UNIVERSITAS FRIBURGENSIS
FACULTY OF ECONOMICS AND SOCIAL SCIENCES

Michael Beer

Hedonic Elementary Price Indices

Axiomatic Foundation and Estimation Techniques

swissuniversity.ch

Hedonic Elementary Price Indices

Hedonic Elementary Price Indices

Axiomatic Foundation and Estimation Techniques

Thesis

presented to the Faculty of Economics and Social Sciences
at the University of Fribourg Switzerland
in fulfillment of the requirements for the degree of
Doctor of Economics and Social Sciences

by

Michael Beer

from Trub (Berne).

Accepted by the Faculty Council on 18 December 2006 at the proposal of
Professor Dr Hans Wolfgang Brachinger (first advisor) and
Professor Dr Dr h.c. mult. Wolfgang Eichhorn (second advisor)

Fribourg 2006

The Faculty of Economics and Social Sciences at the University of Fribourg neither approves nor disapproves the opinions expressed in a doctoral thesis. They are to be considered those of the author. (Decision of the Faculty Council of 23 January 1990)

Typeset with LaTeX in Latin Modern and Univers.

ISBN 978-3-033-01099-4

To my godmother

Preface

> *Few statistics produced by governmental agencies have greater impact than* price indexes. *Statisticians, however, seldom study these indexes. This is not because index construction is straightforward. Indeed, in terms of the controversies surrounding price indexes and the complexity of the data integration they require, they might make a claim to be the* most *complicated of statistics.*
>
> Dorfman et al., 1999

My choice of price statistics as a subject for research was motivated by my first advisor, Professor Dr Hans Wolfgang Brachinger. Our discussions all along the development of the thesis were always intense and fruitful, and many concepts presented here became much sharper and clearer through this kind of exchange. I am indebted to him for his support and encouragement.

Similarly, it was a great honour for me that Professor Dr Dr h.c. mult. Wolfgang Eichhorn accepted to be the second advisor. With his seminal work on the axiomatic approach to index number theory, he established the foundations for a series of research programmes that mine became part of.

During the four and a half years of this project, various people contributed to its progress in various ways. The first to be mentioned are the professors and fellow collaborators at the Department of Quantitative Economics in particular and at the Faculty of Economics and Social Sciences at the University of Fribourg in general who stimulated my research by asking the right questions at the right time. Special thanks go to my colleagues at the Seminar of Statistics for the convivial atmosphere.

This piece of work would have been onerous to complete if I had not had the opportunity to apply the theoretic concepts to real-world data. The openness and the interest the collaborators at AutoScout24 showed in the topic have been of great value for my work. The same holds for the

experts at the price division of the Swiss Federal Statistical Office I got into contact with. Further thanks are owed to Dr Pawel Bednarek who generously rendered every assistance in the maintenance of our Linux data server.

My doctoral thesis is part of a research project funded by the Swiss National Science Foundation as well as the Swiss Federal Statistical Office. I am grateful for their support.

Last but not least, my deepest gratitude goes to my friends, my fiancée, my family, most particularly to my mother whom I would have wished to live and accompany me so much longer, and to my father who has given me all the support I needed throughout the various stages of my life.

Michael Beer, Freiburg
September 2006

Contents

1 Introduction 1

I An Axiomatic Framework for Hedonic Elementary Price Indices

2 Elementary Aggregates and the Hedonic Hypothesis 9
2.1 Goods and elementary aggregates 9
2.2 Objects and characteristics 10
2.3 Characteristics and prices 13

3 Hedonic Elementary Price Indices 17
3.1 Elementary price indices and quality change 17
3.1.1 Bilateral comparison of time periods 17
3.1.2 Elementary price index concepts 18
3.1.3 Measuring the pure price change 19
3.1.4 Conventional methods of quality adjustment 20
3.1.5 Summary and critical comments 21
3.2 Hedonic elementary price indices 22
3.2.1 The hedonic econometric model revisited 22
3.2.2 Hedonic elementary price indices 23
3.3 A stochastic approach to hedonic elementary price indices 26
3.3.1 The sampling approach 26
3.3.2 Universal formulae for hedonic elementary price indices 30

4 An Axiomatic Approach to Hedonic Elementary Price Indices 33
4.1 Bilateral hedonic elementary price indices 33
4.2 Axioms for hedonic elementary price indices 34

4.3 Testing the index definitions . . . 38
4.4 Conclusion . . . 45

II Estimation Techniques for Hedonic Elementary Price Indices

5 Estimating the Hedonic Function . . . 49
5.1 General considerations . . . 49
5.2 Model selection . . . 50
5.3 Regression models . . . 53
5.3.1 Linear regression . . . 53
5.3.2 PLS regression . . . 55
5.4 Model assessment . . . 57
5.5 Further econometric considerations . . . 58

