochastic* RFM model. The model is closely related to the models that we discussed in Chap. 8 [Eq.(8.25)]. The distribution of the number of responses (R) to m_i solicitations, given that customer i has a true response probability of π_i can be specified as:

$$P\left(R = r_i \mid m_i, \pi_i\right) = \binom{m_i}{r_i} \pi_i^{r_i} (1-\pi_i)^{m_i - r_i} \tag{9.10}$$

where r_i is the number of responses of customer i. The distribution of the response probabilities π_i is heterogeneous across customers, and Columbo and Jiang (1999) assume that the π_i's have a beta distribution [compare Sect. 8.2.3, Eq. (8.44)]:

$$f(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}. \tag{9.11}$$

The observed number of responses (r_i) to m_i solicitations can be shown to follow a Beta-Binomial distribution[16]:

$$P\left(R = r_i \mid m_i, \alpha, \beta\right) = \binom{m_i}{r_i} \frac{\Gamma(\alpha+\beta)\Gamma(m_i - r_i + \beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+m_i)}. \tag{9.12}$$

Once α and β are estimated, the expected response probability for customer i can be calculated as:

$$\mathrm{E}\left(\pi_i \mid r_i, m_i, \alpha, \beta\right) = \frac{\alpha + r_i}{\alpha + \beta + m_i}. \tag{9.13}$$

Equation (9.13) only requires the number of solicitations (m_i) and the number of responses (r_i). The selection at customers can be based on the estimated response probabilities for each consumer i.

Other models that are used to select "new" customers are the so-called probability models in which the probability that a customer will react to an invitation (a call or a mail) is related to individual's behavioral characteristics (Fader and Hardie 2009).

The models for customer selection discussed so far in this subsection are suited for "classical" RFM data, that is, purchase data collected in physical stores. However, one of the main challenges of models for customer acquisition is that the transaction history is not available for prospects. Hence, the researcher is left with less informative variables such as demographics and psychographics for profiling the top tier customers and identifying prospects that resemble these top tier customers. The models developed in the online marketing literature, for example,

[16] Compare Eq. (8.45).

Moe and Fader (2004), provide insights on how to deal with this challenge by exploiting clickstream data or data on other non-purchase behavior.

Just as purchase data can be collected in physical stores, it can also be collected in virtual stores, with the only difference that the data set of the virtual store, entails more information. In particular, in online shopping environments, it can not only be observed if, when, and what the customer purchased but also customer movements "through the store" can be tracked; that is, what items visitors looked at, how long they considered their decisions, in which sequence they bought items, and so on. The availability of these click-stream data has led to another stream in research focusing on the development of models that account for the richness of these data sets and allow scholars to obtain customer insights at an unprecedented level of granularity. Many of these models, at least implicitly, address customers' future purchasing probability. For example, Moe and Fader (2004) offer an individual-level probability model that predicts the visits that are likely to convert to purchases. Not only does their model control for different forms of customer heterogeneity, it also allows shopping behaviour to evolve over time as a function of prior experiences. Contrary to Moe and Fader (2004), Sismeiro and Bucklin (2004) take individual-level sequencing information into account and propose a sequential probit model that predicts online buying by linking the purchase decision to what visitors do and to what they are exposed to while browsing a particular website. Montgomery et al. (2004) propose a dynamic multivariate probit model that utilizes page-level movements through a website to predict of purchase conversion. They model clickstream data on a very disaggregate level, thereby improving the predictive power with regard to understanding which users are likely to make a purchase and which ones are not. Venkatesan et al. (2007) develop a selection model using CLV as the selection criterion. Their model consist of different parts:

1. a joint model for purchase timing and purchase quantity;
2. a model that rank-orders customers by their maximized CLV;
3. a selection model that selects customers until the budget constraint is exhausted.

The first model predicts the CLVs for customers. The second and third model use these predicted values to allocate the firm's resources.

9.4.2 Models for Customer Acquisition

Once the decisions regarding which customers to focus have been taken, the next question is how to allocate resources among marketing variables for leveraging customer acquisition. Firms use various types of marketing activities for customer acquisition, which differ according to the communication channel through which a prospect is acquired and the message that is used to attract the prospect (Reinartz and Venkatesan 2008). At the acquisition channel level, firms can acquire customers

directly (i.e., marketing induced customer acquisition) through one of the following channels: personal selling, mass media (e.g., radio and television), direct marketing channels (e.g., direct mail and telemarketing), Internet, and retail outlets (Bolton et al. 2004). At the same time, firms can use different messages (in terms of content and design) to attract different customers. For example, messages may contain brand-related information or price-related information.

Under the notion that different acquisition channels lead to different "qualities" of customers (Villanueva et al. 2008), researchers have modeled the effectiveness of different acquisition channels and have developed models to allocate the acquisition budget more efficiently. Most of the applied models are probability models, which incorporate covariates to explain variation in selected customer probability metrics. For example, Verhoef and Donkers (2005) use variants of probit models to explore how retention rates and cross-selling opportunities differ among the various acquisition channels that a financial services provider uses.

Content-related and design-related attributes of the acquisition message are other key elements of any marketing-induced acquisition campaign. Several researchers have modeled the effects of price discounts on various customer metrics and have provided models to improve price-related decision making in the context of customer acquisition management. For example, Anderson and Simester (2004) use a Poisson count model (Sect. 8.3) to conclude that customers acquired through catalogs with more discounted items have higher long-term value.

We discuss a model which has been developed by Reinartz et al. (2005) to balance acquisition and retention resources to maximize customer profitability. This normative model consists of three parts:

- an acquisition model;
- a relationship duration (retention) model, and
- a customer profitability model.

Reinartz et al. (2005) specify a so-called probit two-stage least squares model which consists of the following equations[17]:

$$z_i^* = \alpha_s' v_i + \mu_{is} \tag{9.14}$$

$$z_i = \begin{cases} 1 & \text{if } z_i^* > 0 \\ 0 & \text{if } z_i^* \leq 0 \end{cases} \tag{9.15}$$

$$y_{Di} = \begin{cases} \beta_{Ds}' x_{Di} + \varepsilon_{Dis} & \text{if } z_i = 1 \\ 0 \text{ otherwise.} \end{cases} \tag{9.16}$$

$$y_{Li} = \begin{cases} \beta_{Ls}' x_{Li} + \gamma_s' y_{Di} + \varepsilon_{Lis} & \text{if } z_i = 1 \\ 0 \text{ otherwise.} \end{cases} \tag{9.17}$$

[17] We closely follow Reinartz et al. (2005). A related model that studies acquisition and retention simultaneously has been developed by Schweidel et al. (2008).

where

z_i^* = a latent variable indicating customer i's utility to engage in a relationship with the firm,

z_i = an indicator variable showing whether customer i is acquired ($z_i = 1$) or not ($z_i = 0$),

v_i = a vector of covariates affecting the acquisition of customer i,

y_{Di} = the duration of customer i's relationship with the firm,

x_{Di} = a vector of covariates affecting the duration of customer i's relationship with the firm,

y_{Li} = the cumulative profitability of customer i,

x_{Li} = a vector of covariates affecting customer i's lifetime value,

$\alpha_s, \beta_{Ls}, \beta_{Ds}$ = segment-specific parameters, and

$\mu_{is}, \varepsilon_{Lis}$, and ε_{Dis} = error terms.

Equations (9.14) and (9.15) form the acquisition model. Equation (9.16) is the relationship duration model and (9.17) is the cumulative profitability equation. This is a recursive simultaneous system of equations model (compare Sect. 6.5). Equations (9.14) and (9.15) constitute a probit model (compare Sect. 8.2.2) that determines the selection/acquisition process. The duration (9.16) and profitability (9.17) are observed only if the customer is acquired. Equations (9.16) and (9.17) are conditional regressions determined partly by the acquisition likelihood of a customer. Because of the structure of the recursive system of relations (9.14)–(9.17) the model can be estimated in stages.[18]

Reinartz et al. (2005) calibrate the model using data from a large multinational B-to-B high-tech manufacturer. The customers are B-2-B and B-2-C firms. The products are durable goods. The marketing manager of this firm has the following information at her disposal: data of each purchase, number of proactive manufacturers, initiated marketing campaigns before that date, type of the campaign (telephone, face-to-face, e-mail, website). Explanatory variables in (9.14) the variables are different types of pre-acquisition *contacts*. Additional variables are the *amount* of acquisition *dollars* spent for each prospect. The standardized parameter estimates of the acquisition relation are shown in Table 9.2.

[18]The linkages among the three equations are captured by the error structure of this probit model two-stage least squares model. Unbiased estimates are obtained, the inverse Mills ratio is estimated and included in the duration model. Compare Sect. 8.5 (Footnote 24).

Table 9.2 Standardized parameter estimates of the acquisition relation (9.14)

Variable	Parameter estimates
Acquisition dollars	-0.559^{**}
Squared acq. dollars	-0.012^{*}
Telephone	0.298^{**}
Face-to-face	0.452^{**}
E-mail	0.271^{*}
Telephone × e-mail	0.086^{**}
Face-to-face × e-mail	0.052^{**}
Web	0.376^{**}
Industry type (B-to-B/B-to-C)	0.306^{*}
Annual revenue (of the customer)	0.414^{**}
Size of the demanding firm (employees)	0.370^{**}

$^{*}p < 0.10$; $^{**}p < 0.05$
Source: Reinartz et al. (2005, p. 71)

From Table 9.2 we conclude that:

- all communication models have a positive impact on acquisition, face-to-face communication having the greatest impact;
- there are positive synergies between telephone and e-mail and face-to-face and e-mail;
- as the firm increases the acquisition budget, the associated acquisition rate will be *less* responsive.

We return to the retention equation (9.16) in Sect. 9.6.

9.5 Models for Customer Development

CLV's can grow through many marketing activities.[19] This ultimately results in growth in sales and margins (profitability), cross-buying, and upgrading (up-selling). Many studies that propose models for customer development have appeared in recent years.[20] Venkatesan and Kumar (2004) use a regression model based on past transactions and marketing mix variables to predict the contribution margin. Using systems of equations within an extended service-profit chain framework, Bowman and Narayandas (2004) link customer management efforts to customer profitability.

[19] We closely follow Bijmolt et al. (2010).

[20] See Blattberg et al. (2008, Chapter 21); Reinartz and Venkatesan (2008); Bijmolt et al. (2010).

Another class of models investigates cross-selling (e.g., Kamakura 2008). Building on the idea that customers have predictable life cycles and, as a result, buy certain products before others, Li et al. (2005) model the demand for multiple products of the same provider (of banking services) over time. To this end, they use a multivariate probit model. Lemon and Von Wangenheim (2009) develop a "dynamic" model of cross-buying across loyalty program partnerships using data form European airlines. Reinartz et al. (2008) apply a causal model to investigate the direction of the relationship between cross-buying and behavioral loyalty. They find that purchasing items for multiple categories is a consequence of behavioral loyalty, and not an antecedent.

Other approaches to cross-selling comprise recommendation systems. The hidden Markov model of Netzer et al. (2008) that we discussed in Sect. 8.2.4 is another example of a customer development model. The model guides the firm's marketing decisions to alter the long-term buying behavior of its customers.

An important development for consumers is their increased opportunity to collect information and to order products from many channels. These channels include the Internet, call centers, sales forces, catalogs, retail stores, interactive television, and so on. Customers not only have more opportunities to contact firms, but the number of opportunities for home delivery increased also. Companies such as Peapod allow their customers to organize, in a customized fashion, their shopping behavior electronically already for some time.

A related topic that receives also much attention is the modeling of channel migration strategies: Hitt and Frei (2002) and Gensler et al. (2012, 2013) use matching methods to determine the effects of channel migration.

9.6 Models for Customer Retention

Customer retention focuses on preventing customer attrition or churn, that is, the termination of the contractual or non-contractual relationship between the customer and the company. Reward/loyalty programs support marketing decisions that may either lead to customer development and/or the reduction of churn.[21] We make a distinction between models that model the effects of reward/loyalty programs and churn prediction models.

9.6.1 Models to Support Loyalty/Reward Programmes

Consumer participation in Loyalty Programs (LPs, also known as loyalty card, rewards scheme, point card, advantage card or club card) continues to increase.[22]

[21] Surveys of research on loyalty programs are Bijmolt et al. (2010); Breugelmans et al. (2014).

[22] This text is based on Bijmolt et al. (2010).

Academic research on LPs has been substantial with numerous issues which seem not to abate over time. For example, the effectiveness of LPs remains a debated issue in the literature. While some studies show positive impact of LPs on customer behavior and firm performance (Liu 2007; Leenheer et al. 2007), other researchers question the effectiveness of LPs (Hartmann and Viard 2008; Shugan 2005).

In most models that determine the effectiveness of LPs, the number of points ("miles flown", "coins") are related to one or more components of customer's CLV. Examples are models by Kopalle and Neslin (2003); Lewis (2004) and Dorotic et al. (2014). Dorotic et al. (2014) determine the effects of LP rewards and LP-related promotions of a multi-vendor loyalty program (MVLP). In a MVLP cardholders receive promotions intended to increase sales of the multiple participating vendors. Dorotic et al. (2014) find low responsiveness of cardholders to loyalty program related promotions.[23]

We now return to Eq. (9.16), the "retention model". The duration of the relationship with the firm is influenced by the amount of money that is spent on retention (retention dollars), the contact channels that are used [as in Eq. (9.14)] and a number of other variables such as the frequency of transactions, the number of categories the customer buys in (cross-buying), the customer's share of wallet with the local firm and the relationship duration. The parameter estimates of the (censored) duration relations are shown in Table 9.3.

Table 9.3 Standardized parameter estimates of the relationship duration model (9.16)

Variable	Parameter estimates
Acquisition dollars	0.501**
Squared acquisition dollars	−0.101*
Telephone	0.328**
Face-to-face	0.381**
E-mail	0.152*
Telephone × e-mail	0.093**
Face-to-face × e-mail	0.077**
Web	0.386**
Frequency	0.417**
Squared frequency	−0.079*
Cross-buying	0.288**
Share-of-wallet	0.335**
Lambda	0.299**

*$p < 0.10$; **$p < 0.05$
Source: Reinartz et al. (2005, p. 71)

Reinartz and colleagues find that many of the communication channels have the same effect on duration as on acquisition. A remarkable difference is the *positive* but decreasing effect of retention dollars on the duration. The parameter lambda, which

[23] A similar study has been performed by Lemon and Von Wangenheim (2009). More specifically they model cross-buying across loyalty programs.

represents the inverse Mills ratio, (0.299) reveals that the duration of a relationship is correlated with the likelihood of acquiring a customer. This finding posits that a customer who is more likely to be acquired is also expected to engage in a longer relationship with the firm.

Reinartz et al. (2005) performed several simulations on the equations for acquisition, duration and customer profitability and derive an optimal resource allocation strategy.[24]

9.6.2 Churn Prediction Models

9.6.2.1 Introduction

Customer churn management[25] focuses on the retention component $r_t(r)$ in Eq. (9.1). Churn (c) is one minus the retention rate: $c = 1 - r$. At the customer level churn refers to the probability the customer leaves the firm in a given period. At the firm level, churn is the percentage of the firm's customer base that leaves in a given time period.

Several studies have investigated the churn or retention drivers in order to provide companies tools on how to improve the effectiveness of retention programs and hereby prolong the lifetime of customers.[26] For instance, Verhoef (2003) found affective commitment and loyalty programs to reduce churn.

In a non-contractual setting, the challenge is to infer whether a customer is still active or not. Most models that have been developed to assess the probability that a customer is still alive are probability models, such as the Pareto/NBD model (Schmittlein et al. 1987). In contractual settings (e.g., mobile phone subscriptions), customer churn is defined as the termination of the contract between the company and its customer. In this context, the churn problem is traditionally stated as a binary issue, where the aim is to predict whether or not a customer is likely to defect during a pre-given time period. Neslin et al. (2006) provide an overview of the binary models that were used by several academics and practitioners in the context of the churn modeling tournament. Various binary choice models have been used in the past. They include logistic regression analysis (Risselada et al. 2010), decision trees, and discriminant analysis (see Kamakura et al. 2005 for a review). An alternative way to tackle customer churn is to model the duration of customer relationship with the firm. This stream of research uses hazard models (Sect. 8.4) to predict the probability of customer defection (Bolton 1998).

A different approach consists of considering customers defections as transient. The "always a share" retention model typically estimates transition probabilities of

[24] We neither discuss these results nor the parameter estimates of the profitability equation (9.17).

[25] Text based on Bijmolt et al. (2010).

[26] Overviews are given by Blattberg et al. (2008, Chapter 24) and Risselada et al. (2010).

customers being in a certain state, where customer defection is defined as one of these states. Transition probabilities are estimated using Markov models (Pfeifer and Carraway 2000).

There a couple of issues that are specific for churn prediction models:

- aggregation of predictions of different models;
- specific validation criteria.

9.6.2.2 Aggregation[27]

Predictions that are based on churn models are biased on in-period or one-period-ahead forecasts. The quality of these predictions depends heavily on the specific sample. Building churn prediction models is a time-consuming and therefore costly operation (Malthouse and Derenthal 2008). To obtain long-term churn predictions firms need a good prediction method. We discuss four ways to obtain "good" predictions. Risselada et al. (2010) study the so-called "staying power" of various churn prediction models. Staying power is defined as the predictive performance of a model in a number of (for example: three) periods after the estimation period. Risselada et al. (2010) examine logit models and classification trees, both with and without applying a bagging procedure. Bagging consists of averaging the results of multiple models that have each been estimated on a bootstrap sample from the original sample. Bootstrap samples are random samples of size n drawn with replacement. The number of original observations in the bootstrap samples is smaller than in the complete sample. Each sample leads to predictions and these predictions can be *averaged*. This is also called "aggregation". The intuition behind aggregating multiple model results is that the quality of a single predictor might depend heavily on the specific sample. Averaging predictions that vary substantially will result in more stable predictors. Risselada et al. (2010) test the models using customer data of two firms from different industries, namely an internet service provider (ISP) and insurance markets. The results show that the classification tree in combination with a bagging procedure outperforms the other three methods. It is shown that the ability to identify high risk customers of this model is similar for the in-period and one-period-ahead forecasts. However, for all methods the staying power is rather low, as the predictive performance deteriorates considerably within a few periods after the estimation period. This is due to the fact that both the parameters estimates change over time and the fact that the variables that are significant differ between periods. Their findings indicate that churn models should be adapted regularly.

9.6.2.3 Validation Criteria

A measure that is commonly used for churn models is the *top-decile lift* (TDL). The TDL is defined as the fraction of churners in the top-decile divided by the fraction of

[27] The following text is based on Risselada et al. (2010).

churners in the whole set. This measure represents the ability of a model to identify those customers that have a high churn probability, the so-called high risk customers. The TDL can be computed as follows:

- use a model to predict churn probabilities;
- rank all customers on these probabilities from high to low risk;
- divide customers in ten groups;
- group 1 consists of these customers that have highest predicted churn probabilities;
- TDL = (actual churn rate of group 1/overall churn rate)*100.

The second measure is the *Gini coefficient*, which does not only focus on the high-risk customers, but considers the performance of the model across all customers.

The Gini coefficient is based on the *cumulative lift curve*,[28] which graphs the cumulative percentage of customers ordered by predicted churn rate (high to low) on the horizontal axis against the observed cumulative percentage of churners on the vertical axis. The solid line in Fig. 9.1 provides an example of a lift curve.

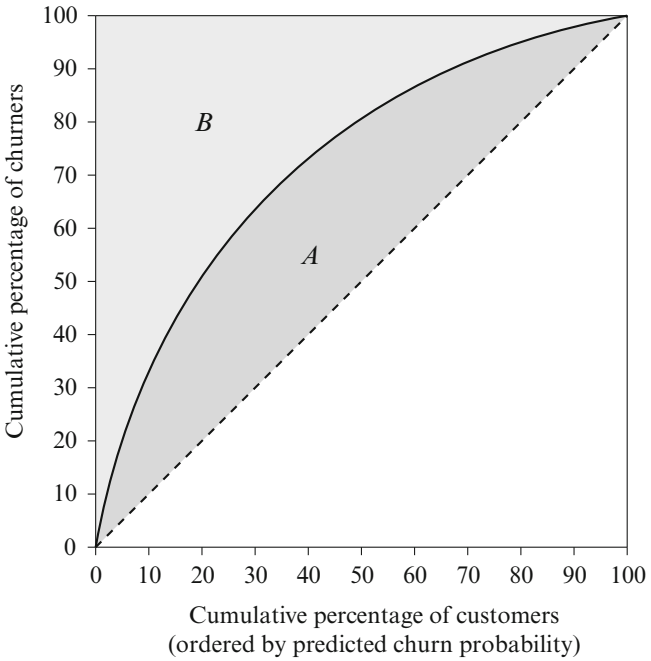

Fig. 9.1 Gini coefficient $(A/(A+B))$

[28] We closely follow Blattberg et al. (2008, pp. 329–320).