6 Estimating Hedonic Elementary Price Indices . . . 61
6.1 Index estimators . . . 61
6.1.1 Reference quality . . . 61
6.1.2 Bilateral hedonic elementary price index estimators . . . 64
6.1.3 Double imputation . . . 64
6.1.4 The time dummy variable method . . . 67
6.2 Bootstrap replications of hedonic elementary price indices . . . 68
6.2.1 Estimation error and confidence intervals . . . 68
6.2.2 Resampling methods . . . 69

III Hedonic Elementary Price Indices for Used Cars

7 Data Source: The AutoScout24 Marketplace . . . 77
7.1 Why used cars? . . . 77
7.2 Data retrieval, storage, structure, and analysis . . . 79
7.2.1 Data import . . . 79
7.2.2 Preparing the data for further analysis . . . 87
7.2.3 Infrastructure for statistical analysis . . . 88
7.3 Descriptive statistics and data filtering . . . 89
7.4 Conceptual limitations of the data set . . . 94

8 Estimating the Hedonic Function for Used Cars **99**
8.1 Linear regression **99**
8.1.1 Simple semi-log model (SSL) **99**
8.1.2 Critical comments **102**
8.1.3 Enhanced semi-log model (ESL) **104**
8.1.4 Simple double-log model (SDL) **105**
8.1.5 Flexible double-log model (FDL) **106**
8.2 PLS regression **106**
8.2.1 Simple PLS model (SPLS) **106**
8.2.2 Enhanced PLS model (EPLS) **108**
8.3 Per-model simple semi-log model (SSL/m) **109**
8.4 Overall examination of the different models **112**
8.5 Concluding remarks **120**

9 Estimating Hedonic Elementary Price Indices for Used Cars **121**
9.1 Hedonic elementary price index estimates **121**
9.2 Estimates of bootstrap confidence intervals **129**

10 Conclusions, Recommendations, and Open Questions **139**

Appendix

A Tables and Figures **149**

B The R package 'hepi' **165**

C Bibliography **179**

CHAPTER 1

Introduction

Consumer price indices and quality change

Price indices play an important role in the economic theory and practice. They provide a measure of the average proportionate change in the prices of a specified set of goods over time (see ILO et al., 2004, para. 1.1; UNITED NATIONS, 1993, para. 16.14). The aim of a *consumer price index* (CPI) in particular is to measure the price evolution of goods and services that households consume.

Consumer price indices may be viewed from an economic standpoint as a measure of the evolution of the cost of living that households are confronted with. They are then defined 'as the ratio of the minimum expenditures needed to attain the given level of utility, or welfare, under two different price regimes' (ILO et al., 2004, para. 1.85). It is this interpretation of a CPI that is applied for indexing salaries, rents, pensions or social security benefits in order to reflect changes in the purchasing power of money. Furthermore, consumer price indices are commonly used as a proxy for the general rate of inflation, although they only take account of goods and services that households consume. In some countries, CPIs are used to set inflation targets for purposes of monetary policy (ILO et al., 2004, para. 1.11).

This wide domain of application of CPIs requires careful consideration with regard to their construction. It is important to recognise that small differences in the movements of a CPI may have considerable financial implications. In consideration of the different purposes and interpretations that such an index is going to have, it might even be necessary to calculate

and publish different CPIs for different uses (ILO et al., 2004, para. 1.12; Brachinger et al., 1999, pp. 33–43).

It is not the aim of this thesis to give an overview of the various approaches to consumer price indices that can be found in the literature. Neither is it to discuss the question of whether it is more appropriate in practice to try and estimate a cost of living index (COLI) as opposed to a cost of goods index (COGI) on a fixed basket of consumer goods. Some of these aspects may, however, reappear punctually in later sections. An all-embracing source of information on consumer price indices in general, including a huge number of detailed reflections in this context, is the Consumer Price Index Manual published by the International Labour Office (ILO et al., 2004).

This thesis is going to step into some fundamental aspects of a problem which is seen as one of the most difficult to handle when compiling price indices. The universe of products that households consume is continually changing. New products appear on the market while existing products disappear. It is then necessary to link price observations for a former item to those of a new item in order to measure the continuing price change (ILO et al., 2004, para. 1.226 ff.). If the quality of products change over time, however, it is necessary to estimate the contribution of the quality change to the observed price change in order not to confuse two different sources of variation. A CPI is meant to measure the overall price change of a set of consumer goods having constant quality.

The famous report of the Boskin Commission estimated the bias in the 'CPI-based measure of the cost of living' due to the 'difficulty of adjusting fully for quality change and the introduction of new products' (Boskin et al., 1998, p. 10) at about 0.6 percentage points per annum. This represents slightly more than half of the total bias of 1.1 percentage points identified therein. In view of the 'extremely large sums of money' (ILO et al., 2004, para. 1.12) concerned by the movements of the CPI, tackling the problem of quality change may thus have a significant impact on the state of finances of households and institutions.