The lift curve is compared against the base case where every customer has the same churn probability, and churn predictions are based on random selection. The lift curve for the base case is expected to look like the dashed line in Fig. 9.1. The more the actual lift curve deviates from the base case, the better the churn model can identify churners and non-churners. For example, in Fig. 9.1, by focusing on the top 10 % high-risk customers, more than 30 % of the total number of churners can be identified. The Gini coefficient is defined as the area between the model's cumulative lift curve and the lift curve that would result from random prediction (area A in Fig. 9.1), divided by the theoretical maximum of this area, which is area A + area B in Fig. 9.1. By construction, the Gini coefficient can only attain values between zero and one, and a higher Gini coefficient indicates a better performance of the churn model.

The Gini coefficient can be calculated as:

$$\text{Gini coefficient} = \sum_{i=1}^{N} (c_i - \hat{c}_i)/(1 - \hat{c}_i) \tag{9.18}$$

where

$\hat{c}_i$ = the proportion of the customers who have a predicted churn probability equal or greater than customer i's,

c_i = the proportion of actual churners who are ranked equal or higher than customer i in their churn probability,

N = the number of customers.

Hence, $\hat{c}_i$ is the position of customer i on the dashed line in Fig. 9.1, and c_i is the position of customer i on the cumulative lift curve.

9.6.2.4 Application

Figure 9.2 shows the average top-decile lifts of the four different models which have been investigated in the study by Risselada et al. (2010). The results have been aggregated across estimation periods for the sake of clarity. The estimation period is denoted by t. The ability of the estimated models to correctly identify high risk customers is decreasing over time, since all lines are downwards sloping. A substantial decrease in period $t+2$ can be observed. Furthermore, the figure shows that the classification trees outperform the logit models in this respect, because both the line of the tree model and the line of the tree + bagging model are above the lines of the logit model. With respect to the effect of applying a bagging procedure, the following can be observed. The logit model does not benefit from this procedure, since both lines overlap in Fig. 9.2. However, the bagging procedure improves the

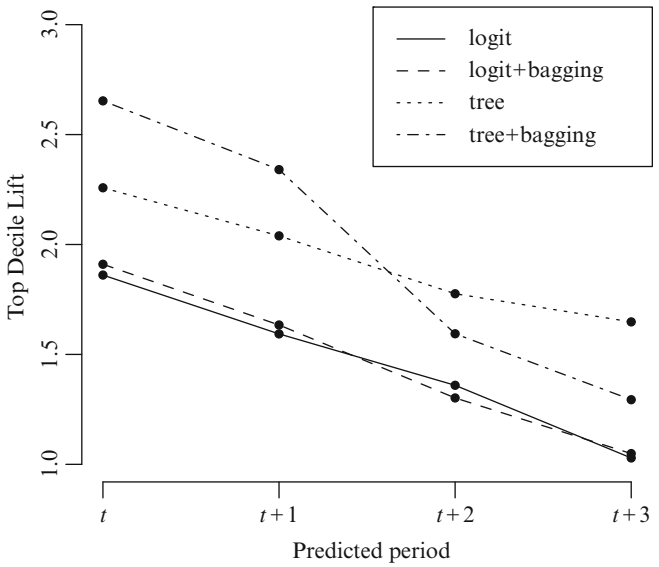

Fig. 9.2 Average top-decile lifts of models estimated at time t (ISP data). *Source*: Risselada et al. (2010, p. 203)

predictive performance of the classification tree substantially. Both for the in-period and the one-period-ahead predictions the TDL is higher for the tree in combination with a bagging procedure than for the single tree.

In Fig. 9.3, the average Gini coefficients of the four models are shown. Similar to what we found for the top-decile lift, the overall performance of all models decreases over time, indicated by the downwards sloping lines. Again, the tree models outperform the logit models and the bagging procedure improves the predictions of the classification trees but has little effect on the logit model results.

We now discuss the estimation results of one of the churn prediction models that have been estimated by Risselada et al. (2010). In Table 9.4 the parameter estimates of the single logit models are presented for each estimation period. The most important observation is that the significance and size of the parameter estimates change over time. Only 4 of the 25 variables (16 %) have a significant effect on churn in all periods. None of these four effects changes in sign. Customers with a higher revenue on their fixed phone line have a higher probability to churn on their internet description and those with the cheapest fixed phone subscription (*type 1*) have a higher churn probability than those with a more expensive subscription. Customers that used carrier pre-select in the past have a higher probability to churn and older people ($age \geq 65$) have a lower churn probability than young people. Three additional variables have a significant effect of the same sign in three of the four periods and four variables have a significant effect in the same direction in only two periods. There are five variables that have a significant effect only in Q1, where the sign of the effect mostly stays the same in the subsequent periods though

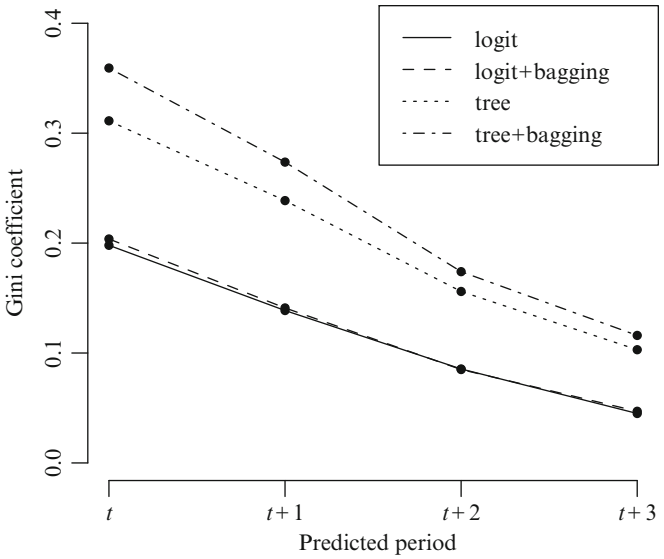

Fig. 9.3 Average Gini coefficient of models estimated at time *t* (ISP data). *Source*: Risselada et al. (2010, p. 204)

the effect is no longer significant. Finally, the effects of two variables, relationship age ISP and connection speed medium are significant in Q1 and Q3, but the sign of the effects is opposite in the two periods; in Q1 the effect is positive, in Q3 it is negative, which clearly indicates low parameter stability.

9.7 Models for Customer Engagement

9.7.1 Customer Engagement and Customer Management

In Sect. 9.1 we define customer engagement as the behavioral manifestation from a customer toward a brand or a firm which goes beyond purchase behavior (Van Doorn et al. 2010).[29] In this past decade much attention in marketing management has been directed to customer engagement and customer value management.

Neglecting behavioral manifestations of this kind can lead to a highly biased perception of a customer's contribution to a firm. For example, Von Wangenheim and Bayón (2007) find that the lack to incorporate WOM in the customer lifetime value (CLV) calculation could lead to an underestimation of the CLV by up to

[29]Parts of this section are based on Bijmolt et al. (2010).

Table 9.4 Parameter estimates of the single logit model (ISP data)

	Period			
Variable	Q1	Q2	Q3	Q4
Revenue fixed phone line (€)	0.1038**	0.1605**	0.1097**	0.1248**
Carrier pre-select	0.1608*	0.4869**	0.5716**	0.3507**
Revenue fixed phone line (€)	0.1038**	0.1605**	0.1097**	0.1248**
Relationship age company (months)	0.0000	−0.0007**	−0.0010**	−0.0012**
Relationship age ISP (months)	0.0084**	0.0009	−0.0019*	0.0001
Connection speed (ref. cat. "slow")				
Medium	0.7834**	0.0199	−0.5348**	−0.1809*
High	0.8992**	0.3996**	−0.1658	−0.1195
Fixed phone subscription (ref. cat. "standard")				
Type 1 (cheapest)	0.7715**	1.0344**	0.9207**	0.4580**
Type 3	−0.2412**	−0.1509*	−0.0198	0.0340
Type 4	−0.2957**	−0.0438	−0.1181	0.0846
Type 5	−0.4219**	−0.2906*	0.1256	0.1229
Household size (ref. cat. "3")				
1	−0.3177**	−0.1016	−0.1293	−0.1103
2	−0.1597*	−0.0426	−0.0191	−0.0756
4	0.0132	−0.0853	−0.1349	−0.0796
5	0.3009**	0.1406	−0.0562	0.0100
>6	0.1219	0.0628	−0.2390	−0.3527*
Age (ref. cat. "25–35")				
<25	0.0626	−0.0240	−0.1151	0.2874*
35–45	0.0909	0.0048	−0.1234	−0.112*
45–55	0.1169	0.0947	−0.0257	−0.0088
55–65	−0.2477**	−0.2320*	−0.1151	−0.1338
≥65	−0.4017**	−0.2286*	−0.3038**	−0.3023**
Income (ref. cat. "1.5 times standard")				
<Standard income	0.2101	0.2119*	0.3205**	0.4043**
Standard income	0.0757	0.1762*	0.1109	0.3189**
2 times standard income	0.0094	−0.1038	−0.0048	0.0173
>2 times standard	−0.2028**	−0.2498**	−0.2147**	−0.0876
Value added services fixed phone line	−0.0057	−0.1553*	−0.0476	−0.0603

* $p < 0.10$; ** $p < 0.05$
Source: Risselada et al. (2010, p. 204)

40 %. Thus, it seems essential to establish measures and models accounting for key behavioral manifestations of customer engagement.

Customer engagement may be generated in different stages of the customer life cycle: customer acquisition, customer development and customer retention (including churn). We discuss models that can be used as supporting tools for each at these stages.

9.7.2 Customer Engagement and Acquisition/Selection

At the heart of the models for customer selection and customer acquisition, discussed in Sect. 9.4, is an objective function used to discriminate among prospects who differ in terms of their response likelihood for a campaign or their purchase level. In the context of customer engagement, these rather purchase-related objectives can readily be replaced by behavioral manifestations other than purchase and be included in the "traditional" modeling approaches. For example, instead of predicting future purchase levels, analytical models may be aligned to predict the number of WOM referrals. Given the assumption that WOM communication positively affects revenues, firms may be interested in targeting customers with a high propensity to WOM. Bowman and Narayandas (2001) estimate two models for predicting WOM. First, using a logistic regression model, they determine whether a WOM referral is made. Then applying a truncated-at-zero NBD model, they estimate the actual number of referrals, given that at least one referral was made. A slightly more convenient approach for predicting WOM is offered by zero-inflation models allowing a joint estimation of the binary and the count model. For example, Von Wangenheim and Bayón (2007) use a zero-inflation Poisson (ZIP) model (Sect. 8.3.2) in which the standard Poisson model is complemented by a logit model. While the logit specification determines whether a referral is made, the Poisson count model subsequently predicts the number of referrals. An appealing side aspect of this model is that it allows for different sets of independent variables predicting the binary and the Poisson model.

At the acquisition channel level, firms not only acquire customers directly but also indirectly through referrals form the prospects' social network. Under the assumption that different acquisition channels lead to different "qualities" of customers (Lewis 2006), firms also need to understand in which way the fact that a customer acquired through WOM impacts its lifetime value. A suitable model for this issue has been proposed by Villanueva et al. (2008) who developed a Vector-Autoregressive model (VAR) model in order to capture WOM effects of a new customer acquisition on customer equity growth.[30]

From a strategic perspective understanding how the process of consumer adoption is affected by information communicated through mass media (external influence) and then spread via WOM (internal influence) has great importance for customer acquisition. Different research methodologies attempt to investigate the role and measurement of WOM:

1. There is a classic line of modeling based on Bass' (1969) well-known diffusion model. This classic line of research attempts to explain how marketing mix strategies affect new product diffusions (Mahajan et al. 2000) and shows that WOM effectively encourages people to start using a product. Usually, these models use aggregate data.

[30] VAR-models are discussed in Volume II.

2. The class of agent-based simulation models is a methodology that is especially useful when agent rules and characteristics can be defined on an individual level, when the population that adopts a new product is heterogeneous, and/or when the topology of the interactions between individuals is complex and heterogeneous. Examples of studies that use agent-based modeling are Goldenberg et al. (2009), Katona et al. (2011), and Van Eck et al. (2011).
3. Discrete choice models. See, for example, Yang et al. (2012) who developed a model to study consumer WOM generation and WOM consumption decisions simultaneously using individual data.
4. VAR models, discussed in Volume II.
5. Spatial models, discussed in Volume II.

The acquisition of new customers can also be realized by seeding strategies. In a seeding strategy one determines which kind of people one wants to sent a message. Should one "seed", "hubs" (well-connected people with a high number of connections to others), "fringes" (who are poorly connected), and "bridges" (who connect to otherwise unconnected parts of the network).[31] The choice of a seeding strategy determines the success of a viral marketing campaign.[32]

Godes and Mayzlin (2009) developed a model to find out what kind of WOM drives *sales*. Their model studies WOM for a restaurant chain which does business in 15 (regional) markets in the United States. The restaurant chain maintains a loyalty program centered around one of their five product line categories. BzzAgent is a marketing agency engaged in the business of creating WOM communication for its clients. Godes and Mayzlin (2009) designed and implemented a field test to determine the effects of WOM of loyal/non-loyal customers, of opinion/non-opinion leaders, of acquaintances/friends and relatives on sales. The field test, which lasted 13 weeks, involved a comparison of the WOM created by the members of the firm's loyalty program on the one hand, and the agency's panel (BzzAgent) who approached the non-members of the loyalty program on the other hand. Godes and Mayzlin (2009) estimated their model using data of the amount of WOM created by each agent over those 13 weeks and the recipients. Agents made reports and determined the type of relationship. Their (aggregate market-level) model has the following structure:

$$S_{it} = \sum_{j\in\{C,N\}} \left[\sum_{r\in R} w_j^r WOM_{ij,t-1}^r \right] + \sum_{i=2}^{15} \mu_i + \sum_{t=2}^{12} \tau_t + \varepsilon_{it} \tag{9.19}$$

where

S_{it} = sales of a certain category in market i in week t,

[31] See Trusov et al. (2010); Hinz et al. (2011).

[32] Van der Lans et al. (2010).

$WOM^{r}_{i\,j,t-1}$ = WOM the total number of reports filed in week $t-1$ in market i that reflect WOM from someone in condition j (customer or non-customer) to a person or people with whom they have a relationship that can be characterized by r,

$r \in R$ = friend, relative, acquaintance, stranger, other: the set of possible relationships between the sender and receiver of WOM information,

$j \in \{C, N\}$ = the customer condition of the sender where a distinction is made between a customer (C) and a non-customer (N),

μ_i = fixed effects for the regional market (where the restaurant chain operates), $i = 1, \ldots, 15$,

τ_t = fixed effects for the week, $t = 1, \ldots, 13$, and

ε_{it} = a disturbance term.

Godes and Mayzlin find that different forms of WOM have different effects on sales. So, for example, the WOM created by the agency (non-customers) has a significant and measurable effect on sales. These effects are higher than the effects of WOM created by customers.

Another area to account for customer engagement in the context of acquisition models is to predict consumers' willingness/ability to engage in co-creation activities for new product development (Fuchs et al. 2010). Since customers often vary highly in their willingness and ability to participate in co-creation tasks (Hoyer et al. 2010), firms are becoming increasingly interested in preselection mechanisms to identify segments of consumers who might be particularly willing and able to participate (Hoffman et al. 2010). Here, customer selection models incorporating major drivers of customer willingness and ability to co-create, such as scoring models can be applied. In addition, co-creation, especially in an online environment, is likely to produce large volumes of customer input that often requires a firm "screening millions of idea" (Hoyer et al. 2010). In order to overcome this problem, firms are interested in suitable preselection models. An example of such a selection model is a logistic regression model developed by Bayus (2013). Bayus studied 2 years of publicly available data form Dell's IdeaStorm community. Dell (computers) and Starbucks (coffee) have large communities that suggest, discuss and vote on thousands of new product and service ideas. Dell is outsourcing their ideation efforts in an attempt to get fresh ideas into their innovation process. Dell uses crowdsourcing:

> …"the act of taking a task once performed by an employee and outsourcing it to a large, undefined group of people external to the company in the form of an open call." (Bayus 2013, p. 226)

The selection and acquisition of customers is heavily influenced by online reviews, search (engine) advertising (Ghose and Yang 2009; Joo et al. 2014; Anderson and Simester (2014); Zenetti et al. 2014), brand posts on brand fan pages (De Vries et al. 2012), and blogs (Gopinath et al. 2013).

A number of models determine the effects of (online) product ratings on sales.[33] Research shows that positive online reviews are less valued than negative reviews. The selection/acquisition of customers based on online reviews is also influenced by social dynamics. Ratings behavior is significantly influenced by previously posted ratings. Sridhar and Srinivasan (2012) developed a (nested) ordered logit model in this respect. Moe and Trusov (2011) develop a rating model based on a hazard model. Ho-Dac et al. (2013) demonstrated that online customer reviews have no significant impact on the sales of strong brands. They discovered a positive feedback loop between (high) sales and the number of positive online customer reviews for weaker brands. Ho-Dac et al. (2013) estimated a three-equation model to account for endogeneity between the dependent variables such as sales and positive and negative online reviews.

9.7.3 Customer Engagement and Customer Development

Accounting for customer engagement in the development stage requires understanding how behavioral manifestations such as WOM, co-creation activities, and complaining behavior impact a CLV.

With regard to WOM, Goldenberg et al. (2007) explored the effects of individual and network-level negative WOM on profits using an agent-based model. They found that the effect of negative WOM on the Net Present Value (NPV) of the firm is substantial, even when the initial number of dissatisfied customers is relatively small. Trusov et al. (2009) study the effect of WOM on member growth at an Internet social networking site. The authors find that WOM-elasticities are approximately 20 times higher than that of marketing events and ten times than of media appearances.

Brand communities can create value among networked firm-facing actors, as such, the active management and stimulation to co-create is another important task. Bagozzi and Dholakia (2006) investigated the antecedents and purchase consequences of customer participation in brand communities. To disentangle the many variables that play a role in these interactions between community members, brands, and purchase they use Structural Equation Models[34] on the trade-off between the costs for a firm to stimulate consumer participation (e.g. financial rewards) and the benefits that firms receive (Hoyer et al. 2010).

A further manifestation of customer engagement, which is likely to affect CLV are customer complaints. On the one hand, firms recognize that complaints represent

[33] See for overviews: Sridhar and Srinivasan (2012); Chen and Lurie (2013).

[34] Structural Equation Models are discussed in Volume II.

an opportunity to remedy products or service-related problems and to positively influence subsequent customer behavior. There is considerable evidence that dealing effectively with complaints can have a dramatic impact on customers' evaluations of customer experiences as well as enhance their likelihood of repurchase and limit the spread of damaging negative WOM.

9.7.4 Customer Engagement and Retention

Traditional approaches to manage customer retention have focused on predicting which customers are most likely to churn and then target auctions to those customers to induce them to stay. However, they have generally mostly ignored the notion of customer engagement to the firm when making such decisions. Nitzan and Libai (2011) explore the role of customers' social network in their defection from a service provider. They find that exposure to a defecting neighbor is associated with an increase of 80 % in the defection hazard. Their findings are based on a (proportional) hazard model (see Sect. 8.4).

Other churn prediction models have been developed by Ascarza and Hardie (2013) and Haenlein (2013). Haenlein (2013) focuses on the importance of social interaction in the customer retention process within a social network. His study provides evidence for social interactions in customer churn decisions. Heanlein demonstrates that a focal actor is more likely to defect from a provider if other individuals to whom that actor is socially connected have previously defected from the provider.

9.8 Summary of Database Marketing Models

In this chapter, we provided key examples of database marketing models helpful to enhance customer acquisition, development, retention, engagement and marketing productivity. We explained the concept of customer lifetime value (CLV) as the cornerstone of database marketing and to demonstrate the value of marketing in increasing CLV and thus firm profits. Future research may adapt database marketing to further sustainable competitive advantage. With the increase in Big Data on customer-specific actions and reactions, we expect database marketing models to continue to thrive.