Hedonic elementary price indices

The prevalent CPI concepts used in practice usually structure the basket of consumer goods in a hierarchical manner. Individual price observations are transformed into a final index value through a sequence of aggregation steps. In the first stage, the price evolution is individually observed for restricted groups of homogeneous products, the so-called *elementary expenditure aggregates*. These aggregates usually serve as strata for data collection. A price index for an elementary expenditure aggregate is called an *elementary price in-*

dex. In a second stage, these elementary price indices are weighted by their quantitative relevance on the market and then averaged, or aggregated, to a higher-level price index (see ILO et al., 2004, para. 9.6 ff.; Brachinger et al., 1999, pp. 72–9).

The issue of adjusting the price measurements for quality change, as introduced above, appears on the base level of the aggregation structure of a CPI. The quality equivalence of two items is important when their prices are directly compared. Such a comparison, however, is the main aim of an *elementary* price index. Higher-level indices take exclusively the elementary price indices as inputs, along with 'weights derived from the values of elementary aggregates in some earlier year or years' (ILO et al., 2004, para. 9.77). There is thus neither a possibility nor a need for adjusting higher-level price indices for quality change. Quality adjustment is an issue of elementary price indices.

One possible manner of tackling the problems mentioned above, namely the handling of differences and changes in quality of the items within an elementary expenditure aggregate, is based on the so-called *hedonic approach*. Its main idea is to identify the quality of a product—or, in other words, its 'potential contribution ... to the welfare and happiness of its purchasers and the community' (Court, 1939, p. 107)—with a vector of product characteristics. A regression equation is then estimated relating the latter to the price of the product. Once such a relationship between characteristics and price of a product is established, the price of any similar item can be predicted by introducing its characteristics into the estimated *hedonic (regression) function*.

With regard to price index estimation, this technique allows to impute prices of items for which an observed price is not available in some of the considered time periods. If, for example, a specific product disappears from the market, it is usually possible to estimate its hypothetical price based on observations of similar products which are still available. The same is true for new items and their imputed prices in previous periods. More generally, the hedonic approach allows to control for quality differences over time within the framework of elementary price indices.

Hedonic elementary price indices, i.e. elementary price indices where the quality adjustments are based on the hedonic approach, have been developed as an alternative to conventional quality adjustment methods that are still widely used in practice. (See e.g. ILO et al., 2004, Chap. 7 or Triplett, 2004, Chap. II for a comprehensive overview on the conventional methods, and Brachinger et al., 1999 or Bundesamt für Statistik, 2006 for details regarding their application in the Swiss CPI.) Early applications of

hedonic elementary price indices date back to the 1920s and 1930s and have been studied with rapidly increasing interest since then. An overview on the development of literature and research is given by Triplett (2004, App. A to Chap. III) or White et al. (2004, p. 3), for example. Nowadays, the hedonic index approach is frequently seen as the 'most intellectually satisfying of the various quality-adjustment methods because it appeals to an underlying economic structure rather than to opportunistic proxies' (Hulten, 2003, p. 9).

Aim and structure of the thesis

The aim of this research project is, first, to develop an axiomatic framework for hedonic elementary price indices and, secondly, to study techniques for their estimation. In contrast to standard reference works on hedonic indices, such as the one by Triplett (2004), we try to follow a normative rather than a historical or explorative approach. Therefore, this thesis does not intend to give a complete overview neither on the hedonic elementary price index literature nor on the current practice of national statistical institutes. We will see, however, that many of the concepts currently in use nicely fit into the general framework developed in the following chapters.

In *Part I* of the thesis, a precise definition of what could be regarded as an ideal type of a hedonic elementary price index is established and discussed from an axiomatic point of view. Our interest lies predominantly in a stochastic approach where a hedonic index is interpreted as a relationship between several probability distributions. *Chapter 2* deals with the formal definition of characteristics and elementary aggregates. The intention here is to characterise the key elements of hedonic functions. *Chapter 3* introduces the time dimension and makes the step towards quality-adjusted elementary price indices. The hedonic idea is brought into the framework, and two alternative universal formulae of hedonic elementary price indices are established. *Chapter 4*, finally, qualifies and further specifies these general formulae based on a list of axioms describing the characteristics an ideal hedonic index should have. Sufficient conditions for indices satisfying specific axioms are worked out.

In *Part II*, several techniques for estimating hedonic functions and hedonic elementary price indices are investigated from a theoretic point of view. *Chapter 5* deals with the task of estimating the hedonic function from a random sample of items of the elementary aggregate concerned. Linear and partial least squares regression models are described, and a criterion for assessing the predictive power of a model is introduced. The main interest in this second part, however, lies in the bilateral hedonic elementary price

index estimators which are treated in *Chapter 6*. Starting from the universal formulae developed in the first part, several alternative implementations of estimators are discussed. Moreover, three bootstrap resampling methods are introduced with the aim of estimating confidence intervals for hedonic elementary price indices.