References

Anderson, E.T., Simester, D.I.: Long-run effects of promotion depth on new versus established customers: Three field studies. Mark. Sci. **23**, 4–20 (2004)

Anderson, E.T., Simester, D.I.: Reviews without a purchase: Low ratings, loyal customers, and deception. J. Mark. Res. **51**, 249–269 (2014)
Ascarza, E., Hardie, B.G.S.: A joint model of usage and churn in contractual settings. Mark. Sci. **32**, 570–590 (2013)
Bagozzi, R.P., Dholakia, U.M.: Open source software user communities: A study of participation in Linux user groups. Manag. Sci. **52**, 345–362 (2006)
Bass, F.M.: A new product growth model for consumer durables. Manag. Sci. **15**, 215–227 (1969)
Bayus, B.L.: Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. Manag. Sci. **59**, 226–244 (2013)
Bijmolt, T.H.A., Dorotic, M., Verhoef, P.C.: Loyalty programs: Generalizations on their adoption, effectiveness and design. Found. Trends Mark. **5**(4), 197–258 (2010)
Bijmolt, T.H.A., Leeflang, P.S.H., Block, F., Eisenbeiss, M., Hardie, B.G.S., Lemmens, A., Saffert, P.: Analytics for customer engagement. J. Serv. Res. **13**, 341–356 (2010)
Blattberg, R.C., Kim, B.D., Neslin, S.A.: Database Marketing: Analyzing and Managing Customers. Springer Science, New York (2008)
Blattberg, R.C., Malthouse, E.C., Neslin, S.A.: Customer lifetime value: Empirical generalizations and some conceptual questions. J. Interact. Mark. **23**, 157–168 (2009)
Bolton, R.N.: A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. Mark. Sci. **17**, 45–65 (1998)
Bolton, R.N., Lemon, K.N., Verhoef, P.C.: The theoretical underpinnings of customer asset management: A framework and propositions for future research. J. Acad. Mark. Sci. **32**, 271–292 (2004)
Bowman, D., Narayandas, D.: Managing customer-initiated contacts with manufacturers: The impact on share of category requirements and word-of-mouth behavior. J. Mark. Res. **38**, 281–297 (2001)
Bowman, D., Narayandas, D.: Linking customer management effort to customer profitability in business markets. J. Mark. Res. **41**, 433–447 (2004)
Breugelmans, E., Bijmolt, T.H.A., Zhang, J., Basso, L.J., Dorotic, M., Kopalle, P., Minnema, A., Mijnlieff, W.J., Wünderlich, N.V.: Advancing research on loyalty programs: A future research agenda. Mark. Lett. (2014, forthcoming)
Chan, T.Y., Wu, C., Xie, Y.: Measuring the lifetime value of customers acquired from Google search advertising. Mark. Sci. **30**, 837–850 (2011)
Chen, Z., Lurie, N.H.: Temporal contiguity and negativity bias in the impact of online word of mouth. J. Mark. Res. **50**, 463–476 (2013)
Columbo, R.A., Jiang, W.: A stochastic RFM model. J. Interact. Mark. **13**(3), 2–12 (1999)
David Shepard Associates: The New Direct Marketing: How to Implement A Profit-Driven Database Marketing Strategy, 3rd edn. McGraw-Hill, New York (1999)
De Vries, L., Gensler, S., Leeflang, P.S.H.: Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. J. Interact. Mark. **26**, 83–91 (2012)
Deshpandé, R.: Paradigms lost: On theory and method in research in marketing. J. Mark. **47**(4), 101–110 (1983)
Donkers, B., Verhoef, P.C., de Jong, M.G.: Modeling CLV: A test of competing models in the insurance industry. Quant. Mark. Econ. **5**(2), 163–190 (2007)
Dorotic, M., Verhoef, P.C., Fok, D., Bijmolt, T.H.A.: Reward redemption effects in a loyalty program when customers choose how much and when to redeem. Int. J. Res. Mark. (2014, forthcoming)
Fader, P.S., Hardie, B.G.S.: Probability models for customer-base analysis. J. Interact. Mark. **23**, 61–69 (2009)
Fuchs, C., Prandelli, E., Schreier, M.: The psychological effects of empowerment strategies on consumers' product demand. J. Mark. **74**(1), 65–79 (2010)
Gensler, S., Leeflang, P.S.H., Skiera, B.: Impact of online channel use on customer revenues and costs to serve: Considering product portfolios and self-selection. Int. J. Res. Mark. **29**, 192–201 (2012)

Gensler, S., Leeflang, P.S.H., Skiera, B.: Comparing the methods to separate treatment from self-selection effects in an online banking setting. J. Bus. Res. **66**(9), 1272–1278 (2013)
Ghose, A., Yang, S.: An empirical analysis of search engine advertising: Sponsored search in electronic markets. Manag. Sci. **55**, 1605–1622 (2009)
Godes, D., Mayzlin, D.: Firm-created word-of-mouth communications: Evidence from a field test. Mark. Sci. **28**, 721–739 (2009)
Goldenberg, J., Libai, B., Moldovan, S., Muller, E.: The NPV of bad news. Int. J. Res. Mark. **24**, 186–200 (2007)
Goldenberg, J., Lowengart, O., Shapira, D.: Zooming in: Self-emergence of movements in new product growth. Mark. Sci. **28**, 274–292 (2009)
Gopinath, S., Chintagunta, P.K., Venkataraman, S.: Blogs, advertising, and local-market movie box office performance. Manag. Sci. **59**, 2635–2654 (2013)
Gupta, S.: Customer-based valuation. J. Interact. Mark. **23**, 169–178 (2009)
Gupta, S., Hanssens, D.M., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., Sriram, S.: Modeling customer lifetime value. J. Serv. Res. **9**, 139–155 (2006)
Gupta, S., Lehmann, D.R.: Managing Customers as Investments. Wharton School Publishing, Philadelphia (2005)
Gupta, S., Lehmann, D.R.: Models of customer value. In: B. Wierenga (ed.) Handbook of Decision Models. Springer, New York (2008)
Gupta, S., Lehmann, D.R., Stuart, J.A.: Valuing customers. J. Mark. Res. **41**, 7–18 (2004)
Gupta, S., Zeithaml, V.: Customer metrics and their impact on financial performance. Mark. Sci. **25**, 718–739 (2006)
Haenlein, M.: Social interactions in customer churn decisions: The impact of relationship directionality. Int. J. Res. Mark. **30**, 236–248 (2013)
Hartmann, W.R., Viard, V.B.: Do frequency reward programs create switching costs? A dynamic structural analysis of demand in a reward program. Quant. Mark. Econ. **6**(2), 109–137 (2008)
Hinz, O., Skiera, B., Barrot, C., Becker, J.U.: Seeding strategies for viral marketing: An empirical comparison. J. Mark. **75**(6), 55–71 (2011)
Hitt, L.M., Frei, F.X.: Do better customers utilize electronic distribution channels? Manag. Sci. **48**, 732–748 (2002)
Ho-Dac, N.N., Carson, S.J., Moore, W.L.: The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? J. Mark. **77**(6), 37–53 (2013)
Hoffman, D.L., Kopalle, P.K., Novak, T.P.: The 'right' consumers for better concepts: Identifying and using consumers high in emergent nature to further develop new product concepts. J. Mark. Res. **47**, 854–865 (2010)
Hoyer, W.D., Chandy, R., Dorotic, M., Krafft, M., Singh, S.S.: Consumer co-creation in new product development. J. Serv. Res. **13**, 283–296 (2010)
Joo, M., Wilbur, K.C., Cowgill, B., Zhu, Y.: Television advertising and online search. Manag. Sci. **60**, 56–73 (2014)
Kamakura, W.A.: Cross-selling: Offering the right product to the right customer at the right time. J. Relat. Mark. **6**(3/4), 41–58 (2008)
Kamakura, W.A., Mela, C.F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P.C., Wedel, M., Wilcox, R.: Choice models and customer relationship management. Mark. Lett. **16**, 279–291 (2005)
Kamakura, W.A., Wedel, M., de Rosad, F., Mazzon, J.A.: Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. Int. J. Res. Mark. **20**, 45–65 (2003)
Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. Bus. Horizons **53**(1), 59–68 (2010)
Katona, Z., Zubcsek, P.P., Sarvary, M.: Network effects and personal influences: The diffusion of an online social network. J. Mark. Res. **48**, 425–443 (2011)
Kopalle, P.K., Neslin, S.A.: The economic viability of frequency reward programs in a strategic competitive environment. Rev. Mark. Sci. **1**(1), 1–39 (2003)

Leenheer, J., Van Heerde, H.J., Bijmolt, T.H.A., Smidts, A.: Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. Int. J. Res. Mark. **24**, 31–47 (2007)
Lemon, K., Von Wangenheim, F.: The reinforcing effects of loyalty program partnerships and core service usage: A longitudinal analysis. J. Serv. Res. **11**, 357–370 (2009)
Lewis, M.: The influence of loyalty programs and short-term promotions on customer retention. J. Mark. Res. **41**, 281–292 (2004)
Lewis, M.: Customer acquisition promotions and customer asset value. J. Mark. Res. **43**, 195–203 (2006)
Li, S., Sun, B., Wilcox, R.T.: Cross-selling sequentially ordered products: An application to consumer banking services. J. Mark. Res. **42**, 233–239 (2005)
Liu, Y.: The long-term impact of loyalty programs on consumer purchase behavior and loyalty. J. Mark. **71**(4), 19–35 (2007)
Mahajan, V., Muller, E., Wind, Y.: New-Product Diffusion Models. Kluwer, Boston (2000)
Malthouse, E.C., Blattberg, R.C.: Can we predict customer lifetime value? J. Interact. Mark. **19**, 2–16 (2005)
Malthouse, E.C., Derenthal, K.M.: Improving predictive scoring models through model aggregation. J. Interact. Mark. **22**, 51–68 (2008)
Moe, W.W., Fader, P.S.: Dynamic conversion behavior at e-commerce sites. Manag. Sci. **50**, 326–335 (2004)
Moe, W.W., Trusov, M.: The value of social dynamics in online product ratings forums. J. Mark. Res. **48**, 444–456 (2011)
Montgomery, A.L., Li, S., Srinivasan, K., Liechty, J.C.: Modeling online browsing and path analysis using clickstream data. Mark. Sci. **23**, 579–595 (2004)
Neslin, S.A., Gupta, S., Kamakura, W., Junxiang, L., Mason, C.H.: Defection detection: Measuring and understanding the predictive accuracy of customer churn models. J. Mark. Res. **43**, 204–211 (2006)
Netzer, O., Lattin, J., Srinivasan, V.: A hidden Markov model of customer relationship dynamics. Mark. Sci. **27**, 185–204 (2008)
Nitzan, I., Libai, B.: Social effects on customer retention. J. Mark. **75**(6), 24–38 (2011)
Pfeifer, P., Carraway, R.: Modeling customer relationships as Markov chains. J. Interact. Mark. **14**, 43–55 (2000)
Reinartz, W., Kumar, V.: The impact of customer relationship characteristics on profitable lifetime duration. J. Mark. **67**(January), 77–99 (2003)
Reinartz, W., Thomas, J.S., Bascoul, G.: Investigating cross-buying and customer loyalty. J. Interact. Mark. **22**, 5–20 (2008)
Reinartz, W., Thomas, J.S., Kumar, V.: Balancing acquisition and retention resources to maximize customer profitability. J. Mark. **69**(1), 63–79 (2005)
Reinartz, W.J., Venkatesan, R.: Decision models for customer relationship management (CRM). In: B. Wierenga (ed.) Handbook of Marketing Decision Models. Springer, New York (2008)
Risselada, H., Verhoef, P.C., Bijmolt, T.H.A.: Staying power of churn prediction models. J. Interact. Mark. **24**, 198–208 (2010)
Rossi, P.E., Allenby, G.M., McCulloch, R.: Bayesian Statistics and Marketing. Wiley, Hoboken (2005)
Rust, R.T., Huang, M.H.: The service revolution and the transformation of marketing science. Mark. Sci. **33**, 206–221 (2014)
Rust, R.T., Kumar, V., Venkatesan, R.: Will the frog change into a prince? Predicting future customer profitability. Int. J. Res. Mark. **28**, 281–294 (2011)
Rust, R.T., Lemon, K.N., Zeithaml, V.A.: Return on marketing: Using customer equity to focus marketing strategy. J. Mark. **68**(1), 109–127 (2004)
Schmittlein, D.C., Morrison, D.G., Colombo, R.: Counting your customers: Who are they and what will they do next? Manag. Sci. **33**, 1–24 (1987)
Schulze, C., Skiera, B., Wiesel, T.: Linking customer and financial metrics to shareholder value: The leverage effect in customer-based valuation. J. Mark. **76**(2), 17–32 (2012)

Schweidel, D.A., Fader, P.S., Bradlow, E.T.: A bivariate timing model of customer acquisition and retention. Mark. Sci. **27**, 829–843 (2008)
Shugan, S.M.: Comments on competitive responsiveness. Mark. Sci. **24**, 3–7 (2005)
Sismeiro, C., Bucklin, R.E.: Modeling purchase behavior at an e-commerce web site: A task-completion approach. J. Mark. Res. **41**, 306–323 (2004)
Sridhar, S., Srinivasan, R.: Social influence effects in online product ratings. J. Mark. **76**(5), 70–88 (2012)
Treacy, M., Wiersema, F.: Customer intimacy and other value disciplines. Harv. Bus. Rev. **71**(January/February), 84–93 (1993)
Trusov, M., Bodapati, A.V., Bucklin, R.E.: Determining influential users in internet social networks. J. Mark. Res. **47**, 643–658 (2010)
Trusov, M., Bucklin, R.E., Pauwels, K.: Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. J. Mark. **73**(5), 90–102 (2009)
Van der Lans, R., Van Bruggen, G., Eliashberg, J., Wierenga, B.: A viral branching model for predicting the spread of electronic word of mouth. Mark. Sci. **29**(2), 348–365 (2010)
Van Doorn, J., Lemon, K., Mittal, V., Nass, S., Pick, D., Pimer, P., Verhoef, P.C.: Customer engagement behavior: Theoretical foundations and research directions. J. Serv. Res. **13**, 253–266 (2010)
Van Eck, P.S., Jager, W., Leeflang, P.S.H.: Opinion leaders' role in innovation diffusion: A simulation study. J. Prod. Innov. Manag. **28**, 187–203 (2011)
Venkatesan, R., Kumar, V.: A customer lifetime value framework for customer selection and resource allocation strategy. J. Mark. **68**(4), 106–125 (2004)
Venkatesan, R., Kumar, V., Bohling, T.: Optimal customer relationship management using Bayesian decision theory: An application for customer selection. J. Mark. Res. **44**, 579–594 (2007)
Verhoef, P.C.: Understanding the effect of customer relationship management efforts on customer retention and customer share development. J. Mark. **67**(4), 30–45 (2003)
Verhoef, P.C., Donkers, B.: The effects of acquisition channels on customer loyalty and cross-buying. J. Interact. Mark. **19**, 31–43 (2005)
Verhoef, P.C., Franses, P.H., Hoekstra, J.C.: The impact of satisfaction and payment equity on cross-buying: A dynamic model for a multi-service provider. J. Retail. **77**, 359–378 (2001)
Verhoef, P.C., Van Doorn, J., Dorotic, M.: Customer value management: An overview and research agenda. Mark.: J. Res. Manag. **2**, 51–68 (2007)
Villanueva, J., Yoo, S., Hanssens, D.M.: The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. J. Mark. Res. **45**, 48–59 (2008)
Von Wangenheim, F., Bayón, T.: The chain from customer satisfaction via word-of-mouth referrals to new customer acquisition. J. Acad. Mark. Sci. **35**, 233–249 (2007)
Wiesel, T., Skiera, B., Villanueva, J.: Customer equity: An integral part of financial reporting. J. Mark. **72**(2), 1–14 (2008)
Yang, S., Hu, M.M., Winer, R.S., Assael, H., Chen, X.: An empirical study of word-of-mouth generation and consumption. Mark. Sci. **31**, 952–963 (2012)
Zenetti, G., Bijmolt, T.H.A., Leeflang, P.S.H., Klapper, D.: Search engine advertising effectiveness in a multimedia campaign. Int. J. Electron. Comm. **18**(3), 7–38 (2014)

Chapter 10
Use: Implementation Issues

10.1 Introduction

In this chapter we discuss several issues that are related to the actual use of a model. We first examine the determinants of model implementation. We categorize the dimensions that contribute to the likelihood of implementation as follows:

- model-related dimensions;
- organization-related dimensions;
- implementation-strategy dimensions.

The likelihood of implementation depends on the *model* itself. In Chap. 2, we formulated criteria that a model should satisfy: a model must be simple, complete, adaptive and robust. If a model satisfies these criteria it has a good chance of being implemented. However, such model structure "requirements" are necessary but not sufficient. The model parameters must be estimated reliably, an aspect we examined in detail in Chaps. 4, 6 and 8. Model acceptance will also be determined by any validation history, in the sense that successful validation experience will positively contribute to acceptance (Chap. 5). Model structure, parameter estimation and model validation are all part of the validity component of the probability of success in implementation. Another model-related dimension is the cost-benefit trade-off. A model may do very well in the sense of being correct and complete and yet it has no chance of being implemented if the model benefits do not exceed the costs. In Sect. 10.2 we discuss model-related dimensions that affect model use.

The probability of model implementation also depends on a number of contingencies related to the *organization* for which the model building project is carried out: *organizational validity*. Organizational validity comprises personal, interpersonal and organizational factors. Personal factors include the characteristics of the user, interpersonal factors relate to the interface between model user and model builder, and organizational factors consist of characteristics of the organization, and its environment. We discuss organizational validity in Sect. 10.3.

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0_10

The basic strategy for implementation is the continuous model builder–model user interface, underlying a process view of model building as described in Sect. 1.5. Two aspects of the model-building process deserve further comment: model scope and the evolutionary nature of model building. We examine these aspects along with the model's ease of use, another element of *implementation strategy*, in Sect. 10.4.

Finally, we spend attention to marketing management support systems, dashboards and metrics in Sect. 10.5

10.2 Model Related Dimensions

10.2.1 Cost–Benefit Considerations

The development and use of models is justified if the (expected) benefits exceed the (expected) costs. For a firm that has adopted a model, this implies that profit with the model should be greater than without it. Of course this is not easily operationalized. *After estimation and validation*, we can quantify the benefits and make a comparison with the costs incurred (see Fig. 1.5).

A decision maker, in choosing a strategy which is based on model assumptions, weighs two major factors:

- the *benefits* to be derived from the strategy;
- the *cognitive effort* and *costs* associated with:
  - *building* a model, and
  - *using* that model.

The costs of model utilization are important for the model builder and the model user. If the perceived or actual costs of using a proposed model are excessive, it may be advisable to formulate a simpler model. The model builder should focus closely on the effort required for model use, because barriers to model use will lead to rejection of the model. Counterproductive use may occur if the model is more complex than needed. Model complexity by itself, however, does not have to deter the user. Most users do not need to see the model details and they often do not want to see the complexities. Rather, they want a model that provides output they can understand and relate to. This output should be valid and reliable, and if the model allows the user to play relevant "what if" games, the user should not have to confront the model complexities. This is one of the advantages of using dashboards, that are discussed in Sect. 10.5.

10.2.1.1 Benefits

In Sect. 1.4, we made a distinction between direct benefits and side benefits. Direct benefits are the improvements in decisions that result from model use. Side benefits are those generated from model use that were not intended or expected.

The quantification of direct benefits is difficult for a number of reasons.

- Often the benefits of a model are determined on the same data used for model testing and parameter estimation. Such a comparison between situations in which the model is used versus not used would be biased in favor of the model.[1]
- For normative models it is possible to compare marketing decisions based on model output with the decisions that would have been made in the absence of a model. However, for descriptive models no optimal decisions are implied. One could propose to compare the estimated parameters with what managerial judgments would have generated, but such a comparison does not quantify the (possible) benefits from model use. If an estimated demand model is used for decision making, then it should be possible to compare the quality of the decisions with what would have been done without the model. Obviously, these comparisons are not straightforward.
- In all instances in which comparisons are made, the implicit assumption is that the model represents reality. In a limited way this is a testable proposition. The other way to think of it is that the benefits cannot be determined until after decisions have been made (or could have been made) based on the model. Indeed, we advocate that this be done repeatedly so that the model's limitations get clarified.

Examples are found in the literature that demonstrate how benefits are determined for specific models.[2]

10.2.1.2 Costs

We consider the following cost components:

(a) initial development costs;
(b) maintenance costs;
(c) costs inherent to model use;
(d) costs of marketing data.

(a) *Initial Development Costs*

The cost of initial model development tends to be fixed. It is incurred once, when the project is first undertaken by the marketing science department or information

[1] See, for example, Doyle and Saunders (1990).

[2] See, for example, Bult and Wansbeek (1995); Kumar et al. (2009); Wiesel et al. (2011).

system group within the firm. Alternatively, an outside consultant may be asked to develop and test a model for a fixed fee (Lilien 2011). If a model already exists, and can be rented from a consulting firm, then the entire cost becomes variable in that it depends on, for example, the frequency of use.

(b) *Maintenance Costs*

Model development is not a one-shot event. Maintenance costs relate to updating the model, such as, changing its structure, updating the parameter estimates, etc. These costs will be partly fixed, partly variable, in the sense that the frequency of structural change will depend on use intensity as well as dynamics in the marketplace.

(c) *Costs Inherent to Model Use*

Costs inherent to model use are:

- managerial *time;*
- the *cognitive effort* required of the decision maker to interact with the model, and
- the *cognitive effort* required to process the information generated by the system.

The managerial time required for model development and model use needs to be considered as a cost. This time may be assessed in terms of the manager's salary. This time cost can be compared against the possible reduction in time that results from having the model do part of the job the manager used to do. For example, managers may spend less time on programmed and structured activities such as inventory management, media allocation, sales force management and control, and judging the consequences of alternative marketing programs. They will have more time available for unstructured activities such as the creation of entirely new marketing programs. This reallocation of time can result in important benefits.

(d) *Costs of Marketing Data*

The model builder has to specify a model that is theoretically appropriate and substantively meaningful. It also needs to be subjected to data. In order for the model to provide relevant output, the model builder needs data that fit the properties of the model. The data have to be collected or purchased.

10.2.2 Supply and Demand of Marketing Response Models

10.2.2.1 Introduction

In the past 30 years market response models have diffused in the practitioners' community.[3] Leading firms, especially in consumer goods and services, database marketing companies and traditional market research companies develop and use

[3]The text of this subsection is based on Hanssens et al. (2005).

increasingly sophisticated models and analyses. The successful implementation of models depends on data availability, the methodology used, and other characteristics. It appears, however, that sophistication in *model specification* and estimation are often not conducive to acceptance. Research on actual model use is scarce.[4]

Marketing practice commonly focuses on relatively simple approaches such as data splitting, cross-tabulations and/or univariate frequencies.

It appears that many models appearing in the academic literature have little relation to marketing practice. Such models often deal with specific problems, are more descriptive than prescriptive, and include complexities that reduce the chance of implementation in practice. There is a tension between the objectives of academic research and the needs of managers. Research tends to be favored for publication in the premier academic journals if it meets the standards of the academic community. Neither relevance to real-world problems nor likelihood of implementation is normally a critical consideration (Leeflang 2011).

In what follows we concentrate on marketing response models that may be used in marketing practice. We provide an overview of the demand and supply of these models in Table 10.1. We distinguish standardized models, empirical generalizations, models of the firm and idiosyncratic models at the supply side. Model outcomes are used by marketing managers, senior executives, researchers, public policy officials and attorneys. Most existing models are intended for the benefit of marketing managers.

Table 10.1 Demand and supply of market response models

	Supply			
Demand	Idiosyncratic models	Standardized models	Generalizations	Models of the firm
Marketing management	±	++	++	–
Senior executives	+	±	++	++
Public policy and litigation	++	–	–	–

++ highly relevant, + relevant, ± maybe relevant, – not relevant
The stated relevance originates from the authors
Source: Hanssens et al. (2005, p. 425)

Much of marketing decision making is of a repetitive or a tactical nature. For example, advertising expenditures, sales promotion budgets, shelf space allocations, prices, margins, etc. have to be determined for each period. The consideration of changes in decisions is facilitated by the development of ever more detailed databases, for example those developed by AC Nielsen, IRI (Information Resources Inc.), IMS Health (the leading global provider of market information to the pharmaceutical and health care industries), and GfK. The availability of these

[4] An excellent survey of applications of marketing science models is given in Lilien et al. (2013). See also Vriens (2012, Chapter 4).

databases also makes it easier to justify the use of econometric modelling (e.g. bimonthly audit data would not permit the estimation of deal effect curves).[5] Furthermore, the increasing frequency and amount of marketplace feedback also demands a systematic approach for data analysis.

10.2.2.2 Supply Side

Idiosyncratic Models

Idiosyncratic models are developed to tackle specific marketing problems. Most marketing models are idiosyncratic models. Examples are the customization of marketing efforts,[6] the role of social media,[7] banner advertising,[8] search engines,[9] weblogs,[10] and the impact of user-generated content[11] in consumer decision making processes. These are unique models for unique applications and are usually rather complex. As another example we refer to models that have been developed to overcome a product-harm crisis (Van Heerde et al. 2007; Cleeren et al. 2008; Cleeren et al. 2013).

Standardized Models

Standardized models have become important tools to improve the quality of tactical marketing decisions at functional levels such as *brand management*. We define a standardized model as a set of one or more relations where the mathematical form and the relevant variables are fixed. A variation consists of the use of subsets of relations as modules. This is attractive if the relevance of modules depends on, say, client factors. In a module-based approach, the structure of each module is fixed. Of course the estimated equations will often still vary somewhat between applications. For example, predictor variables can be deleted from the relations based on initial empirical results. Standardized models are calibrated with data obtained in a standardized way (audits, panels, surveys), covering standardized time periods. Outcomes are reported in a standardized format such as tables with predicted own-item sales indices for all possible combinations of display/feature and specific price points (SCAN*PRO, see Sects. 7.3.2.2 and 6.8.7) or predicted market

[5] Van Heerde et al. (2001).

[6] See Ansari and Mela (2003); Valenzuela et al. (2009).

[7] Stephen and Galak (2012); De Vries et al. (2014).

[8] Manchanda et al. (2006).

[9] Zenetti et al. (2014).

[10] Onishi and Manchanda (2012).

[11] See also the special issue of Marketing Science: vol. 31, no. 3.

shares for new products (ASSESSOR, Urban 1993). SCAN*PRO was developed by Nielsen based on clients' needs for quantified expressions of the impact of temporary price cuts. The availability of more detailed data (at the SKU level for many metropolitan areas) at more frequent intervals (weekly versus bimonthly) avoided many aggregation concerns that used to hamper model estimation, and at the same time mandated a different approach for managers to interpret market feedback. IRI created similar models (Abraham and Lodish 1990).

Wide applicability of these models is not possible without the availability of detailed data sets for many products and access to appropriate software and estimation methods. The model building exercise is often a compromise between a desire to have complete representations of marketplace phenomena and the need to have simple equations. The model builder and the model user must understand how results can be interpreted, what limitations pertain to the model, and in what manner the model can be extended to accommodate unique circumstances. To achieve implementation of model results, the structure of standardized models is often simple and robust. We provide a few examples of standardized models in Table 10.2.

One benefit of standardization is that both model builders and users can learn under which conditions the model fails so that the base model can be adjusted over time: evolutionary model building: see Sect. 10.4.2.

Table 10.2 Supply: examples of standardized models

Questions	Models	Suppliers
1 Sales effects	SCAN*PRO	AC Nielsen
	PROMOTION SCAN	IRI
	MICRO TEST (new product)	Research International
	BPTO (Brand Price Trade-off Analysis)	Research International
	MEDIA DRIVER	IRI
2 Interaction effects	SCAN*PRO	AC Nielsen
	PROMOTION SCAN	IRI
3 Competition	SCAN*PRO	AC Nielsen
4 Category demand	ASSORTMAN	AC Nielsen
5 Short versus long term effects	RANGE OPTIMIZER	Research International
	EQUITY ENGINE	Research International
	LT MEDIA DRIVER	IRI

Source: Hanssens et al. (2005, p. 425)

Generalizations

Managers benefit from having performance *benchmarks* relative to the competition. The use of benchmarks in market response is subject to the uncertainty inherent in parameter estimates. Empirical generalizations, derived from meta-analyses of

market response estimates, provide one basis for benchmarks. For example, extant research includes average price and advertising elasticities, and decompositions of sales effects resulting from temporary price cuts. The assumption is that brands, product categories and markets are comparable at a general level. However, the analyses also allow for systematic variation across brand/model settings in an identifiable manner. We discussed a number of these generalizations in Sect. 1.3.2.

Models of the Firm

The empowerment of consumers as active parties in the customization of products and services and the increasing focus on the creation of customer value bode well for the strategic importance of marketing.[12] This evolution also poses a new challenge for market response modellers. Recognition of the value of models in the boardroom requires a broader focus that includes the long-term sales impact, cross-functional relations, such as supply-chain effects that facilitate consideration of profit, and the impact on capital markets. Here we touch the marketing/finance interface.

After the start of the financial crisis in 2007 marketers are increasingly being challenged to measure and communicate the value created by their actions on shareholder value.[13] Such impact on firm valuation may occur through improving the

- magnitude;
- speed, and
- safety,

of cash flows.[14] Especially the demand regarding accountability requires marketing to look beyond its impact on top-line performance to its effects on firm value. Figure 10.1 presents a framework of how such links may be established through marketing effects on company cash flows and firm value.

In Fig. 10.1, we posit two process mechanisms that make explicit the contribution of marketing to firm value creation through the creation of market-based assets (see also, Joshi and Hanssens 2010). Specifically, we follow the framework of Srinivasan et al. (2011) in delineating two mechanisms whereby marketing affects firm value. First is an indirect route wherein the market-based assets enhance firm value indirectly through their effects on cash flows. Second is a direct route that considers the market-based asset serves as information. In this route, the marketing acts as a visible signal of the financial well-being of the firm, and directly influences firm value.

[12] See Sect. 9.7.

[13] This text is based on Srinivasan et al. (2011).

[14] Srivastava et al. (1998).

Cash Flow Framework: Marketing as an Asset

The efficient market hypothesis implies that stock prices reflect all known information about the firm's future earnings prospects (Fama 1970). For instance, investors may expect the firm to maintain its usual level of advertising and price promotions. Developments (in the form of unexpected changes) that positively affect future cash flows result in increases in stock price while those negatively affecting cash flows result in decreases. Srivastava et al. (1998) argue that shareholders are motivated by three distinct behaviors of cash flows: magnitude, speed and volatility. We discuss, in turn, how marketing may influence each component.

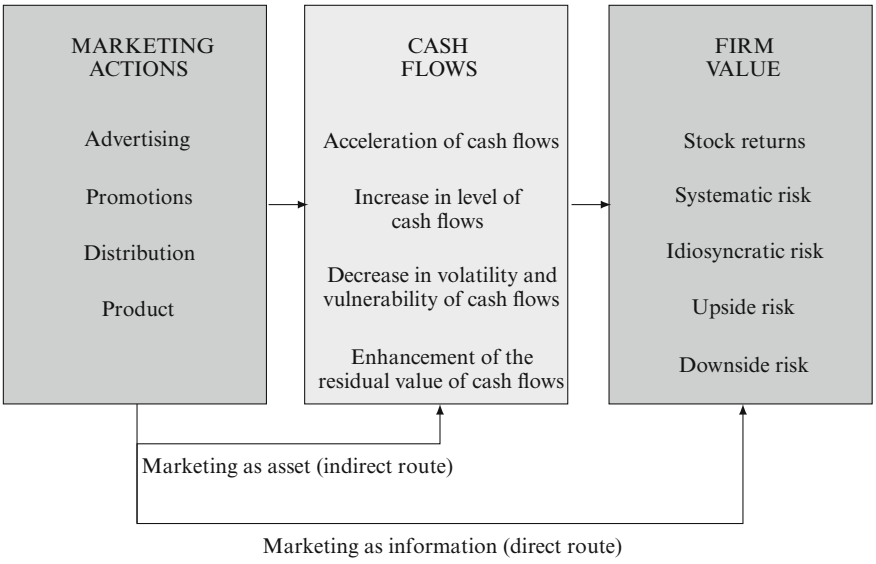

Fig. 10.1 Conceptual framework: the marketing finance interface. *Source*: Srinivasan et al. (2011, p. 89)

Magnitude: Enhancing cash flows. Marketing actions, which can involve substantial costs in the short run, can increase shareholder value by enhancing the level of *cash flows* (i.e., more cash), notably by increasing revenues and lowering costs. Most previous studies in marketing literature demonstrate this link; whether the marketing action is new product introductions, price promotions, new distribution/communication channels or advertising.

Speed: Accelerating cash flows. Marketing investments can enhance shareholder value by accelerating cash flows (i.e., faster cash). The acceleration of cash flows is affected by the faster speed of customer-value generating marketing initiatives relative to competition. For example, when a product extension announcement occurs, investors will decide to buy or sell company stock based on expectations of how the new extension will affect future cash flows (Lane and Jacobson 1995).

If investors value such extensions, this will result in increases in stock price. Apple's announcement of successive generations of the iPhone (3G, 3GS, 4G) resulted in acceleration of cash flows for Apple. There is limited empirical evidence (Pauwels et al. 2004; Srinivasan et al. 2009) thus far, that relate cash flow acceleration to shareholder value.

Safety: Lowering volatility of cash flows. Finally, marketing investments can increase shareholder value by lowering the vulnerability and volatility of these cash flows (i.e., safer cash), which lowers the firm's systematic risk and results in a lower cost of capital or discount rate (e.g., Srivastava et al. 1998; McAlister et al. 2007). Osinga et al. (2011)'s findings suggest that collective benefits of direct-to-consumer advertising (DTCA) expenditures for pharmaceuticals insulate a firm's stock from market downturns and thus lower its systematic risk. Thus, all else equal, cash flows that are predictable and stable have a higher net present value and thus create more shareholder wealth.

Signaling Framework: Marketing as Information

Research suggests that stock markets reflect an environment of information asymmetry between firms and investors (Myers and Majluf 1984). The signaling framework contends that economic information that is uniquely known by management (e.g., competitive viability) will be conveyed to shareholders through various signals, one of which is the customer-as-investor-signal. This is the second route by which marketing can affect firm value as noted by Srinivasan et al. (2011). Several studies in marketing (Joshi and Hanssens 2010; Osinga et al. 2011) use a signaling perspective to explain the effects of marketing on firm stock prices. Research suggests that firms with strong advertising are well-known and this reputation effect signals lower risks of the firm's stock to the investors (McAlister et al. 2007; Rego et al. 2009; Srinivasan et al. 2011; Xiong et al. 2013; Vitorino 2014).[15] Also, advertising increases demand for a firm's stock as it enhances the firm's salience for individual investors (Barber and Odean 2008). The finding of Osinga et al. (2011) of the positive effect of DTCA on returns is also consistent with the marketing as information route given that advertising expenditures grew rapidly after the DTCA regulation relaxation for pharmaceuticals, herewith enhancing the visibility of, and attention for, pharmaceutical firms. Furthermore, DTCA serves as an informational mechanism for individual investors, for example about new product launches and hence should enhance investor involvement with the company. Such involvement may cause individual investors to pay more attention to firm-specific news, such as clinical concerns, which would result in a stronger investor response to company news about stock returns.

[15]Other marketing efforts also have significant effects on stock prices/value of the firm. Examples are the timing of the introduction of innovations (Moorman et al. 2012) and brand quality (Bharadwaj et al. 2011); see Srinivasan and Hanssens (2009) for a survey.

Other Business Functions

The consideration of marketing models by senior executives requires that the work is closely aligned with other business functions, in particular with operations management. For example, response models have been used extensively to estimate the impact of sales promotions on sales. For a discussion of the challenges in the integration of marketing and operations management, see Karmarkar (1996) and the Special Issue on "Marketing and Operations Management Interfaces and Coordination" of *Management Science* 50(4) (2004).

10.2.2.3 Demand Side

In the previous section we already discussed the role of models for marketing managers and senior executives. Here we spend specific attention to the contributions models may have to solve *public policy problems* and to the resolution of *legal problems*.

Public Policy

Due to increased sophistication in marketing models and access to unique databases, academic researchers in marketing have an enhanced opportunity to contribute to the resolution of *policy* issues. The effect of advertising on consumer demand for products deemed to be harmful to social interests (public health, consumer safety, etc.) has received much attention. Examples of such products are tobacco[16] and alcoholic drinks. Models used by public-policy makers have been critically evaluated by marketing model builders. Such models may apply to several levels of demand. For example, brand sales models are used to estimate the effects of advertising bans, sales promotion bans, price regulations, etc. on brand sales. It is often particularly relevant to determine differences in effects between smaller or newer brands and market leaders. A recurring question in the litigation of advertising and tobacco products is whether manufacturers are accountable for damages in proportion to their brands' market shares or to their advertising spending levels.

Industry sales models have been used to estimate the effects of public policy decisions at a broader level. Relevant questions are:

- What is the effect of an advertising ban on total category consumer demand?
- What is the effect of a taxation-induced price increase on demand and on tax revenues?

[16] Capella et al. (2011).

These questions apply readily to tobacco products and alcoholic drinks, but also to gasoline and electricity markets. Examples of published market response models include Hu et al. (1995) for cigarettes, and Ornstein and Hanssens (1987) for distilled spirits and beer. Interaction effects such as the effect of advertising on price sensitivity are relevant to public policy questions as well. We discussed this issue extensively for pharmaceuticals in Sect. 7.3.2.3. There we discussed a study of the effect of marketing expenditures for prescription drugs on the price-elasticity of demand in the short- and the long-term. If marketing dampens the price elasticity, policy makers are inclined to believe that marketing should be restricted. However, if marketing improves the ability of a physician to identify the most suitable treatment for the disease of a particular patient, the price elasticity should also move toward zero.

Relatively little attention so far has been given to cross-category effects of an advertising ban. Duffy (1991) examined this by calibrating a model that accounts for interactions between beer, spirits, wine, and tobacco. He found a strong negative relation between tobacco and beer consumption. Thus, the reduced consumption of tobacco that might follow an advertising ban would be expected to result in a consumption increase in another "harmful" product category.

Recent time-series models focus on the effects of market shake-ups. Market shake-ups are events that are expected to change the market structure such as future values of relevant variables (sales, market shares) and/or relationships between variables (such as competitive reaction functions). Market shake-ups may appear at different levels of aggregation. A change in legislation (macro level), the acceptance of a new technology (industry level) or the entry of a radical new entrant (brand level) may have a permanent effect on the market structure. Time series models are useful to detect the long-run consequences of such events.[17]

The Federal Trade Commission (FTC) considers the impact on competition of potential mergers and acquisitions. For example, whether Nestle's acquisition of Dreyer's premium ice cream would have deleterious effects on competition was considered by economists in terms of whether premium ice cream constituted a separate market. Market-level scanner data were used to estimate own- and cross-brand price elasticities. However, since most of the price variation represents temporary discounting, adjustments should be made for cross-period (or stockpiling) effects.

Litigation

Market response models have been used in litigation for a long time. For example, in the 1960s, the Lydia Pinkham vegetable compound sales and advertising data became available for research due to a court case.[18] Rumors about a brand, catastrophic events and unfair competition are examples of factors that may have a

[17] Kornelis et al. (2008).

[18] Palda (1964).

negative effect on brand sales. Marketing mix models provide a basis for estimation of damage caused by these factors. In one case a broadcasting station spread improper news about a brand and was found liable by the court. Damage to the brand was to be estimated based on a marketing mix model. However, the model result revealed that increased competition and lower marketing investments were the critical factors that accounted for the decrease in brand sales. Thus, the news item had little negative effect on sales (Exota, Netherlands). In contrast, the well-publicized sudden acceleration rumor around the Audi 5000 in the US had a strongly negative sales impact, not only on sales of this product, but on the entire Audi brand.[19]

In another case, AKZO(-Nobel) (Germany) was alleged to have supplied harmful ingredients to a Unilever subsidiary for the production of the brand Iglo (frozen food, ice creams). Two people were poisoned after consuming Iglo's frozen food. Due to an incomplete recall procedure, two more people became sick from Iglo's products in a subsequent period. Hence the manufacturer of Iglo incurred two successive catastrophic events, both of which could affect the brand's sales, category sales (frozen food), sales of other Iglo products (cross category effects) and Iglo's brand equity. For the estimation of impact, no data on marketing variables were available so that time series models were applied to category sales, brand sales and market shares with two intervention dummy variables. The separation of the two events provided the basis for the allocation of the total damage between the supplier and the manufacturer.

10.3 Organizational Validity

In this section we discuss personal, interpersonal and organizational factors that are part of organizational validity. Research on model implementation, database systems and on computerized decision support systems (Sect. 10.5) has identified these factors as being influential.

10.3.1 Personal Factors

Models should ideally be custom built, and developed in accordance with the *integrative complexity* of the model user. *Integrative complexity* is the ability of an individual to integrate information on multiple dimensions in a complex fashion. Integratively more complex individuals can perform higher levels of information processing than integratively simple individuals. Thus, a model should be developed for a specific user in such a way that the model fits the manner in which the user makes decisions.

[19] Sullivan (1990).

Successful implementation also depends on "*user involvement*" and "*personal stake*".[20] Not surprisingly, greater *user involvement* leads to higher implementation success rates. User involvement is also central to the degree of interaction between developer (researcher) and user (manager). The more involvement between the two sides in terms of quantity and quality of interaction, the more likely it is that mutual confidence develops (see below).

Personal stake is the extent to which a model leads to better performance for the model user. If the model can be shown to improve the quality of a manager's decision, and if improved decision making leads to better performance, then the model is more likely to be implemented.

The personal characteristics of the decision maker that affect implementation success[21] include general intelligence, work experience, length of time at the company and in the job, education, personality, decision style and attitude toward a model.

10.3.2 Interpersonal Factors: The Model User–Model Builder Interface

Churchman and Schainblatt (1965) proposed the following matrix to represent four distinct *views* of the Model User (MU)–Model Builder (MB) interface.

Table 10.3 Model User (MU)–Model Builder (MB) interface

MU understands MB	MB understands MU	MB does not understand MU
	Mutual understanding	Communication
MU does not understand MB	Persuasion	Separate functions

Source: Churchman and Schainblatt (1965)

"MU understands MB" means that the MU reacts to what the MB is trying to do in a manner that improves the manager's chances of successfully exercising the duties assigned to the MU; see Table 10.3. We discuss distinct elements of these views below.

1. *Mutual Understanding*

This position represents an ideal set of characteristics. The model user understands the pros and cons of making decisions based on model output, and the model builder knows the various perspectives considered by the model user. The mutual understanding should lead to increased confidence about the outcome of the model-building process, and this will facilitate acceptance and use of the model.

[20] Hanssens et al. (2001, p. 327).

[21] See also Wierenga et al. (1997).

2. *Communication*

In this view, the model user directs the model builder. Essentially, the user has decided that a model should improve the quality of decisions, and the user has identified at least conceptually the structure to be used by the model builder. However, the model builder will not have a complete understanding of the user's perspectives. Of course, it is critical that the user has a good understanding of elements from statistics, econometrics and operations research. The model builder depends on extensive and thorough communication of the model user's needs and expectations.