Part III, finally, takes the market of used cars in Switzerland as an example for implementing the concepts discussed in the previous chapters. *Chapter 7* describes the data and the steps that were necessary for transforming the raw data into a form that is suitable for estimating hedonic functions and indices. Furthermore, the conceptual limitations of this specific data set are discussed. *Chapter 8* tackles the problem of an empirical estimation of the hedonic models introduced in Part II. Several implementations are compared and discussed. In *Chapter 9*, these hedonic models are then used to estimate bilateral hedonic elementary price indices for used cars as well as appropriate confidence intervals for the period of October 2004 until March 2006.

The thesis ends with a list of summarising conclusions and recommendations as well as some open questions for further research.

PART I

An Axiomatic Framework for Hedonic Elementary Price Indices

CHAPTER 2

Elementary Aggregates and the Hedonic Hypothesis

2.1 Goods and elementary aggregates

The general aim of a consumer price index is to measure 'changes in the prices of consumption goods and services' (ILO et al., 2004, para. 3.1 ff.). Here, the primary emphasis lies on 'consumption', and a *consumption good or service* is defined as 'one that members of households use, directly or indirectly, to satisfy their own personal needs and wants.' In order to build a formal framework for price indices, however, it is first necessary to discuss how the term 'goods and services' has to be interpreted in this context.

The notion of a *good* has a long history in the economic literature. According to Milgate (1987), the plural term *goods* appeared in the English language with the meaning of 'objects or things which confer some advantage or produce some desirable effect upon their owner'. Later, it was confronted with *commodities* which was less tightly attached to the concept of providing *utility* than *goods* originally was, but rather to *exchangeable value* in a more general sense. Only recently, and largely influenced by the German tradition, has the noun *good* in singular form entered the formal language of English economics.

For the purpose of price index research, it is sufficient to follow the practice of, e.g., Hirshleifer (1980, p. 17), who introduces the notion of *commodities* as a synonym of *goods and services*. 'Goods', he writes, 'as distinguished from services are physical things (wares or merchandise)' while

'services represent a flow of benefits over a period of time'. He then admits that commodities and goods can also be thought of as 'synonymous words covering also desired consumption services.'

The approach we are going to adopt here follows the observation of LANCASTER (1971) that 'the "goods" of traditional theory are typically such aggregates as "automobiles," "food," "clothing," rather than individual goods as strictly defined.' We already discussed in the introduction that a consumer price index is usually built upon elementary indices for a list of *elementary expenditure aggregates*. These are defined in function of the applied aggregation structure such as COICOP (Classification of Individual Consumption according to Purpose, see ILO et al., 2004, paras. 3.162–8). An elementary aggregate, in this context, consists of 'a small and relatively homogeneous set of products defined within the consumption classification used in the CPI' (ILO et al., 2004, para. 1.120). The notions of a good and an elementary aggregate are thus often interchangeable.

Finally, a precise definition of the notion of a good with regard to hedonic indices is given by BRACHINGER (2002). He describes a good as being 'characterized by the set of all those models or variants ... which fit under one and the same hedonic equation'. This definition, however, requires a general acceptance of the hedonic hypothesis (see later) for any object even before the notion of a good or an elementary aggregate can be introduced.

From a general point of view, a good can be viewed as a set of objects being very similar in some fundamental properties or serving the same purpose. This idea will serve as a basis for the definition of an elementary aggregate below. It will be specified as a set of objects that share the same well-defined set of *characteristics*. This concept is now going to be formalised.

2.2 Objects and characteristics

Let $\mathcal{O}$ denote the set of all objects supplied on a market. Here, the notion of an *object* means physically tangible items as well as services and other immaterial entities to which a price can be assigned. Each of these objects exhibits a set of *characteristics*. Examples of such characteristics might be the volume or the physical mass of the object. Others might comprise the location of sale or any after-sales services. In general, the nature and level of measurement of characteristics can be very diverse.

It is obvious that not every characteristic can be observed for any given object $o \in \mathcal{O}$. Processor speed, for example, is a characteristic of computers but not of clothes or bicycles. In general, each characteristic m is only

defined on a specific subset $\mathcal{O}_m$ of $\mathcal{O}$. As a rule, it should be possible to quantify the values of any characteristic such that they form a subset of the Euclidean real number space. This leads to the following definition.

Definition 2.1. *A **characteristic** m is a real-valued function $m : \mathcal{O}_m \longrightarrow \mathbb{R}$ defined on a non-empty subset $\mathcal{O}_m$ of $\mathcal{O}$. The set $\mathcal{O}_m$ is called the **domain** of m and, for each $o \in \mathcal{O}_m$, $m(o)$ will be called the m-**value** of o.* ◇

For the sake of simplicity, the m-value of o may also be called its *m-characteristic*, or just characteristic if it is clear which characteristic is meant.