3. *Persuasion*

This view applies when the model user is uninformed about the role of models in decision making. The problem of model implementation is then one of the model builder selling the features of the model. Among the drawbacks of this position is that model advocates often promise more than can be delivered. This excess in promise can show both in the model not fitting the decision making context and the model not performing at the promised level of accuracy.

The underlying rationale for this position is that model users (managers) are too busy to learn about models and do not have the patience to discuss details relevant to model development. It is then the task of the model builder to understand the manager well enough for the model to be accepted in principle and upon model completion based on the superiority of results. The persuasion task will require that the model builder understands the personality of the manager so that resistance to change can be overcome.

4. *Separate Functions*

In this view the functions of model builder and model user are essentially separate and separable. The model builder has the responsibility of generating a workable model. This model may be intended for use in a large number of settings. Once the model is completed, its purpose, its function, and its results will be presented to managers who either accept or reject its use. A modest amount of customization may be provided, dependent upon the heterogeneity in user needs, data characteristics, etc.

Based on several studies, it appears that the *mutual understanding* position is the most effective interface. This should be true especially if the entire model-building exercise takes place within an organization. The *communication* position may characterize situations in which managers oversee a group of individuals hired to provide model-building expertise. The last two positions represent cases where, for example, consultants or market researchers offer their services (Lilien 2011).

Mutual understanding and *communication*, are also characterized by a relatively high degree of user involvement. In the context of model building, user involvement depends on variables such as[22]:

- the user's understanding of model-building arguments and model characteristics;
- the user's evaluation of the quality of a model and its supporting mechanisms (people, hardware, data);
- the user's knowledge about approaches for conflict resolution (between participants in model development).

Factors that inhibit user involvement and prevent facile communication among the participants stem from fundamental differences between the personal and other characteristics of model builders and model users.

10.3.3 Organizational Factors

In Sects. 10.3.1 and 10.3.2 we considered organizational validity in terms of the behavior of the *members* of the organization involved in a model-building project. In this section we discuss implementation issues related to organizational structure and to relations between members who work at different levels in an organizational hierarchy.

The primary organizational factors that contribute to the likelihood of model implementation are:

- top management support;
- a match between model structure and the decision making structure;
- the position of the model builders in the firm.

Top management support refers to (top or divisional) management support, or lack of resistance, for models. The importance of management support stems from the *resistance to change* which is nearly always present when a model is to be implemented. This resistance is not just due to technical aspects of the model or lack of understanding. Managers may believe that models present a threat to their job. Implementation of a model may indeed reduce a manager's flexibility. The manager may feel that the model imposes a rigid structure and multiple constraints on her decision-making authority. The model may be seen as a formal mechanism to control the manager's performance. Thus instead of relying solely on one's judgment, the manager now may start with model output.

The second organizational factor is the *match between model structure* and *organizational structure*. An often-voiced criticism is that models are partial representations of reality. For example, advertising budgeting models generally miss explicit consideration of media allocation, and media planning models often

[22]See Hanssens et al. (2001, p. 327).

take the advertising budget as given. To see whether this is a serious pitfall, it is important to take the structure of the organization into account. If advertising budget decisions and media allocation decisions are made at different levels in the organizational hierarchy, then model implementation will be facilitated if the advertising budget model omits media allocation considerations. We elaborate on this point in Sect. 10.4.

The third organizational factor relates to the *position* of the *model builders* (or the information systems unit) *within the organization*. The following characteristics have been found to determine model acceptance: the size of the unit, its internal structure, the organizational and technical capabilities of the group, its reputation, the life cycle stage of the unit and the place of the unit in the organizational structure. We consider here only the last two points, the others being largely self-explanatory.

The marketing science (model-building) unit's life cycle stage affects implementation. In early stages, when strong organizational support is not yet present, implementation is more difficult than in later stages, when successful performance should create positive word-of-mouth and requests for other models from satisfied users.

In some firms marketing science projects are completed in the market research department. Alternatively, this department is responsible for the purchase of models from suppliers such as ACNielsen, IRI, GfK, Research International, etc. The latter situation pertains especially to smaller firms.[23] In those firms marketing science may be a one-person operation. A marketing scientist/model builder is likely to report to a product manager or marketing manager. Occasionally the model builder reports directly to the vice president for marketing. The involvement of higher levels in the organization occurs especially when models are developed on an ad-hoc basis to solve a specific problem. In many firms marketing research, analysis, and competitive intelligence are separate functions. Practitioners suggest to combine these functions and to adopt an insights approach (Vriens 2012, p. xxi). Decision makers are more and more asking for more and better actionable insights. Interpretation of the different functions could stimulate the generation of these insights.

10.4 Implementation Strategy Dimensions

10.4.1 Introduction

The proposed implementation strategy incorporates learning through evolutionary model building (see Sect. 10.4.2). We return to the topic of the relation between model structure and organization structure in Sect. 10.4.3 where we discuss issues

[23] See, for example, Wiesel et al. (2011).

dealing with *model scope*. We end this section by discussing how interactive computer systems can satisfy the *ease of use* requirement in Sect. 10.4.4.

10.4.2 Evolutionary Model Building

In Sect. 2.2.3 evolutionary model building was introduced as a means to overcome inconsistency between the model structure criteria of simplicity and completeness. Simplicity is desired so that managers understand models. We note that understanding does not imply that the user must understand the mathematical, statistical, and technical aspects of the model. Rather the user must understand what the model does and does not do.

As an implementation strategy, evolutionary model building increases the likelihood of model acceptance for several reasons. First, evolutionary model building implies continuous user involvement, which should lead to reduced resistance to change. Second, it leads to a communication pattern that is more favorable to model acceptance. Since the user must understand the model's strengths and weaknesses at a conceptual level, evolutionary model building represents a model user–model builder interface that corresponds to the "persuasion" position in the Churchman–Schainblatt framework (see Table 10.3). Third, there exists an optimal match between the environmental complexity of a model and the integrative complexity of the user. By adopting an evolutionary model-building approach the user can and will ask that the model's environmental complexity is consistent with the user's integrative complexity. This should lead to an optimal or near-optimal match.

The adjective "evolutionary" further implies that there are dynamics involved. That is, a model seen by a user as adequately representing the environment at one point in time may not receive the same favorable evaluation after learning has taken place. Experience with the model may lead the model builder and the user to the realization that some aspects should be added, and that other aspects should be changed or even deleted. Thus, the optimal match between model and model user may change over time as a result of learning. Thus, an implementation strategy should allow for corrective action. Learning will also be operative in overcoming resistance to change. As such, evolutionary model building is more than just an adaptation to a user's wishes. As an implementation strategy it may also be instrumental in attitude change.

We discussed the SCAN*PRO model developed by Wittink et al. (2011) in several sections. The original model is defined in Sect. 6.8.7. More complex and more realistic SCAN*PRO-versions were developed by Foekens et al. (1994) as part of an evolutionary model building process. The changes involve relaxation of the homogeneous parameters assumption. Chain-specific (heterogeneous parameters) were shown to provide better fit and better forecasting accuracy. Van Heerde et al. (2001) estimated a SCAN*PRO model by a semiparametric method to avoid inappropriate constraints on functional forms. This evolution is appealing in the

sense that a user may favor an initial version that is fully parametric because of its simplicity.

Other evolutionary versions of the SCAN*PRO model are:

- a SCAN*PRO model with varying parameters[24];
- a dynamic SCAN*PRO model with leads and lags.[25]

The leads and lags are critical so that the expanded model allows sales effects from promotions to be decomposed into, for example, brand switching, stockpiling and other effects.[26] We also mention a step in which competitive reaction- and feedback effects are added to the SCAN*PRO-model (Horváth et al. 2005). Finally, we mention the study in which the effects of including heterogeneity on the quality of the predictions based on the SCAN*PRO model are determined. This study is discussed in Sect. 6.8.7.

Thus far, evolutionary model building has been proposed as a process of gradually moving from a relatively simple representation of a given problem to a more complex representation. There may, however, also be an evolution in the types of problems being modeled. Urban (1974, p. 9) found that evolutionary progress often means not only model changes, but also the identification of new model needs. The chances of successful implementation are seriously impaired if the model building is very advanced and the user has little or no experience with models and model building. In that sense, early model-building efforts might concern well-structured problems, which are easy to systematize. After positive experience with such simple problems, the model building gradually evolves toward relatively unstructured problems.

10.4.3 Model Scope

In this section we first examine the global versus local model-building controversy. We then discuss general versus detailed descriptions of marketing variables and hierarchical linking (the linking of different decision points).

Model scope is seen as an element of implementation strategy because it relates to the matching of model structure to organizational structure.

[24] Foekens et al. (1999).

[25] Van Heerde et al. (2000).

[26] Van Heerde et al. (2004). The process of evolutionary model building is also worked out in detail for models that represent competitive actions and reactions in Leeflang (2008).

10.4.3.1 Global Versus Local Models

Should one build models for the entire firm (global) or for parts of the company (local)? This problem has multiple aspects. First, successful model building often starts at the lower levels in an organization. A major reason is that structured problems are more likely to be found at the operational levels in the organization, and as indicated before, structured problems lend themselves more easily to modeling. Also the appropriate level of aggregation is more easily determined for structured than for unstructured problems. In that sense, one might argue that trying to model a firm as a whole is a risky undertaking, and that it is better to build and use local models, such as a media selection model, a marketing mix model, or a new product model, and to link the (local) submodels over time. Links and feedbacks thus come into existence through a natural process rather than being forced onto an existing structure.[27]

The business world has embraced the notion that the functional areas of the firm, such as marketing and production and marketing and finance, should not act as independent units. Increasingly individual activities, nominally belonging to different functional areas, are coordinated and sometimes integrated. We discussed this already in Sect. 10.2.2.2.

10.4.3.2 General Versus Detailed Descriptors of Marketing Variables

For a brand sales model such as that described by Eq. (7.12), one might ask why the model does not include specific characteristics of the promotion (such as the content of the feature message, the quality of the display, the media used, etc.) or specific descriptors of distribution (such as the types of distribution outlets, the store characteristics, the store locations, etc.). Instead of specifying the marketing variables in *general* or broad terms (e.g. Gross Rating Points for advertising or percent of outlets in which the item is sold for distribution), one could decompose the variable into its constituent parts and use those components as separate predictor variables in the model.

Another approach is to specify models where the parameters are functions of factors that affect the parameter. This approach is known as the specification of varying parameter models or hierarchical model building (see Volume II).

10.4.4 Ease of Use

By ease of use we mean that the model is:

[27] See, for example, Fischer et al. (2011).

- easy to control, and
- easy to communicate with.

"Easy to control" means that the model should be constructed in such a way that it tends to behave as the manager would want it to. One interpretation of this is that ideally, if the manager believes that a price increase of 10 % will lead to a 20 % decrease in sales, then the model should predict just that. If the model produces predictions that vary wildly from the manager's experience or judgment, it is likely that the manager will reject the model. It is critical that the model is tested so that it not only provides acceptable predictions from the model builder's perspective but especially from the user's perspective. Of course, some education of the user will be needed if the user's intuition is demonstrably wrong.

The "easy to communicate with"-characteristic has many aspects. One is that model builders should be able to communicate their ideas in a manner that suits the model user. Thus, the model builder must have the user's perspective in mind to enhance the likelihood of successful model implementation.

Easy communication also means that it should be easy for the user to specify model inputs, and for the model to provide output quickly. It may be helpful to program the model in a worksheet that allows the model user to "play around" with relevant inputs to see how important output variables respond. Such worksheets can be quite effective in bringing about ease of use. They:

- reduce barriers between model and user;
- aid the learning process by immediate response;
- make immediate availability of information possible;
- encourage the user to examine a large number of possible plans;
- can be made available to a number of different users.

The degree of interaction afforded by a decision aid is often an important precursor to usage. And active participation in model development, including the construction of a frame for output measures by the user, tends to lead to increased commitment, acceptance, support and use.

10.5 Marketing Management Support Systems (MMSS), Dashboards and Metrics

10.5.1 Introduction

In this section we discuss three issues that have impact on the ease of use of models that support marketing decisions. In Chap. 9 we already discussed models that are used for customer management/database marketing. In Sect. 10.5.2 we discuss other 'standardized' *systems* which have been developed to support a number of more specific marketing decisions: Marketing Management Support Systems (MMSS).

Nowadays we observe that more and more companies use so-called *dashboards*. These dashboards replace the MMSS over time. Dashboards are discussed in Sect. 10.5.3. Finally, we spend attention to relevant *marketing metrics* in Sect. 10.5.4.

10.5.2 Marketing Management Support Systems (MMSS)

The use of electronic means to record purchases in supermarkets and other retail outlets has led to an exponential increase in the availability of data for modeling purposes.[28] McCann and Gallagher (1990, p. 10) suggest that the change from bimonthly store audit data about brands to weekly scanner data results in a 10,000 fold increase in available data (scanner data also measure more variables for a larger number of geographic areas at the UPC/SKU-level). Assume a brand group would normally spend five person-days analyzing one report based on bimonthly store audit data. Analyzing means here detecting important changes, explaining changes, discussing reasons for explanations, etc. This group would have to increase its size about 1,000 times or become 1,000 times more efficient (or some combination of these) in order to analyze weekly store-level scanner data in the same manner as the audit data.

Given the new "Big Data" (Sect. 3.5.6) that are now available the challenge to store and analyze data to support marketing decisions has become immense. To prevent this explosion in manpower, much of the analysis of more detailed and more frequently available scanner data has to become automated. One way to accomplish this is through the development and application of Marketing Management Support Systems (MMSS). The following types of "computerized Marketing Management Support Systems" can be distinguished:

- MarKeting Information System (MKIS);
- Marketing Decision Support System (MDSS);
- Marketing Knowledge-Based System (MKBS);
- Marketing Case-Based Reasoning systems (MCBR);
- Marketing Neural Nets (MNN).

A MKIS harnesses marketing-related information to facilitate its use within the firm (Lilien and Rangaswamy 2004, pp. 372–387). A MKIS consists of a database (or databank) with marketing data and statistical methods that can be used to analyze these data (statistical methods bank). The statistical analyses are used to transform the data into information for marketing decisions.

[28] We do not discuss the benefits from MMSS here. See, in that respect, Kayande et al. (2009), who demonstrate that Model-Based Decision Support Systems (MDSS) improve firm's performance in many contexts that are data rich, entail uncertainty, and require repetitive decisions. See also Becker et al. (2010); Wierenga (2011).

A MDSS differs from a MKIS in that it contains a model base (modelbank) in addition to a databank and a statistical methods bank. The purpose of a MDSS can be described as the collection, analysis, and presentation of information for immediate use in marketing decisions. MKIS and MDSS belong to the category of data-driven, quantitative MMSSs. Data-driven MMSS deal with structured marketing problems. Many MMSS are based on customer databases to enhance marketing productivity through more effective acquisition, retention and development of customers (see Chap. 9).

Other MMSS deal with specific marketing problems as will be clear from the list of well-known data driven MMSS below:

- ASSESSOR (Silk and Urban 1978), a measurement and modeling system designed to estimate the sales potential of new packaged goods before they are test-marketed;
- BRANDAID (Little 1975), a marketing mix model to assist board managers to compose their marketing mix;
- CALLPLAN (Lodish 1971)[29] a MMSS to allocate sales force effective and to support sales force decisions;
- TRACKER (Blattberg 1978), a model to predict the potential sales of new products;
- SCAN*PRO (Wittink et al. 2011, see also Chap. 7), a MMSS/model that can be used to evaluate sales promotions;
- SH.A.R.P (Bultez and Naert 1988), a supermarket shelf space allocation model;
- CHAN4CAST: a mutichannel, multiregion sales forecasting model and decision support system for consumer packaged goods (Divakar et al. 2005).

Many marketing decisions are weakly structured. Examples are decisions about a copy strategy, the planning of a sales promotion campaign, the way international marketing negotiations should be positioned etc. Knowledge driven MMSS are used to support these weakly structured marketing problems (Wierenga et al. 1997; Wierenga 2011). MKBS/MCBR are examples of knowledge driven MMBS.

A MKBS is a more advanced system than the previous two in the sense that theoretical knowledge and empirical generalizations are included. A restricted version of MKBS is an "expert system" which is related to the concept of "artificial intelligence" (AI). AI is concerned with the creation of computer programs that exhibit intelligent behavior. The program solves problems by applying knowledge and reasoning that mimics human problem solving. The expert system approach is one of the earliest techniques for the creation of intelligent programs.

Typically, an expert system focuses on a detailed description (model) of the problem-solving behavior of experts in a specific area. The expert system is a knowledge-driven system because it encompasses and uses knowledge from experts but also

[29] See also Sinha and Zoltners (2001) and Lilien and Rangaswamy (2004, pp. 373–378).

other knowledge sources such as information available from books and articles, empirical results and experience. Examples of such systems are[30]:

- PEP, a system designed to support the planning of consumer sales promotion campaigns (Bayer et al. 1988);
- ADCAD (ADvertising Communication Approach Design), a system to assist managers with the definition of marketing and advertising objectives, with advertising decisions regarding the positioning of a product/service, with the message, etc. (Burke et al. 1990);
- ADDUCE, a system that predicts how consumers will respond to advertisements from a description of ad, audience, brand and market characteristics (Burke 1991);
- BMA (Brand Manager's Assistant), a knowledge based system to assist brand management (McCann et al. 1991);
- NEGOTEX (NEGOTiation EXpert), a system that provides guidelines to individuals or teams preparing for international marketing negotiations (Rangaswamy et al. 1989).
- BRANDFAME: an expert system that assists brand managers (Wierenga et al. 2000; Wierenga et al. 2008)

A Marketing Case-Based Reasoning system (MCBR) is based on the fact that analogical reasoning is a natural way to approach problems. The analogizing power of a decision maker can be strengthened by a MCBR, a system that stores historical cases with all the relevant data kept intact. The ADDUCE-system infers how consumers will react to a new advertisement by searching relevant past advertisements. Thus, it can be interpreted as a knowledge system with a case-based reasoning system as one of its components. This indicates that the different systems are not necessarily distinct. McIntyre et al. (1993) built a MCBR to forecast the retail sales for a given promotion based on historical analogs from a case base.

Neural networks can be used to model the way human beings attach meaning to a set of stimuli or signals. Artificial neural networks can be trained to make the same types of associations between inputs and outputs as human beings do. An important feature of a neural network is its ability to learn. Marketing Neural Nets (MNNs) may be useful for the recognition of new product opportunities and to learn to distinguish between successful and less successful sales promotion campaigns. Other applications are the design of mailing campaigns and the selection of potential prospects. For extensive overviews see, for example, Hruschka (2008); Blattberg et al. (2008).

[30] See also Van Bruggen (1993).

10.5.3 Dashboards

Many MMSS work with marketing dashboards that bring the firm's key marketing metrics (Sect. 10.5.4) into a single display.[31] Many firms have created such a dashboards either by themselves or together with a dashboard service provider. The term "marketing dashboard" is borrowed from a vehicle dashboard which reports the few metrics the driver needs to know. Dashboards are related to MMSS that provide managers with guidance on decisions such as promotion activities and sales force allocation (see the discussion above). Hence the *dashboard display* is the output of a larger *dashboard system*. Figure 10.2 provides an illustration of a dashboard.

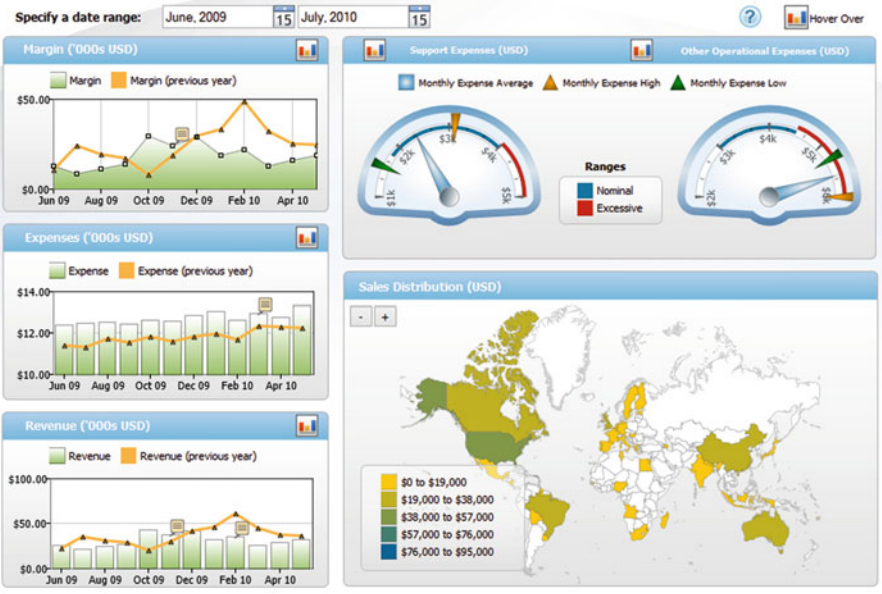

Fig. 10.2 Example of a Marketing Dashboard. *Source*: http://www.dundas.com

As many as 40 % of large US–UK companies report substantial efforts in this area (Clark et al. 2006). Many services firms use such dashboards to track marketing effectiveness and guide decision making, in industries such as business communication (e.g. Avaya), pay-TV broadcasting (e.g. British Sky), consumer credit (e.g. Capital One), online services (e.g. Google), gaming (e.g. Harrah's), hospitality (e.g. Hilton), investment banking (e.g. Morgan Stanley), systems integration (e.g. Unisys), and mutual funds (e.g. Vanguard). Moreover, dashboard providers and

[31] We closely follow Pauwels et al. (2009). See also Peters et al. (2013), who discuss social media metrics.

advisers themselves are offering a service, whether within their company (e.g. to their CMO or CEO) or to the client company.

The key elements of a dashboard include the summarization and integration of key performance metrics with underlying drivers to communicate performance throughout the organization. We define a dashboard as a relatively small collection of interconnected key performance metrics and underlying performance drivers that reflect both short- and long-term interests to be viewed in common throughout the organization.

The creation of a dashboard integrates the information of different functional areas, such as finance, accounting, R&D etc., and marketing. Integration is an important characteristic of dashboards in three ways:

- *Data*: Understanding the firm's market and its position within the market requires information and data from diverse sources at different levels of aggregation and covering different time periods. The dashboard provides a common organizing framework.
- *Processes*: The dashboard helps management relate inputs, such as marketing expenditures, to market performance measures and ultimately to financial performance, such as profits, cash flows, and shareholder value, thus building a bridge between internal and external reporting.
- *Viewpoints*: Whether assessing the market, performance, or planning, a dashboard allows different executives, in different departments and locations, to share the same, equally measured input, that is, to view the firm's market situation in the same light.

Pauwels et al. (2009) maintain that the integration of data proves easier than integrating processes and viewpoints. Performance information needs context, such as benchmarking, previous results, plan, or competition (Ambler et al. 2004). For multiservice or multiunit companies, performance is commonly across different services, market segments, or units. Visually, dashboards do this through devices such as gauges, charts, and tables, often color-coded for easy summarization (Bauer 2004; Lehmann and Reibstein 2006).

A dashboard enforces *consistency* in measures and measurement procedures across department and business units. Second, a dashboard helps to *monitor* performance. Monitoring in turn may be both evaluative (who or what performed well?) and developmental (what have we learned?). Dashboard metrics are early indicators of performance, and if a dip occurs in, for example, the "trust and privacy" metric, the company takes corrective action. Third, a dashboard may be used to *plan* (what should our goals and strategies be for the future given where we are now?). For instance, Ameritrade started with corporate scorecards from the strategic planning department to develop a dashboard that plugs in to the planning cycle and is tied to quarterly bonuses. Fourth, a dashboard may be used to *communicate* to important stakeholders. In particular, it communicates not only what the performance is but also what an organization values as performance by the choice of metrics on the dashboard.

10.5.4 Metrics

In many firms, marketers have a difficult time justifying their expenditures in terms of direct return on investment.[32] In other words, the inability to account for marketings contribution has undermined its standing within the firm. As McGovern et al. (2004, p. 74) state:

> "the [marketing] field is chock-a-block with creative thinkers, yet it's short on people who lean toward an analytic, left-brain approach to the discipline."

Two aspects are relevant in this respect. First, many marketers do not measure the effect of their actions, because they are unable or unwilling to do so or because they do not use the appropriate metrics and/or methods. As a consequence, many advertisements have no effect on sales (Vakratsas and Ambler 1999), sales promotions have no persistent influence on sales at either the brand or the category sales level (Nijs et al. 2001), and new products suffer from low success rates. Not surprisingly, CEOs cannot get clear, compelling answers about marketing's impact (see Kumar 2004, p. vii). Marketing productivity could increase if managers were able to measure it. Calls for more attention for accountability, marketing metrics, and dashboard marketing may be helpful in this respect (Farris et al. 2006). Second, appropriate specifications of metrics, especially metrics that measure long-term or persistent effects, are lacking. McGovern et al. (2004) argue that many managers do not know what to measure or how to interpret the results. For example, a manager might collect customer satisfaction scores and customer retention rates, but if he or she cannot explain these scores (in relation to marketing activities), the data are not very useful (Leeflang and Wittink 2000). Accountability also involves a determination of the effects of marketing activities on the value of the firm. Some recently successful attempts have helped determine these effects (e.g., Gupta et al. 2004; Pauwels et al. 2004; McAlister et al. 2007; Tellis and Johnson 2007; Srinivasan and Hanssens 2009; Joshi and Hanssens 2010; Osinga et al. 2011). The importance of accountability has been acknowledged widely (Lehmann 2004; Rust et al. 2004). Moorman and Rust (1999) show a positive relationship between accountability and the marketing department's influence within the firm, and O'Sullivan and Abela (2007) report that top management is more satisfied with marketing when it is more accountable. These findings are confirmed in studies by Verhoef and Leeflang (2009) and Verhoef et al. (2011).

The attention for accountability has stimulated many authors[33] to develop "new" concepts such as "metrics". A metric is a measuring system that quantifies a trend or characteristic. Marketing metrics usually refer to

- a monetary value (expenditures);
- a percentage (percentage of stores that stock your brand);

[32] We closely follow Verhoef and Leeflang (2009). See also Leeflang et al. (2014).

[33] Ambler (2003); Jeffery (2010); Farris et al. (2006, 2010).

- a count (number of unit sales);
- a rating (satisfaction);
- an index (price index).

We distinguish metrics that refer to performance measures and metrics that influence these performances (exogenous variables). The metrics that measure marketing's performance (the endogeneous variables) have evolved over time, as illustrated in Table 10.4. As Table 10.4 shows, there are many studies that relate marketing efforts to firm performance measures, such as customer life time value, customer equity, and even firm value (usually in the form of stock prices and volatility in stock prices). These studies demonstrate the importance and contribution of marketing efforts to firm value and probably (we hope) can help marketing regain its position in the boardroom as we discussed before.

Farris et al. (2006) define numerous metrics for exogeneous variables. Table 10.5 gives an overview of the most important metrics.

Ambler (2003) and Jeffery (2010) specify additional metrics. Farris et al. (2006) also specify metrics which are the outcomes of additional analyses such as:

- analyses in which *elasticities* are determined;
- analyses in which the baseline sales (total sales less incremental sales generated by a marketing program) are determined;
- outcomes of break-even analyses;
- outcomes of profitability analyses, etc.

Hence, Farris et al. (2006) also put price, advertising elasticities, margins per unit, break-even points in their list of metrics.

Many specific metrics have been specified in the area of database marketing. The (CRM) database contains variables, such as customer equity, customer lifetime value (CLV or LTV), retention rates, churn rates, etc.[34]

Other metrics that are useful to collect are the so-called *mind-set metrics*. Examples are advertising awareness, brand consideration and brand liking. Srinivasan et al. (2010) demonstrate that mind-set metrics can be used as advance warning signals that allow enough time for managerial action before market performance itself is altered. Mind-set metrics have also shown their value as diagnostic measures. An elaborate discussion on the use of marketing *and* financial metrics is provided by Mintz and Currim (2013).

[34]Compare, for example, Blattberg et al. (2008, Chapters 8 and 9).

Table 10.4 Performance measures in models over time

Metric	Example	Publications
Sales:		
• Product class	• Demand for electricity	Van Helden et al. (1987)
• Brand level	• Demand for cigarettes	Leeflang and Reuyl (1984)
	• Demand for pharmaceuticals	
	– Market share	Fischer et al. (2010)
	– Units	Leeflang and Wieringa (2010)
Profit	• Optimizing price, advertising expenditures etc.	Dorfman and Steiner (1954); Verdoorn (1956); Best (2004)
Brand equity	• Impact of marketing expenditures on equity	Yoo et al. (2000)
Customer satisfaction	• Impact of marketing expenditures on satisfaction	De Wulf et al. (2001); Gomez et al. (2004)
CLV/customer equity	• Budget allocation	Niraj et al. (2001); Reinartz et al. (2005)
WOM/net promoter score	• Impact of satisfaction on WOM	Arnett et al. (2003); Villanueva et al. (2008)
Gratitude	• Impact of investments in relationship marketing on gratitude	Palmatier et al. (2009); Wetzel et al. (2014)
Firm value	Impact on	
	• Brand equity	Madden et al. (2006)
	• Customer satisfaction	Fornell et al. (2006)
	• WoM	Luo (2009)
	• Customer equity	Kumar and Shah (2009)
	• Advertising, innovations, promotions	Srinivasan and Hanssens (2009)
	• Direct-to-consumer advertising on firm value	Osinga et al. (2011)

Source: Leeflang (2011, p. 82)

Table 10.5 Overview of metrics based on Farris et al. (2006)

Products and portfolio management
• Cannibalization rate, sales less from existing products/sales new products • Brand equity measures • Quality measures
Sales force and channel management
• Numeric distribution (percentage of outlets, in a defined universe, that stock a particular brand/product) • ACV = All Commodity Volume (numeric distribution, weighted by penetrated outlets' share of sales of *all* product categories) • PCV = Product Category Volume (see ACV, but now obtained for a particular *product category*) • Facings • Out-of-stock • Inventories • DPP = Direct Product Profitability (adjusted gross margin of products, less direct product costs) • Sales Force Compensations (money per sales person) • Sales Force Efforts (size of the sales force)
Pricing strategy
• Price per SKU (Stock Keeping Unit) • Discounts • Reservation price (maximum amount an individual is willing to pay for a product) • Cost price per SKU
Sales promotions
• Redemption rates (coupons redeemed divided by coupons distributed) • Costs for coupons and rebates • Percentage sales with coupon • Percent sales on deal • Percent time on deal (percentage of time during which temporary promotions are offered) • Pass-through rate (promotional discounts provided by the trade to consumers divided by discounts provided to the trade by manufacturers)
Advertising media and web metrics
• Gross Rating Points (GRP's) • Cost per Thousand Impressions (CPM) • Net Reach • Average Frequency • Share of Voice • Page views (the number of times a web page is served) • Click-through Rate • Cost per Click • Cost per Order • Cost per Customer Acquired • Visits • Visitors

References

Abraham, M.M., Lodish, L.M.: Getting the most out of advertising and promotion. Harv. Bus. Rev. **68**(3), 50–60 (1990)

Ambler, T.: Marketing and the Bottom Line, 2nd edn. Pearson Education, Edinburgh (2003)

Ambler, T., Kokkinaki, F., Puntoni, S.: Assessing marketing performance: Reasons for metrics selection. J. Mark. Manag. **20**(3/4), 475–498 (2004)

Ansari, A., Mela, C.F.: E-customization. J. Mark. Res. **40**, 131–145 (2003)

Arnett, D.B., German, S.D., Hunt, S.D.: The identity salience model of relationships marketing success: The case of nonprofit marketing. J. Mark. **67**(2), 89–105 (2003)

Barber, B., Odean, Y.: All that glitters: The effect of attention on the buying behavior of individual and institutional investors. Rev. Financ. Stud. **21**(2), 785–818 (2008)

Bauer, K.: The CPM dashboard: The visuals. DM Rev. **14**(5), 41 (2004)

Bayer, J., Lawrence, S., Keon, J.W.: PEP: An expert system for promotion marketing. In: E. Turban, P.R. Watkins (eds.) Applied Expert Systems. Elsevier Science, North-Holland, Amsterdam (1988)

Becker, J.U., Greve, G., Albers, S.: Left behind expectations. how to prevent CRM implementations from failing. GfK Mark. Intell. Rev. **2**(2), 35–41 (2010)

Best, R.J.: Market-Based Management. Prentice Hall, Upper Saddle River (2004)

Bharadwaj, P., Tuli, K.R., Bonfrer, A.: The impact of brand quality on shareholder wealth. J. Mark. **75**(5), 88–104 (2011)

Blattberg, R.C.: Tracker: An early test market forecasting and diagnostic model for new product planning. J. Mark. Res. **15**, 192–202 (1978)

Blattberg, R.C., Kim, B.D., Neslin, S.A.: Database Marketing: Analyzing and Managing Customers. Springer Science, New York (2008)

Bult, J.R., Wansbeek, T.J.: Optimal selection for direct mail. Mark. Sci. **14**, 378–394 (1995)

Bultez, A.V., Naert, P.A.: SH.A.R.P: Shelf allocation for retailers' profit. Mark. Sci. **7**, 211–231 (1988)

Burke, R.R.: Reasoning with empirical marketing knowledge. Int. J. Res. Mark. **8**, 75–90 (1991)

Burke, R.R., Rangaswamy, A., Wind, Y., Eliashberg, J.: A knowledge based system for advertising design. Mark. Sci. **9**, 212–229 (1990)

Capella, M., Webster, C., Kinard, B.R.: A review of cigarette advertising. Int. J. Res. Mark. **28**, 269–279 (2011)

Churchman, C.W., Schainblatt, A.: The researcher and the manager: A dialectic of implementation. Manag. Sci.**11**, B69–B87 (1965)

Clark, B.H., Abela, A.V., Ambler, T.: Behind the wheel. Mark. Manag.**15**(3), 18–23 (2006)

Cleeren, K., Dekimpe, M.G., Helsen, K.: Weathering product-harm crises. J. Acad. Mark. Sci. **36**, 262–270 (2008)

Cleeren, K., Van Heerde, H.J., Dekimpe, M.G.: How brands and categories can overcome product-harm crises. J. Mark. **77**(2), 58–77 (2013)

De Vries, L., Gensler, S., Leeflang, P.S.H.: Effects of owned and earned social media and paid traditional media on consumer mind-set metrics and acquisition. Working Paper, University of Groningen, Groningen (2014)

De Wulf, K., Odekerken-Schröder, G., Iacobucci, D.J.: Investments in consumer relationships: A cross-country and cross-industry exploration. J. Mark. **65**(4), 33–50 (2001)

Divakar, S., Ratchford, B., Shankar, V.: CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. Mark. Sci. **24**, 334–350 (2005)

Dorfman, R., Steiner, P.O.: Optimal advertising and optimal quality. Am. Econ. Rev. **44**, 826–836 (1954)

Doyle, P., Saunders, J.: Multiproduct advertising budgeting. Mark. Sci. **9**, 97–113 (1990)

Duffy, M.: Advertising and the consumption of tobacco and alcoholic drinks: A system-wide analysis. Scottish J. Polit. Econ. **38**, 369–385 (1991)

Fama, E.F.: Efficient capital markets: A review of theory and empirical work. J. Finance **25**, 383–417 (1970)
Farris, P.W., Bendle, N.T., Pfeifer, P.E., Reibstein, D.J.: Marketing Metrics: 50+ Metrics Every Executive Should Master. Wharton School Publishing, Upper Saddle River (2006)
Farris, P.W., Bendle, N.T., Pfeifer, P.E., Reibstein, D.J.: Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Pearson Education, Upper Saddle River (2010)
Fischer, M., Albers, S., Wagner, N., Frie, M.: Dynamic marketing budget allocation across countries, products, and marketing activities. Mark. Sci. **30**, 568–585 (2011)
Fischer, M., Leeflang, P.S.H., Verhoef, P.C.: Drivers of peak sales for pharmaceutical brands. Quant. Mark. Econ. **8**(4), 429–460 (2010)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation. Int. J. Forecast. **10**, 245–261 (1994)
Foekens, E.W., Leeflang, P.S.H., Wittink, D.R.: Varying parameter models to accommodate dynamic promotion effects. J. Econom. **89**, 249–268 (1999)
Fornell, C., Mithas, S., Morgeson III, F.V., Krishnan, M.S.: Customer satisfaction and stock prices: High returns, low risk. J. Mark. **70**(1), 3–14 (2006)
Gomez, F.I., McLaughlin, E.W., Wittink, D.R.: Customer satisfaction and retail sales performance: An empirical investigation. J. Retail. **80**, 265–278 (2004)
Gupta, S., Lehmann, D.R., Stuart, J.A.: Valuing customers. J. Mark. Res. **41**, 7–18 (2004)
Hanssens, D.M., Leeflang, P.S.H., Wittink, D.R.: Market response models and marketing practice. Appl. Stoch. Models Bus. Ind. **21**, 423–434 (2005)
Hanssens, D.M., Parsons, L.J., Schultz, R.L.: Market Response Models: Econometric and Time Series Analysis. Kluwer Academic, Boston (2001)
Horváth, C., Leeflang, P.S.H., Wieringa, J.E., Wittink, D.R.: Competitive reaction- and feedback effects based on VARX models of pooled store data. Int. J. Res. Mark. **22**, 415–426 (2005)
Hruschka, H.: Neural nets and genetic algorithms in marketing. In: B. Wierenga (ed.) Handbook of Marketing Decision Models. Springer, New York (2008)
Hu, T., Sung, H., Keeler, T.E.: The state antismoking campaign and the industry response: The effects of advertising on cigarette consumption in California. Am. Econ. Rev. **85**, 85–90 (1995)
Jeffery, M.: Data-Driven Marketing: The 15 Metrics Everyone in Marketing Should Know. Wiley, Hoboken (2010)
Joshi, A.M., Hanssens, D.M.: The direct and indirect effects of advertising spending on firm value. J. Mark. **74**(1), 20–33 (2010)
Karmarkar, K.S.: Integrative research in marketing and operations management. J. Mark. Res. **33**, 125–133 (1996)
Kayande, U., De Bruyn, A., Lilien, G.L., Rangaswamy, A., Van Bruggen, G.H.: How incorporating feedback mechanisms in a DSS affects DSS evaluations. Inform. Syst. Res. **24**(4), 527–546 (2009)
Kornelis, M., Dekimpe, M.G., Leeflang, P.S.H.: Does competitive entry structurally change key marketing metrics? Int. J. Res. Mark. **25**, 173–182 (2008)
Kumar, N.: Marketing as Strategy. Harvard Business School Press, Boston (2004)
Kumar, V., Fan, J., Gulati, R., Venkat, P.: Marketing-mix recommendations to manage value growth at P&G Asia-Pacific. Mark. Sci. **28**, 645–655 (2009)
Kumar, V., Shah, D.: Expanding the role of marketing: From customer equity to market capitalization. J. Mark. **73**, 119–136 (2009)
Lane, V., Jacobson, R.: Stock market reactions to brand extension announcements: The effects of brand attitude and familiarity. J. Mark. **59**(1), 63–77 (1995)
Leeflang, P.S.H.: Modeling competitive reaction effects. Schmalenbach Bus. Rev. **60**, 322–358 (2008)
Leeflang, P.S.H.: Paving the way for distinguished marketing. Int. J. Res. Mark. **28**, 76–88 (2011)
Leeflang, P.S.H., Reuyl, J.C.: On the predictive power of market share attraction models. J. Mark. Res. **11**, 211–215 (1984)

Leeflang, P.S.H., Verhoef, P.C., Dahlström, P., Freundt, T.: Challenges and solutions for marketing in a digital era. Eur. Manag. J. **32**(1), 1–12 (2014)
Leeflang, P.S.H., Wieringa, J.E.: Modeling the effects of pharmaceutical marketing. Mark. Lett. **21**, 121–133 (2010)
Leeflang, P.S.H., Wittink, D.R.: Building models for marketing decisions: Past, present and future. Int. J. Res. Mark. **17**, 105–126 (2000)
Lehmann, D.R.: Metrics for making marketing matter. J. Mark. **68**(4), 73–75 (2004)
Lehmann, D.R., Reibstein, D.J.: Marketing Metrics and Financial Performance. Marketing Science Institute Monograph (2006)
Lilien, G.L.: Bridging the academic - practitioner divide in marketing decision models. J. Mark. **75**(4), 196–210 (2011)
Lilien, G.L., Rangaswamy, A.: Marketing Engineering, 2nd rev. edn. Trafford Publishing, Bloomington (2004)
Lilien, G.L., Roberts, J., Shankar, V.: Effective marketing science applications: Insights from ISMS-MSI practice prize finalist papers and projects. Mark. Sci. **32**, 229–245 (2013)
Little, J.D.C.: BRANDAID: A marketing-mix model, part 1: Structure. Oper. Res. **23**, 628–655 (1975)
Lodish, L.M.: CALLPLAN: An interactive salesman's call planning system. Manag. Sci.**18**, B25–B40 (1971)
Luo, X.: Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. Mark. Sci. **28**, 148–165 (2009)
Madden, T.J., Fehle, F., Fournier, S.: Brands matter: An empirical demonstration of the creation of shareholder value through branding. J. Acad. Mark. Sci. **34**, 224–235 (2006)
Manchanda, P., Dubé, J.P., Goh, K.Y., Chintangunta, P.: The effects of banner advertising on internet purchasing. J. Mark. Res. **43**, 98–108 (2006)
McAlister, L., Srinivasan, R., Kim, M.: Advertising, research and development, and systematic risk of the firm. J. Mark. **71**(1), 35–48 (2007)
McCann, J.M., Gallagher, J.P.: Expert Systems for Scanner Data Environments. Kluwer Academic, Boston (1990)
McCann, J.M., Lahti, W.G., Hill, J.: The brand managers assistant: A knowledge-based system approach to brand management. Int. J. Res. Mark. **8**, 51–73 (1991)
McGovern, G.J., Court, D., Quelch, J.A., Crawford, B.: Bringing customers back into the boardroom. Harv. Bus. Rev. **82**(11), 70–80 (2004)
McIntyre, S.H., Achabal, D.D., Miller, C.M.: Applying case-based reasoning to forecasting retail sales. J. Retail. **69**, 372–398 (1993)
Mintz, O., Currim, I.: What drives managerial use of marketing and financial metrics and does metric use affect performance of marketing-mix activities? J. Mark. **77**(2), 17–40 (2013)
Moorman, C., Rust, R.T.: The role of marketing. J. Mark. **63**(Special Issue), 180–197 (1999)
Moorman, C., Wies, S., Mizik, N., Spencer, F.J.: Firm innovation and the ratchet effect among consumer packaged goods firms. Mark. Sci. **31**, 934–951 (2012)
Myers, S.C., Majluf, N.S.: Corporate financing and investment decisions when firms have information that investors do not have. J. Financ. Econ. **13**(2), 187–221 (1984)
Nijs, V.R., Dekimpe, M.G., Steenkamp, J.B.E.M., Hanssens, D.M.: The category-demand effects of price promotions. Mark. Sci. **20**, 1–22 (2001)
Niraj, R., Gupta, M., Narasimhan, C.: Customer profitability in the supply chain. J. Mark. **65**(3), 1–16 (2001)
Onishi, H., Manchanda, P.: Marketing activity, blogging and sales. Int. J. Res. Mark. **29**, 221–234 (2012)
Ornstein, S.I., Hanssens, D.M.: Resale price maintenance: Output increasing or restricting? The case of distilled spirits in the United States. J. Ind. Econ. **36**, 1–18 (1987)
Osinga, E.C., Leeflang, P.S.H., Srinivasan, S., Wieringa, J.E.: Why do firms invest in consumer advertising with limited sales response? A shareholder perspective. J. Mark. **75**(1), 109–124 (2011)