Notation. The set of all characteristics as defined in Def. 2.1 will be denoted by $\mathcal{M} := \{m : \mathcal{O}_m \longrightarrow \mathbb{R} \,|\, \mathcal{O}_m \subset \mathcal{O}, \mathcal{O}_m \neq \emptyset\}$.

Having established the notion of a characteristic, it is now possible to specify the notion of an elementary aggregate.

Definition 2.2. *An **elementary aggregate** $\mathcal{G}$ is a set of objects in $\mathcal{O}$ having the following properties:*

1. *The set $\mathcal{M}_\mathcal{G}$ of the characteristics defined for all elements of $\mathcal{G}$ is not empty, i.e.*
$$\mathcal{M}_\mathcal{G} := \{m \in \mathcal{M} \,|\, \mathcal{O}_m \supset \mathcal{G}\} \neq \emptyset\,.$$
2. *The intersection of the domains of all characteristics contained in $\mathcal{M}_\mathcal{G}$ is a subset of $\mathcal{G}$, i.e.*
$$\bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}_m \subset \mathcal{G}\,.$$

*Each object $o \in \mathcal{G}$ will be called an **item** of the elementary aggregate $\mathcal{G}$. The elements of $\mathcal{M}_\mathcal{G}$ are called **distinguishing characteristics** of $\mathcal{G}$.* ◇

As a result of property 1, two objects supplied on the market will only be classified as items of one elementary aggregate if they share at least one common characteristic. Conversely, if two objects do *not* belong to the same elementary aggregate, there must be a characteristic that is observable for one of these objects but not for the other. Moreover, the second property ensures that an elementary aggregate embraces each object carrying all characteristics of $\mathcal{M}_\mathcal{G}$.

The following proposition shows that each elementary aggregate has some kind of maximality property in the sense that its distinguishing characteristics fully determine the items of the aggregate. In other words, there is no

item of an elementary aggregate that is not contained in the intersection of the domains of all distinguishing characteristics.

Proposition 2.3. *Each elementary aggregate $\mathcal{G}$ equals the intersection of the domains of its distinguishing characteristics, i.e.*

$$\bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}_m = \mathcal{G}\,. \qquad \diamond$$

Proof. We have $\mathcal{G} \subset \mathcal{O}_m$ for all $m \in \mathcal{M}_\mathcal{G}$. Therefore, $\mathcal{G} \subset \bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}_m$. The inclusion in the other direction is given by property 2 of Def. 2.2 and hence equality holds. ∎

It should be noted that an elementary aggregate in the sense of Def. 2.2 may still comprise a lot of different items. Thus we define the term 'elementary aggregate' in a more general sense than it is usually used in practice. However, it follows from Proposition 2.3 that supplementing the set $\mathcal{M}_\mathcal{G}$ of distinguishing characteristics of an elementary aggregate with additional characteristics leads to a reduction of $\mathcal{G}$. The elementary expenditure aggregates known in price statistics (see ILO et al., 2004, para. 9.7) are thus embraced by the current definition.

As a consequence of Proposition 2.3, it is possible to construct elementary aggregates from samples of individual objects. Let $\mathcal{O}^* \subset \mathcal{O}$ be any set of objects. These might be, e.g., different models of personal computers. Let $\mathcal{M}_{\mathcal{O}^*} := \{m \in \mathcal{M} \,|\, \mathcal{O}_m \supset \mathcal{O}^*\}$ be the set of all common characteristics of these objects and assume that $\mathcal{M}_{\mathcal{O}^*}$ is not empty. Then, it is possible to specify an elementary aggregate $\mathcal{G}(\mathcal{O}^*)$ induced by $\mathcal{O}^*$. The induced aggregate is defined as the intersection of the domains of all characteristics in $\mathcal{M}_{\mathcal{O}^*}$, i.e.

$$\mathcal{G}(\mathcal{O}^*) := \bigcap_{m \in \mathcal{M}_{\mathcal{O}^*}} \mathcal{O}_m\,. \tag{2.1}$$

The set $\mathcal{O}^*$ is thus extended by all those objects on the market that can be specified by at least the same characteristics as the objects fixed in $\mathcal{O}^*$. Obviously, by means of (2.1), any given set of characteristics $\mathcal{M}$ induces an elementary aggregate $\mathcal{G}(\mathcal{M}) := \bigcap_{m \in \mathcal{M}} \mathcal{O}_m$.