O'Sullivan, D., Abela, A.V.: Marketing performance measurement ability and firm performance. J. Mark. **71**(2), 79–83 (2007)
Palda, K.S.: The Measurement of Cumulative Advertising Effects. Prentice-Hall, Englewood Cliffs (1964)
Palmatier, R.W., Jarvis, C.B., Bechkoff, J.R., Kardes, F.R.: The role of customer gratitude in relationship marketing. J. Mark. **73**(5), 1–18 (2009)
Pauwels, K., Ambler, T., Clark, B.H., LaPointe, P., Reibstein, D., Skiera, B., Wierenga, B., Wiesel, T.: Dashboards as a service why, what, how, and what research is needed? J. Serv. Res. **12**, 175–189 (2009)
Pauwels, K.H., Silva-Risso, J., Srinivasan, S., Hanssens, D.M.: New products, sales promotions, and firm value: The case of the automobile industry. J. Mark. **68**(4), 142–156 (2004)
Peters, K., Chen, Y., Kaplan, A.M., Ognibeni, B., Pauwels, K.: Social media metrics – A framework and guidelines for managing social media. J. Interact. Mark. **27**, 281–298 (2013)
Rangaswamy, A., Eliashberg, J., Burke, R.R., Wind, Y.: Developing marketing expert systems: An application to international negotiations. J. Mark. **53**(4), 24–39 (1989)
Rego, L.L., Billett, M.T., Morgan, N.A.: Consumer-based brand equity and firm risk. J. Mark. **73**(6), 47–60 (2009)
Reinartz, W., Thomas, J.S., Kumar, V.: Balancing acquisition and retention resources to maximize customer profitability. J. Mark. **69**(1), 63–79 (2005)
Rust, R.T., Ambler, T., Carpenter, G.S., Kumar, V., Srivastava, R.K.: Measuring marketing productivity: Current knowledge and future direction. J. Mark. **69**(4), 76–89 (2004)
Silk, A.J., Urban, G.L.: Pre-test-market evaluation of new packaged goods: A model and measurement methodology. J. Mark. Res. **15**, 171–191 (1978)
Sinha, P., Zoltners, A.: Sales-force decision models: Insights from 25 years of implementation. Interfaces **31**, 8–44 (2001)
Srinivasan, S., Hanssens, D.M.: Marketing and firm value: Metrics, methods, findings, and future directions. J. Mark. Res. **46**, 293–312 (2009)
Srinivasan, S., Hsu, L., Fournier, S.: Branding and firm value. In: S. Bharadwaj, S. Ganesan (eds.) Handbook of Marketing and Finance. Edward Elgar (2011)
Srinivasan, S., Pauwels, K., Silva-Risso, J., Hanssens, D.M.: Product innovations, advertising, and stock returns. J. Mark. **73**(1), 24–43 (2009)
Srinivasan, S., Pauwels, K., Wiesel, T.: The marketing/finance interface. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in Honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)
Srinivasan, S., Vanhuele, M., Pauwels, K.: Mind-set metrics in market response models: An integrative approach. J. Mark. Res. **27**, 672–684 (2010)
Srivastava, R.K., Shervani, T.A., Fahey, L.: Market based assets and shareholder value: A framework for analysis. J. Mark. **62**(1), 2–18 (1998)
Stephen, A.T., Galak, J.: The effects of traditional and social earned media on sales: A study of a microlending marketplace. J. Mark. Res. **49**, 624–639 (2012)
Sullivan, M.: Measuring image spillovers in umbrella-branded products. J. Bus. **63**, 309–329 (1990)
Tellis, G.J., Johnson, J.J.: The value of quality. Mark. Sci. **26**, 758–773 (2007)
Urban, G.L.: Building models for decision makers. Interfaces **4**, 1–11 (1974)
Urban, G.L.: Pretest market forecasting. In: J. Eliashberg, G.L. Lilien (eds.) Handbooks in Operations Research and Management Science Marketing, vol. 5. North-Holland, Amsterdam (1993)
Vakratsas, D., Ambler, T.: How advertising works: What do we really know? J. Mark. **63**(1), 26–43 (1999)
Valenzuela, A., Dhar, R., Zettelmeyer, F.: Contingent response to self-customization procedures: Implications for decision satisfaction and choice. J. Mark. Res. **46**, 754–763 (2009)
Van Bruggen, G.H.: The effectiveness of marketing management support systems. Unpublished Ph.D. thesis, Erasmus University, Rotterdam (1993)

Van Heerde, H.J., Helsen, K., Dekimpe, M.G.: The impact of a product-harm crisis on marketing effectiveness. Mark. Sci. **26**, 230–245 (2007)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: The estimation of pre- and postpromotion dips with store-level scanner data. J. Mark. Res. **37**, 383–395 (2000)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Semiparametric analysis to estimate the deal effect curve. J. Mark. Res. **38**, 197–215 (2001)
Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R.: Decomposing the sales promotion bump with store data. Mark. Sci. **23**, 317–334 (2004)
Van Helden, G.J., Leeflang, P.S.H., Sterken, E.: Estimation of the demand for electricity. Appl. Econ. **19**(1), 69–82 (1987)
Verdoorn, P.J.: Marketing from the producer's point of view. J. Mark. **20**(3), 221–235 (1956)
Verhoef, P.C., Leeflang, P.S.H.: Understanding marketing department's influence within the firm. J. Mark. **73**(2), 14–37 (2009)
Verhoef, P.C., Leeflang, P.S.H., Natter, M., Reiner, J., Baker, W., Grinstein, A., Gustafsson, A., Morrison, P., Saunders, J.: A cross-national investigation into the marketing department's influence within the firm: Toward initial empirical generalizations. J. Int. Mark. **19**(3), 59–86 (2011)
Villanueva, J., Yoo, S., Hanssens, D.M.: The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. J. Mark. Res. **45**, 48–59 (2008)
Vitorino, M.A.: Understanding the effect of advertising on stock returns and firm value: Theory and evidence from a structural model. Manag. Sci.**60**, 227–245 (2014)
Vriens, M.: The Insights Advantage: Knowing How to Win. iUniverse, Bloomington (2012)
Wetzel, H.A., Hammerschmidt, M., Zablah, A.R.: Gratitude versus entitlement: A dual process model of the profitability implications of customer prioritization. J. Mark. **78**(1), 1–19 (2014)
Wierenga, B.: The ROI of marketing management support systems. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in Honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)
Wierenga, B., Dalebout, A., Dutta, A.: BRANDFAME: A marketing management support system for the brand manager. In: B. Wierenga, G. Van Bruggen (eds.) Marketing Management Support Systems: Principles, Tools, and Implementation. Kluwer Academic, Boston (2000)
Wierenga, B., Van Bruggen, G.H.: The integration of marketing problem solving modes and marketing management support systems. J. Mark. **6**(3), 21–37 (1997)
Wierenga, B., Van Bruggen, G.H., Althuizen, N.A.P.: Advances in marketing management support systems. In: B. Wierenga (ed.) Handbook of Marketing Decision Models. Springer, New York (2008)
Wiesel, T., Pauwels, K., Arts, J.: Marketing's profit impact: Quantifying online and off-line funnel progression. Mark. Sci. **30**, 604–611 (2011)
Wittink, D.R., Addona, M.J., Hawkes, W.J., Porter, J.C.: SCAN*PRO: The estimation, validation and use of promotional effects based on scanner data. In: J.E. Wieringa, J.C. Hoekstra, P.C. Verhoef (eds.) Liber Amicorum in Honor of Peter S. H. Leeflang. University of Groningen, Groningen (2011)
Xiong, G., Bharadwaj, S.: Asymmetric roles of advertising and marketing capability in financial returns to news: Turning bad into good and good into great. J. Mark. Res. **50**, 706–724 (2013)
Yoo, B., Donthu, N., Lee, S.: An examination of selected marketing mix elements and brand equity. J. Acad. Mark. Sci. **28**, 195–211 (2000)
Zenetti, G., Bijmolt, T.H.A., Leeflang, P.S.H., Klapper, D.: Search engine advertising effectiveness in a multimedia campaign. Int. J. Electron. Comm. **18**(3), 7–38 (2014)

Appendix A
Matrix Algebra

A.1 Matrices and Simple Matrix Operations

A matrix is a rectangular array of elements. A typical matrix has m rows and n columns, for example the matrix A given as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

By the *order* of a matrix we mean the number of rows and columns a matrix has, where we always name the rows of the matrix first. So, for example, the matrix A has order $m \times n$, because it has m rows and n columns. We call an entry of the matrix an *element* and indicate a typical element by a_{ij}, where i denotes the row and j denotes the column of the element. The element on the first row in the first column of A is for example a_{11}, while the fourth element on the sixth row of A is denoted by a_{64}. In this element notation, we can also denote A as $A = (a_{ij})_{m \times n}$.

Matrix *addition* and *subtraction* are defined similarly to their scalar counterparts. We can only add matrices of the same size, that is with the same number of rows and the same number of columns. In the new matrix, the sum of the separate elements is the element of the new matrix. So, for example

$$\begin{bmatrix} 4 & -1 & 0 \\ 9 & 2 & 5 \\ -4 & 0 & 11 \end{bmatrix} + \begin{bmatrix} 2 & 7 & 1 \\ -3 & 0 & -5 \\ 1 & 8 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 6 & 1 \\ 6 & 2 & 0 \\ -3 & 8 & 15 \end{bmatrix},$$

P.S.H. Leeflang et al., *Modeling Markets*, International Series
in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0

while

$$\begin{bmatrix} 3 & 1 & 7 \\ 2 & -1 & 9 \end{bmatrix} + \begin{bmatrix} 10 & 3 \\ 8 & -3 \\ 5 & 0 \end{bmatrix}$$

is not defined because the rows and columns of both matrices are unequal. Matrix subtraction is defined in a similar way, where we should realize that $-A$ is equal to

$$-A = \begin{bmatrix} -a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & -a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{m1} & -a_{m2} & \cdots & -a_{mn} \end{bmatrix},$$

that is, each element of A is multiplied by -1. Matrix subtraction is thus nothing else then adding a matrix where each element is multiplied by -1 to another matrix. An element of the new matrix therefore consists of the difference of the elements of the matrices to be subtracted, for example

$$\begin{bmatrix} 4 & -1 & 0 \\ 9 & 2 & 5 \\ -4 & 0 & 11 \end{bmatrix} - \begin{bmatrix} 2 & 7 & 1 \\ -3 & 0 & -5 \\ 1 & 8 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -8 & -1 \\ 12 & 2 & 10 \\ -5 & -8 & 7 \end{bmatrix}.$$

Again, the number of rows and columns of each matrix should be equal. We have already seen that multiplication of a matrix by -1 boils down to multiplying each element by -1. This result holds in general and is known as *scalar multiplication*. The product of a matrix A and a scalar r is defined as

$$r \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} ra_{11} & ra_{12} & \cdots & ra_{1n} \\ ra_{21} & ra_{22} & \cdots & ra_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ra_{m1} & ra_{m2} & \cdots & ra_{mn} \end{bmatrix},$$

so for example we would get

$$4 \begin{bmatrix} 3 & 1 & 7 \\ 2 & -1 & 9 \end{bmatrix} = \begin{bmatrix} 12 & 4 & 28 \\ 8 & -4 & 36 \end{bmatrix}.$$

Another operation we can define is the *transpose* of a matrix, denoted as A' or A^T. We will use the notation A' throughout this text. When we transpose a matrix, the first row of the matrix becomes the first column, the second row becomes the second column, and so on. So, for example we would get

$$A = \begin{bmatrix} 3 & 1 & 7 \\ 2 & -1 & 9 \end{bmatrix}, \qquad A' = \begin{bmatrix} 3 & 2 \\ 1 & -1 \\ 7 & 9 \end{bmatrix}$$

for a matrix A and its transpose A'. If a matrix and its transpose are the same ($A = A'$), we call that matrix a *symmetric matrix*. A symmetric matrix is by definition a square matrix (that is, $m = n$), because otherwise the orders of the matrix and its transpose would be different even. An example of a symmetric matrix is

$$A = \begin{bmatrix} 3 & 1 & 7 \\ 1 & 6 & 4 \\ 7 & 4 & 11 \end{bmatrix} = A'.$$

A.2 Matrix Multiplication

While we have seen that matrix addition, subtraction and multiplication by a scalar are easily done for matrices, the multiplication of matrices is a little more difficult. In matrix multiplication, the order in which the matrices are multiplied matters and multiplication is not always possible even. If we wish to multiply two matrices A and B, it should hold that the number of columns of A is equal to the number of rows of B. If we let A be a $m \times n$ matrix and B a $n \times k$ matrix, the product of these matrices $C = A \times B$ is an $m \times k$ matrix with typical elements c_{ij} defined as

$$c_{ij} = \sum_{s=1}^{n} a_{is} b_{sj},$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, k$. That is, we get a typical element c_{ij} by multiplying the element a_{is} in the ith row with the element b_{sj} in the jth column and then adding all these products. To illustrate this, we will give two examples:

$$\begin{aligned} AB &= \begin{bmatrix} 3 & 1 & 7 \\ 2 & -1 & 9 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 7 & -4 \\ -2 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 3(1) + 1(7) + 7(-2) & 3(0) + 1(-4) + 7(5) \\ 2(1) + -1(7) + 9(-2) & 2(0) + -1(-4) + 9(5) \end{bmatrix} \\ &= \begin{bmatrix} -4 & 31 \\ -23 & 49 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
BA &= \begin{bmatrix} 1 & 0 \\ 7 & -4 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} 3 & 1 & 7 \\ 2 & -1 & 9 \end{bmatrix} \\
&= \begin{bmatrix} 1(3)+0(2) & 1(1)+0(-1) & 1(7)+0(9) \\ 7(3)+-4(2) & 7(1)+-4(-1) & 7(7)+-4(9) \\ -2(3)+5(2) & -2(1)+5(-1) & -2(7)+5(9) \end{bmatrix} \\
&= \begin{bmatrix} 3 & 1 & 7 \\ 13 & 11 & 13 \\ 4 & -7 & 31 \end{bmatrix}
\end{aligned}
$$

Notice that in this case both products AB and BA exist. However, this needs not always be the case depending on the number of rows and columns both matrices have. Therefore, one should remember that in general it does *not* hold that $AB = BA$.

The *transpose* operation is also defined in matrix multiplication. If we have that $C = AB$, then it holds that $C'=B'A'$. To see why this holds, we look at one element of the matrix C, namely the element $c_{ij} = \sum_{s=1}^{n} a_{is}b_{sj}$. Then we get that

$$
\begin{aligned}
c'_{ij} &= c_{ji} && \text{(Definition of transpose)} \\
&= \sum_{s=1}^{n} a_{js}b_{si} && \text{(Definition of matrix multiplication)} \\
&= \sum_{s=1}^{n} a'_{sj}b'_{is} && \text{(Again applying definition of transpose)} \\
&= \sum_{s=1}^{n} b'_{is}a'_{sj} && \text{(Using that } a \times b = b \times a \text{ for scalars)}
\end{aligned}
$$

and we find that C' is indeed equal to $B'A'$. This result is easily generalized to the multiplication of more than two matrices. Hence if $D = ABC$, we get that $D' = C'B'A'$.

Now that we are familiar with matrix multiplication, we can also define *matrix powers*, for example A^2, A^{361} or in general A^k for any non-negative integer k. Definitions for zero and negative k will follow later. Matrix multiplication can be seen as multiplying a matrix with itself k times. In the case of A^2 we would therefore get that $A^2 = AA$ and in general we have that $A^k = AA \ldots A$(repeat A k times). Notice however that we can only define matrix powers for square matrices, as we multiply the matrix by itself and only square matrices have the correct dimensions to do this. As an example, we will compute the square of the (symmetric) matrix

$$A = \begin{bmatrix} 3 & 1 & 7 \\ 1 & 6 & 4 \\ 7 & 4 & 11 \end{bmatrix}$$

as follows:

$$\begin{aligned} A^2 = AA &= \begin{bmatrix} 3 & 1 & 7 \\ 1 & 6 & 4 \\ 7 & 4 & 11 \end{bmatrix} \begin{bmatrix} 3 & 1 & 7 \\ 1 & 6 & 4 \\ 7 & 4 & 11 \end{bmatrix} \\ &= \begin{bmatrix} 3(3)+1(1)+7(7) & 3(1)+1(6)+7(4) & 3(7)+1(4)+7(11) \\ 1(3)+6(1)+4(7) & 1(1)+6(6)+4(4) & 1(7)+6(4)+4(11) \\ 7(3)+4(1)+11(7) & 7(1)+4(6)+11(4) & 7(7)+4(4)+11(11) \end{bmatrix} \\ &= \begin{bmatrix} 59 & 37 & 102 \\ 37 & 53 & 75 \\ 102 & 75 & 186 \end{bmatrix}. \end{aligned}$$

This example also demonstrates that the square of a symmetric matrix is also symmetric.

A.3 Special Matrices

We have already seen that a matrix with the same number of rows and columns ($m = n$) is called a square matrix. Besides this matrix, there are several other matrices that have special properties. The first matrix we will discuss is the *identity matrix*. This is the square $n \times n$ matrix with ones on the diagonal and zeros on the off-diagonal. We denote it by I_n, where n is the order of the matrix. It thus takes the form

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The identity matrix plays a similar role to that of 1 in scalar algebra. If A is an $m \times n$ matrix, then we have that

$$I_m A = A I_n = A$$

showing that multiplying with the correct identity matrix just returns the matrix A. It is also easily seen that $I_n' = I_n$. Referring to the section on matrix powers, we are also in a position now to define A^0 as the corresponding identity matrix, that is $A^0 = I_n$ if A is a square $n \times n$ matrix.

The identity matrix is a special case of a *diagonal matrix*, that is matrix with entries on the diagonal and zero entries for all the off-diagonal entries. Its general form is

$$D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix},$$

where the a_{ii} are allowed to take on any value. This means that the diagonal entries can be zero as well. It also is not required that the matrix is $n \times n$, as long as the matrix contains entries starting from a_{11} along the diagonal and has zero entries everywhere else, the matrix is called diagonal. For example,

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \end{bmatrix}$$

also is called a diagonal matrix, despite not being an $n \times n$ matrix.

Related to diagonal matrices are *upper* and *lower triangular matrices*. These matrices have entries on the diagonal and either above or below the diagonal, while the other part of the matrix consists of zeros. An example for both is given by

$$A = \begin{bmatrix} -1 & 5 & 2 \\ 0 & -4 & 9 \\ 0 & 0 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 7 & 0 \\ -7 & 2 & 6 \end{bmatrix},$$

where A is an upper triangular matrix and B is a lower triangular matrix. These matrices can be of any size $m \times n$ as well, by a similar argument as for diagonal matrices.

Two other special matrices are the *row matrix* and the *column matrix*, where matrix can also be read as vector in this case. A row matrix has one row, that is $m = 1$, while a column matrix has one column, thus $n = 1$. Examples are

$$A = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 5 \\ 2 \end{pmatrix},$$

where A is a row matrix and B is a column matrix.

Related to matrix powers is the *idempotent matrix*. We call a matrix idempotent if $A \times A = A$ or in general $A^n = A$, that is multiplying the matrix by itself returns the matrix. A trivial example of such a matrix is the previously discussed identity matrix. Another matrix is

$$A = \begin{bmatrix} 6 & -6 \\ 5 & -5 \end{bmatrix}$$

for example. Note that we still require the matrix to be square for it to be idempotent.