Moreover, it is important to note that any finite subset $\{m_1, \ldots, m_K\} \subset \mathcal{M}_\mathcal{G}$ of the distinguishing characteristics of an elementary aggregate $\mathcal{G}$ can be assembled to a vector function

$$\begin{array}{rccl} \boldsymbol{m} & : & \mathcal{G} & \longrightarrow & \mathbb{R}^K \\ & & o & \longmapsto & \boldsymbol{m}(o) := (m_1(o), \ldots, m_K(o))'\,. \end{array}$$

All the items of $\mathcal{G}$ are by these means identified with a K-dimensional vector of characteristics. The identification of objects with a characteristics vector leads to an equivalence relation on $\mathcal{G}$ defined by

$$o_1 \sim_{\boldsymbol{m}} o_2 \quad :\Longleftrightarrow \quad \boldsymbol{m}(o_1) = \boldsymbol{m}(o_2)\,.$$

Two items of an elementary aggregate are thus identified if and only if their $\boldsymbol{m}$-values, i.e. their m_1- to m_K-values coincide. The equivalence classes respective to the relation $\sim_{\boldsymbol{m}}$ will be called $\boldsymbol{m}$-equivalence classes. They partition $\mathcal{G}$ into subsets containing items with equal $\boldsymbol{m}$-values. Finally, the quotient set induced by this equivalence relation will be denoted by $\mathcal{G}/\sim_{\boldsymbol{m}}$.

2.3 Characteristics and prices

In the last section, we identified an object with a list of characteristics and we showed how objects can be grouped into elementary aggregates according to their characteristics. The economic foundation of this approach is the consumption theory originally introduced by Lancaster (1971). It is based on the assumption that the behaviour of economic agents towards consumption goods is completely linked to their characteristics. More precisely, it is assumed that 'one demands not just physical objects, but the qualities with which they are endowed' (Milgate, 1987, p. 546). The consumers' preferences are therefore originally directed towards the characteristics of an object, and the latter determine eventually the consumers' preference structure between individual items of an elementary aggregate.

One aspect already discussed by Lancaster (1971, p. 140 ff.) himself, however, is the fact that there are characteristics (such as the serial number of a car) that are usually irrelevant for a consumer's decision of purchase. Irrelevant characteristics are especially those which are invariant, i.e. which have the same value, for all items of an elementary aggregate. Inversely, he defines a characteristic as *relevant* when ignoring it would change the preference structure between two items.

Lancaster's approach finally suggests that an object's price observed on the market is essentially determined by its relevant characteristics. This assumption is called *hedonic hypothesis* in the literature (Triplett, 1987; United Nations, 1993; Dickie et al., 1997; Brachinger, 2002). The hedonic hypothesis serves as the general basis for all hedonic elementary price indices. In order to build up a solid theory of hedonic elementary price indices, it is necessary to formulate this hypothesis as an econometric model.

14 Hedonic econometric model. *Let $\mathcal{G}$ be an elementary aggregate with distinguishing characteristics $\mathcal{M}_{\mathcal{G}}$. There exists a finite set of characteristics*

$$\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}} = \{m_1, \ldots, m_{K_{\mathcal{G}}}\} \subset \mathcal{M}_{\mathcal{G}}$$

and a function $h_{\mathcal{G}} : \mathbb{R}^{K_{\mathcal{G}}} \longrightarrow \mathbb{R}_{\geq 0}$, such that the price $p(o)$ of any item $o \in \mathcal{G}$ can be written as

$$p(o) = h_{\mathcal{G}}\big(\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}(o)\big) + \epsilon(o) \tag{2.2}$$

with $\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}(o) = (m_1(o), \ldots, m_{K_{\mathcal{G}}}(o))'$. The residual term $\epsilon(o)$ is assumed to be stochastic with expectation

$$\mathrm{E}_{[o]}\big(\epsilon(o)\big) = 0 \tag{2.3}$$

where $[o]$ denotes the $\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}$-equivalence class to which o belongs, $[o] \in \mathcal{G}/\sim_{\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}}$, and $\mathrm{E}_{[o]}$ the expectation built over all items belonging to $[o]$. Moreover, for any pair $o_i, o_j \in \mathcal{G}$, it is assumed that their residual terms $\epsilon(o_i)$ and $\epsilon(o_j)$ satisfy

$$\mathrm{E}_{[o_i,o_j]}\big(\epsilon(o_i)\,\epsilon(o_j)\big) = 0 \tag{2.4}$$

where $\mathrm{E}_{[o_i,o_j]}$ denotes the expectation built over all pairs of items (o_i', o_j') with $o_i' \in [o_i]$ and $o_j' \in [o_j]$.

*The set $\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}}$ will be called the **set of price-relevant characteristics**, and $h_{\mathcal{G}}$ is the **hedonic function** of $\mathcal{G}$.* ◇

This model implements the idea that for an elementary aggregate where the hedonic hypothesis holds, the set of distinguishing characteristics contains a finite subset of price-relevant characteristics. They determine the price up to a residual term which covers the quality-independent price component. The subset of price-relevant characteristics represents an item's *quality*. Assumption (2.3) implies that the hedonic elementary price of an item with a certain quality is given by the average price over all items of the same quality. Assumption (2.4) means that any two residual terms, i.e. any two quality-independent price components are uncorrelated.

The term 'quality' subsumes in this context all characteristics of a product which make it distinguishable from other products from an economic point of view (see e.g. United Nations, 1993, para. 16.105 ff.). Quality differences, however, are not only attributable to differences in the *physical* characteristics of the items. Additionally, there are, for instance, differences in the conditions of sale (guarantee, after-sales service etc.) as well as the location and the timing of delivery, which may lead to unequal qualities. The reason for this is that the marginal utility of a particular item for a

consumer may depend on these circumstances. 'Transporting a good to a location in which it is in greater demand is a process of production in its own right in which the good is transformed into a higher quality good' (United Nations, 1993, para. 16.107).

It is obvious that, for example, electricity or transport provided at peak times represent another quality than the same goods provided at off-peak times. 'The fact that peaks exist shows that purchasers or users attach greater utility to the services at these times, while the marginal costs of production are usually higher at peak times' (United Nations, 1993, para. 16.108). Changes in quality are reflected by changes in price for the different times of delivery. Therefore, 'time of delivery' is susceptible to be a price-relevant characteristic for the respective elementary aggregates.

Conversely, an observed price difference between physically identical products normally reflects a difference in quality in one of the senses described above. Rationally behaving consumers would not be willing to pay a higher price for a product otherwise. However, there are essentially three situations where this statement does not hold (cf. United Nations, 1993, para. 16.112 ff.). First, it is quite common that consumers are not sufficiently informed about the different sales prices of one and the same product at different outlets. Searching for the lowest price would cost them more than the expected price reduction. Second, a regime of price discrimination by the suppliers may lead to a situation where not every consumer is able to buy the same service (and the same quality) at the same price. A prominent example of such a regime is price discounts for students or pensioners, e.g. for cultural manifestations. It is, finally, thinkable that a specific product might be supplied simultaneously at two different prices but there is insufficient supply for the lower-priced variant. This situation may occur when there are parallel markets supplying the product, and one of them is, for instance, subject to official control. As a summary, the price of an item should reflect its quality in most of the cases but there may be exceptions to this rule. These exceptions are modelled by the residual $\epsilon(o)$ in the hedonic econometric model.

It is important to note that the hedonic function implicitly contains information on the market structure or the competitive situation for a certain good. The value of a hedonic function is the average price a consumer needs to pay for an item of a specific quality. In this sense, the hedonic function $h_{\mathcal{G}}$ is defined on the quotient set $\mathcal{G}/\sim_{\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}}$ and attributes a price to each of these equivalence classes. The assumption (2.3) is reasonable if all the equivalence classes $[o]$ are sufficiently homogeneous. The more price-relevant

characteristics there are, the more homogeneous the individual equivalence classes are. By adding additional characteristics, the homogeneity of these classes may be increased and the assumption becomes more realistic.

A microeconomic interpretation of the hedonic function is given, for instance, by Diewert (2003a, p. 321 ff.) going back to Rosen (1974). Starting from a list of assumptions on the consumer's utility functions, he develops a willingness-to-pay value function that indicates the amount of money the consumer is willing to pay for an item with any particular characteristics vector. This eventually leads to a special case of the hedonic equation (2.2).

CHAPTER 3

Hedonic Elementary Price Indices

3.1 Elementary price indices and quality change

3.1.1 Bilateral comparison of time periods

Elementary price indices measure the price evolution within an elementary aggregate between two time periods. There is a base period denoted by 0 which serves as a reference and where any index value is usually normalised to 1 (or 100), and there is a current period denoted by 1 which is confronted to the base period. If the index takes a value larger or smaller than 1 (or 100), this indicates that the prices of the items of the elementary aggregate in the current period are generally higher or lower than in the base period, respectively. A value of 1 (or 100) shows that the price level in the base and current period is the same.

All the concepts introduced here will be focused on the bilateral comparison of two time periods. This is the usual approach for elementary indices. Index values of subsequent current periods are going to be constructed from individual bilateral comparisons where the base period remains fixed. In order to bring in the time dimension and to define elementary price indices, we are now going to extend the framework established in the previous chapter.

Each elementary price index refers to an elementary aggregate $\mathcal{G} = \mathcal{G}^0$ at the base period 0. This aggregate varies over time: new items appear on the market, others disappear. At any time t, the aggregate of interest consists of all items available on the market characterised by the distinguishing characteristics of $\mathcal{G}^0$. This leads to the following definition.