A.4 Matrix Inverse

There is another important matrix operation that we have not yet discussed up until now, namely the *inverse* of a matrix. When we have a square $n \times n$ matrix A, we call the square $n \times n$ matrix B such that

$$AB = BA = I$$

the inverse of A and denote it as A^{-1}. Due to definition of B above, we can also that A is the inverse of B, that is both are inverses of each other. We state here without proof that each matrix can only have at most one inverse and thus the inverse of a matrix is unique.

We stated that a matrix has *at most* one inverse. This already suggests that not every matrix has an inverse. The existence of an inverse is determined by the properties a particular matrix has. Although there are several properties that can be used to determine whether a matrix is invertible, we will focus on one called the *rank* of the matrix. To get to the definition of rank, we start as follows: Let A be an $m \times n$ matrix. Then we say that A has *full column rank* if for any n-vector x it holds that

$$Ax = 0 \Rightarrow x = 0.$$

Here the arrow should be read as '*implies*'. If this holds, we say that the n columns of A are linearly independent. Linear independence in this case thus means that none of the columns can be formed as a linear combination of the other columns of the matrix. Now we say that the *rank* of A is the maximum number of linearly independent columns that A has. This number is less than or equal to the number of columns that A has. Because our goal is to define the conditions under which the inverse of a matrix exists and the inverse only exists for square ($m = n$) matrices, this definition of rank is sufficient for our purposes. For a matrix to be invertible, we require that is has maximum rank or equivalently is of full column rank, that is $\text{rank}(A) = n$. This implies that all columns of the matrix should be linearly independent.

Now that we know when an inverse exists, the question arises how to find it when it does exist. The most general way of finding the inverse matrix is by applying Gaussian elimination on the augmented matrix $[A : I]$. We will give an example for a 3×3 matrix A below. Let

$$A = \begin{bmatrix} 1 & 4 & 0 \\ 2 & 6 & -1 \\ -3 & 0 & 7 \end{bmatrix}.$$

It can be checked that the rank of A is 3, that is the matrix has full rank, as none of the columns can be formed by a linear combination of the other columns. Now we form the augmented matrix and apply Gaussian elimination to A:

$$\begin{aligned}
[A : I] &= \left[\begin{array}{ccc:ccc} 1 & 4 & 0 & 1 & 0 & 0 \\ 2 & 6 & -1 & 0 & 1 & 0 \\ -3 & 0 & 7 & 0 & 0 & 1 \end{array}\right] \\
&= \left[\begin{array}{ccc:ccc} 1 & 4 & 0 & 1 & 0 & 0 \\ 0 & -2 & -1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -9 & 6 & 1 \end{array}\right] \\
&= \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & -21 & 14 & 2 \\ 0 & 1 & 0 & 5.5 & -3.5 & -0.5 \\ 0 & 0 & 1 & -9 & 6 & 1 \end{array}\right].
\end{aligned}$$

To obtain the second line we put the matrix A in what is called the row-echelon form, which is the upper triangular matrix obtained when applying elementary row operations to the matrix A. To go from the second to the third line, we move from the row echelon form to what is called the reduced row echelon form, where also the entries above the diagonal are zero. As we also apply the same operations to the matrix on the right hand side simultaneously, this matrix changes in each step. The matrix on the right hand side then is the inverse of A, that is

$$A^{-1} = \begin{bmatrix} -21 & 14 & 2 \\ 5.5 & -3.5 & -0.5 \\ -9 & 6 & 1 \end{bmatrix}.$$

Using this method it is always possible to obtain the inverse matrix, should one exist. In practice however this procedure gets tedious fast and other methods have been developed to deal with larger matrices. These methods are incorporated in most software packages.

Having defined the inverse matrix, we can also deal with matrix powers that have negative, integer k as exponent. For these matrices it holds that $A^{-k} = (A^{-1})^k = (A^k)^{-1}$, that is we can take the matrix power of the inverse matrix k times or, similarly, take the inverse of the kth matrix power to obtain the matrix A^{-k}. This is a straightforward generalization of the earlier theory discussed on matrix powers.

A.5 Determinants

In the previous section we used the rank of a matrix as a criterion for invertibility of a matrix. We can also use a number called the *determinant* to judge whether we can invert a matrix. The determinant is a unique number associated with each square matrix, which in general is derived in the setting of solving systems of linear equations or in a geometric setting. For the purpose of exposition it is enough to know that the determinant is a unique number which we can compute for a square matrix.

We will start with a 2×2 matrix and compute the determinant. Consider the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

We can compute its determinant denoted by $\det(A)$ as follows: We take the product of the diagonal elements and subtract from that the product of the off-diagonal elements, which gives $\det(A)=a_{11}a_{22} - a_{12}a_{22}$. This is the basic definition of the determinant in the setting of a 2×2 matrix. This is also makes clear why we did not start with 1×1 matrices for our definition, as their determinant would just be the number in the matrix itself. We will however use this definition for the 1×1 matrix to generalize the definition of the determinant. To see how, we write

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}\det(a_{22}) - a_{12}\det(a_{21}).$$

We compute the first part of the determinant by taking the entry a_{11} of the matrix and multiplying it by the determinant from the matrix obtained by deleting the row and column containing the entry a_{11}, which is the 1×1 matrix containing $det(a_{22})$. The second part of the determinant is computed similarly by taking entry a_{12} and multiplying it by the determinant of the matrix that is left when we delete the row and column corresponding to the entry, which leaves the matrix with a_{21} as element. Moreover, observe that the signs given to different parts alternate: The first part gets a plus sign, the second part a minus sign.

With these observations done in the case of 2×2 matrices, we are now in a position to give a general definition of the determinant, which we will illustrate

for the case of 3×3 matrices. In general, let A be an $n \times n$ matrix. Let A_{ij} be the $(n-1) \times (n-1)$ submatrix obtained from A by deleting the ith row and jth column. Let

$$M_{ij} = \det(A_{ij})$$

which is called the (i, j)th *minor* of A. Furthermore, define

$$C_{ij} = (-1)^{i+j} M_{ij}$$

to be the (i, j)th *cofactor* of A. The cofactor is equal the minor multiplied by 1 or -1 depending on whether $i + j$ is odd or even. So, if $i + j$ is even we have $C_{ij} = M_{ij}$, while if $i + j$ is odd we get that $C_{ij} = -M_{ij}$. Check that in the 2×2 case discussed previously we get that $\det(A) = a_{11}M_{11} - a_{12}M_{12} = a_{11}C_{11} + a_{12}C_{12}$. We now have the tools to go beyond 2×2 matrices, as we will show by computing the determinant for a 3×3 matrix.

For a general 3×3 matrix A, we get

$$\begin{aligned}
\det(A) &= \det\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} \\
&= a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13} \\
&= a_{11}\det\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{12}\det\begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} + a_{13}\det\begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \\
&= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.
\end{aligned}$$

Thus, we can compute the determinant by considering smaller and smaller submatrices until we reach a point where we can compute the determinant exactly, which in this case is a 2×2 matrix. This example can directly be generalized to the case of an $n \times n$ matrix, which gives the following definition for the determinant of an $n \times n$ matrix A:

$$\begin{aligned}
\det(A) &= a_{11}C_{11} + a_{12}C_{12} + \ldots + a_{1n}C_{1n} \\
&= a_{11}M_{11} - a_{12}M_{12} + \ldots + (-1)^{n+1}a_{1n}M_{1n}.
\end{aligned}$$

With this general definition known, we can return to the start of this section where we remarked that the determinant can be used to determine whether or not we can invert a matrix. If the determinant of a matrix is not 0, we say that a matrix is invertible. Subsequently, only matrices with determinant 0 cannot be inverted. This provides another straightforward way to check for the invertibility of a matrix.

We will conclude this section with an example. Consider again the matrix

$$A = \begin{bmatrix} 1 & 4 & 0 \\ 2 & 6 & -1 \\ -3 & 0 & 7 \end{bmatrix}$$

for which we have already computed the inverse. We have also shown that this matrix has full rank. Now we will compute the determinant and show its nonzero as follows:

$$\begin{aligned} \det(A) &= \det \begin{bmatrix} 1 & 4 & 0 \\ 2 & 6 & -1 \\ -3 & 0 & 7 \end{bmatrix} \\ &= 1 \times \det \begin{bmatrix} 6 & -1 \\ 0 & 7 \end{bmatrix} - 4 \times \det \begin{bmatrix} 2 & -1 \\ -3 & 7 \end{bmatrix} + 0 \times \det \begin{bmatrix} 2 & 6 \\ -3 & 0 \end{bmatrix} \\ &= 1 \times 42 - 4 \times 11 + 0 \times 18 \\ &= -2. \end{aligned}$$

Hence the determinant of the matrix is -2, which is nonzero, and therefore we can invert the matrix, as we have seen previously.

A.6 Eigenvalues and Eigenvectors

In this section we will study another aspect of square matrices, namely the existence of so-called *eigenvalues* and their corresponding *eigenvectors*. In short, the eigenvalue is a number λ which when subtracted from the diagonal of a square matrix A results in the matrix to be singular. We say that a matrix is *singular* when the inverse of the matrix does not exist. This means a.o. that the determinant is 0 and the rank is less than n, as we have seen previously. This might seem abstract, therefore we give an example. Consider the matrix

$$A = \begin{bmatrix} 5 & 3 & 3 \\ 3 & 5 & 3 \\ 3 & 3 & 5 \end{bmatrix}$$

and now subtract 2 from the diagonal elements, which gives

$$A = \begin{bmatrix} 3 & 3 & 3 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \end{bmatrix}.$$

This matrix is singular as all columns are exactly the same now and the rank of the matrix thus is less than 3. Therefore, the number 2 is an eigenvalue of this matrix.

While it was easy to find the eigenvalue of the matrix above, in general we have to do some more work to find the eigenvalues of a matrix. However, the way to find the eigenvalues is similar to that of computing the determinant, which is something we are capable of doing. The only difference is that, by definition of the eigenvalue, we need to subtract a number λ from the diagonal. The resulting formula will be a polynomial which we can solve for the number λ, thus obtaining the eigenvalues. In fact, we are computing $\det(A - \lambda I) = 0$ when we want to obtain the eigenvalues, with I the identity matrix corresponding to A. To illustrate, consider the following example for a general 2×2 matrix:

$$\begin{aligned} \det(A - \lambda I) &= \det\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = 0 \\ &\Leftrightarrow (a_{11} - \lambda)(a_{22} - \lambda) - (a_{12}a_{21}) = 0 \\ &\Leftrightarrow \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0. \end{aligned}$$

Solving this equation for λ gives us the eigenvalues if they exist. To give a numeric example, we will compute the eigenvalues in the following example:

$$\begin{aligned} A &= \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \\ \det(A - \lambda I) &= \det\begin{bmatrix} 3 - \lambda & 0 \\ 4 & 5 - \lambda \end{bmatrix} = 0 \\ &\Leftrightarrow (3 - \lambda)(5 - \lambda) - 0 = 0 \\ &\Leftrightarrow \lambda^2 - 8\lambda + 15 = 0 \\ &\Leftrightarrow \lambda = \frac{8 \pm \sqrt{64 - 60}}{2} \\ &\Leftrightarrow \lambda = 5 \quad \text{or} \quad \lambda = 3. \end{aligned}$$

Hence, we see that by solving the polynomial equation we find that the eigenvalues are 3 and 5 in this case. By computing the determinant of the matrix $A - \lambda I$ and setting the determinant to 0, we are always able to find the eigenvalues, should they exist. The number of distinct eigenvalues a matrix can have is smaller than or equal to the order of the matrix, which is n. That the number can be smaller is caused by the possibility of one eigenvalue having a multiplicity of 2 or more, which can happen easily when we consider larger matrices. Another aspect that we need to be aware of is that eigenvalues can be complex, which is the case if the polynomial we need to solve has complex solutions. All in all, we can't say a priori how many distinct eigenvalues a matrix has just by looking at it.

Eigenvectors are vectors that correspond to a certain eigenvalue and can be found by solving the equation

$$(A - \lambda I)\, v = 0$$

for a non-zero vector v, as the zero vector is trivially a solution to this equation. For illustration, we will compute the eigenvectors for the matrix previously discussed. We already notice now that eigenvectors are not unique.

$$\begin{aligned}
(A-3I)\, v &= \begin{bmatrix} 3-3 & 0 \\ 4 & 5-3 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \\
&= \begin{bmatrix} 0 & 0 \\ 4 & 2 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \\
&= \begin{bmatrix} 0v_1 + 0v_2 \\ 4v_1 + 2v_2 \end{bmatrix} = 0 \\
&\Rightarrow \mathbf{v} = \begin{pmatrix} 1 \\ -2 \end{pmatrix} \\
(A-5I)\, v &= \begin{bmatrix} 3-5 & 0 \\ 4 & 5-5 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \\
&= \begin{bmatrix} -2 & 0 \\ 4 & 0 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \\
&= \begin{bmatrix} -2v_1 + 0v_2 \\ 4v_1 + 0v_2 \end{bmatrix} = 0 \\
&\Rightarrow v = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.
\end{aligned}$$

We thus have found one eigenvector for each eigenvalue. As we remarked before, eigenvectors are not unique. We could have for example taken $\begin{pmatrix} 2 \\ -4 \end{pmatrix}$ as eigenvector corresponding to the eigenvalue 3 and still have ended up with a solution to the system of equations. We therefore always take the vector with the greatest common divisor as the eigenvector. Also notice that we could have taken the zero vector as eigenvector corresponding to eigenvalue 5. However, as eigenvectors are required to be non-zero, we took 1 as the number to be multiplied with 0 in the equations. We could have taken any other number as well, but it is customary to take 1 in that case.

A.7 Definiteness of a Matrix

In this section we will present some more results for symmetric matrices related to what is called the *definiteness* of the matrix. We will use some results on eigenvalues to establish the definiteness of a matrix, which is why we have postponed discussion of this subject until now.

The notion of the definiteness of a matrix can be seen as the matrix equivalent to the notion of positive and negative real numbers. First we will introduce all the different versions of definiteness and afterwards we will provide some additional results. Let A be a symmetric, $n \times n$ matrix. Then we call A:

- *positive definite* if $x'Ax > 0$ for all $x \neq 0$ with $x \in R^n$,
- *positive semidefinite* if $x'Ax \geq 0$ for all $x \neq 0$ with $x \in R^n$,
- *negative definite* if $x'Ax < 0$ for all $x \neq 0$ with $x \in R^n$,
- *negative semidefinite* if $x'Ax \leq 0$ for all $x \neq 0$ with $x \in R^n$,
- *indefinite* if $x'Ax > 0$ for some $x \neq 0$ and < 0 for some other $n \neq 0$, both $x \in R^n$.

In the above, R^n denotes the collection of all real numbers in the n-dimensional Euclidean space. These five notions are used to determine the definiteness of a matrix. Notice that if a matrix is positive or negative semidefinite, it also is positive or negative definite by definition. Checking these formal definitions can be tedious however, as they require checking results for every non-zero vector x. Luckily, using the eigenvalues of a matrix, we can also determine the definiteness of a matrix. This boils down to finding all eigenvalues and checking their signs. Using the eigenvalues as criterion, we say that a symmetric, $n \times n$ matrix A is

- *positive definite* if and only if all eigenvalues are > 0,
- *positive semidefinite* if and only if all eigenvalues are ≥ 0,
- *negative definite* if and only if all eigenvalues are < 0,
- *negative semidefinite* if and only if all eigenvalues are ≤ 0,
- *indefinite* if A has both eigenvalues that are > 0 and < 0.

Thus, it is enough to know the signs of the eigenvalues to determine the definiteness of a symmetric matrix. We will present some results without proof that can be derived using the notions of definiteness. First, let $A > 0$ denote that A is positive definite. Similarly, $A \geq 0$ then denotes that A is positive semidefinite.

A strong result is that if $A > 0$, it is invertible, which gives us yet another method to check whether we can invert a matrix. Furthermore, then it holds that

$$A > 0 \Leftrightarrow A^{-1} > 0$$

or more generally

$$A \geq B > 0 \Leftrightarrow B^{-1} \geq A^{-1} > 0.$$

These results show that computation of the inverse is not necessary if we are interested in finding out whether or not the inverse is positive (semi)definite.

A.8 Matrix and Vector Differentiation

Vectors and matrices like other mathematical functions can be differentiated. Let x be a (column) vector with dimension $n \times 1$, x' a row vector with dimension $1 \times n$, A a matrix with dimension $n \times m$, γ is a scalar, y is column vector with dimension $m \times 1$. We consider the differentiation of scalars, vectors and matrices, respectively.

First, we show the differentiation of the scalar γ, where γ is defined as

$$\gamma = x'a = a'x = \sum_{i=1}^{n} ax_i$$

and a is a $n \times 1$ vector. The derivative of γ to a column vector x can be written as

$$\frac{\partial \gamma}{\partial x} = \begin{bmatrix} \frac{\partial \gamma}{\partial x_1} \\ \vdots \\ \frac{\partial \gamma}{\partial x_n} \end{bmatrix} = a$$

and similarly

$$\frac{\partial \gamma}{\partial x'} = \begin{bmatrix} \frac{\partial \gamma}{\partial x_1} \cdots \frac{\partial \gamma}{\partial x_n} \end{bmatrix} = a'.$$

We now consider the case where

$$\gamma = x'Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

where A is a $n \times n$ matrix. Differentiation of γ to an element of x_k, $k = i, j = 1, \ldots, n$ gives:

$$\begin{aligned} \frac{\partial \gamma}{\partial x_k} &= \frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j}{\partial x_k} \\ &= \frac{\partial \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1, i \neq j}^{n} \sum_{j=1, i \neq j}^{n} a_{ij} x_i x_j}{\partial x_k} \\ &= 2a_{kk} x_k + 2 \sum_{j=1, j \neq k}^{n} a_{kj} x_j \\ &= 2 \sum_{j=1}^{n} a_{kj} x_j. \end{aligned}$$

If we perform this operator for each k we obtain:

$$\frac{\partial x'Ax}{\partial x} = 2Ax.$$

In a similar way we can derive that

$$\begin{aligned}
\frac{\partial x'Ax}{\partial x'} &= 2x'A', \\
\frac{\partial x'Ay}{\partial x} &= Ay, \\
\frac{\partial x'Ay}{\partial y} &= A'x, \\
\frac{\partial x'Ay}{\partial y'} &= x'A.
\end{aligned}$$

We apply these findings to derive expressions in the (general) linear model. For a given cross section j, the relations between the criterion variable y_{jt} and the predictor variables $x_{\ell jt}$, the unknown parameters $\beta_{\ell j}$ and the disturbance terms can be written as

$$\begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jT} \end{bmatrix} = \begin{bmatrix} x_{1j1} & x_{2j1} & \cdots & x_{Lj1} \\ x_{1j2} & x_{2j2} & \cdots & x_{Lj2} \\ \vdots & \vdots & & \vdots \\ x_{1jT} & x_{2jT} & \cdots & x_{LjT} \end{bmatrix} \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \beta_{Lj} \end{bmatrix} + \begin{bmatrix} u_{j1} \\ u_{j2} \\ \vdots \\ u_{jT} \end{bmatrix}$$

where

y_{jt} = the value of the criterion variable in observation jt,

$x_{\ell jt}$ = the value of the ℓ-th predictor variable, $\ell = 1, \ldots, L$, for observation jt,

$\beta_{\ell j}$ = the unknown parameters, $\ell = 1, \ldots, L$, for cross section j,

u_{jt} = the value of the disturbance term in observation jt,

T = the number of observations per cross section.

In matrix notation this system becomes:

$$y_j = X_j\beta_j + u_j,$$

where

y_j = a column vector of T values for the criterion variable,

X_j = a matrix of order $T \times L$ with values taken by the L predictor variables $x_{1j}, \ldots, x_{Lj}$,

β_j = a column vector of L unknown parameters, and

u_j = a column vector of T disturbance terms.

For convenience we delete the index j and have:

$$y = X\beta + u.$$

The least squares estimates of β are obtained by minimizing the sum of squares of the residuals: $u'u$. We can write $u'u$ as

$$\begin{aligned} G(\beta) &= (y - X\beta)'(y - X\beta) \\ &= y'y - 2\beta' y'X + \beta' X'X\beta. \end{aligned}$$

This expression is minimized by least-squares by differentiating this expression to β:

$$\frac{\partial G(\beta)}{\partial \beta} = -2X'y + 2X'X\beta,$$

which we equate to zero. The result may be written as

$$X'X\beta = X'y,$$

and

$$\beta = (X'X)^{-1}X'y.$$

Remark that even when X is not a square $(n \times n)$ matrix, but has order $m \times n$ then $X'X$ has the order $n \times n$ and, when $X'X$ is not singular, its inverse $(X'X)^{-1}$ exists, at least in principle.

Author Index

P.S.H. Leeflang et al., *Modeling Markets*, International Series in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0

C

D

E

L

M

S

Subject Index

P.S.H. Leeflang et al., *Modeling Markets*, International Series in Quantitative Marketing, DOI 10.1007/978-1-4939-2086-0