18 **Definition 3.1.** *Let $\mathcal{G} = \mathcal{G}^0$ be any elementary aggregate defined relative to the set $\mathcal{O}^0$ of all objects supplied on the market at a base period 0 and let $\mathcal{M}_\mathcal{G}$ be its set of distinguishing characteristics. Let $\mathcal{T}$ be the set of all time periods considered and let $\mathcal{O}^t$ denote the set of all objects supplied on the market at time t.*

Then, for any time $t \in \mathcal{T}$, the ***current aggregate*** *$\mathcal{G}^t = \mathcal{G}^t(\mathcal{M}_\mathcal{G})$ is defined as the elementary aggregate induced by $\mathcal{M}_\mathcal{G}$ on $\mathcal{O}^t$, i.e.*

$$\mathcal{G}^t(\mathcal{M}_\mathcal{G}) := \bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}^t_m \,, \tag{3.1}$$

where $\mathcal{O}^t_m \subset \mathcal{O}^t$ denotes the domain of characteristic m at period t. Moreover, the ***composite elementary aggregate*** *$\mathcal{G}^\mathcal{T}$ for all time periods in $\mathcal{T}$ is defined by*

$$\mathcal{G}^\mathcal{T} := \bigcup_{t \in \mathcal{T}} \mathcal{G}^t(\mathcal{M}_\mathcal{G}) \,. \tag{3.2}$$

◇

Obviously, by means of (3.2), any elementary aggregate $\mathcal{G}$ defined relative to a base period induces a composite elementary aggregate $\mathcal{G}^\mathcal{T}$ for any set $\mathcal{T}$ of time periods. This viewpoint allows the identification of items that newly appeared between a base period 0 and a current period 1 by the difference $\mathcal{G}^1 \setminus \mathcal{G}^0 \subset \mathcal{G}^{\{0,1\}}$. Conversely, the items that disappeared between 0 and 1 can be specified as $\mathcal{G}^0 \setminus \mathcal{G}^1$. The intersection $\mathcal{G}^0 \cap \mathcal{G}^1$ represents all the items available in both periods.

For the specification of an elementary price index, there are basically two concurrent approaches. The first is to form the ratio of the average price level in the current period to the one in the base period. Conversely, the second approach consists of forming an average of the price ratios (or price relatives) of individual items of the elementary aggregate. Both approaches are justified, and hence we are going to investigate both of them in the following paragraphs.

3.1.2 Elementary price index concepts

The first of the two approaches for specifying an elementary price index compares the average price of the elementary aggregate in the current period with the one in the base period. If the average price of the elementary aggregate $\mathcal{G}^t$ at time t is denoted by $\pi^t(\mathcal{G}^t)$, such an elementary price index has the form

$$EPI^{0:1}(\mathcal{G}) = \frac{\pi^1(\mathcal{G}^1)}{\pi^0(\mathcal{G}^0)} \,. \tag{3.3}$$

Depending on the definition of $\pi^t(\mathcal{G}^t)$, this results in different possible implementations of an elementary price index.

In general, the average price $\pi^t(\mathcal{G})$ of any subset $\mathcal{G}$ of the elementary aggregate $\mathcal{G}^t$ may be viewed as a measure of location μ characterising the distribution of the prices at time t of the elements in $\mathcal{G}$. More formally, this implies defining

$$\pi^t(\mathcal{G}) := \mu\big(p^t(\mathcal{G})\big), \tag{3.4}$$

where $p^t(\mathcal{G}) := \bigcup_{o\in\mathcal{G}}\{p^t(o)\}$ and $p^t(o)$ is the observed price of o at time t. The set function $\mu : \boldsymbol{\mathcal{P}}(\mathbb{R}_{\geq 0}) \longrightarrow \mathbb{R}_{\geq 0}$, where $\boldsymbol{\mathcal{P}}(\mathbb{R}_{\geq 0})$ denotes the power set of $\mathbb{R}_{\geq 0}$, may attribute different weights to the individual prices in $p^t(\mathcal{G})$, according to the frequency they occur with among the items of $\mathcal{G}$.

The second approach consists of establishing an elementary price index as an average of price ratios. Formally, this leads to elementary price indices of the form

$$EPI^{0:1}(\mathcal{G}) = \mu\left(\bigcup_{o\in\mathcal{G}^0\cap\mathcal{G}^1}\left\{\frac{p^1(o)}{p^0(o)}\right\}\right), \tag{3.5}$$

where μ, again, is an appropriate measure of location characterising the distribution of price relatives.

3.1.3 Measuring the pure price change

Regardless of the approach that is followed, there is the difficulty of separating two different potential sources of an observed price change: one that is due to inflation and another that is due to any change in quality. An elementary price index is meant to measure only the former of the two, namely the 'pure' price change for a population of constant quality.

An important problem of the index concept (3.3) is that the elementary aggregates involved in both the numerator and the denominator of the ratio differ. This means that, in general, the quality range of the items considered for constructing the index changes between the base and the current period. The observed change of the average price level may thus be partly due to the change in quality of the objects provided on the market and not only due to inflation.

In order to build a quality-adjusted elementary price index, it is therefore necessary to restrict the set of objects involved to those that are available in both time periods. In other words, the proper specification of an elementary price index as a ratio of average prices is

$$EPI^{0:1}(\mathcal{G}) = \frac{\pi^1(\mathcal{G}^0\cap\mathcal{G}^1)}{\pi^0(\mathcal{G}^0\cap\mathcal{G}^1)} \tag{3.6}$$

rather than (3.3). By restricting the range of items over which the averages are taken, one ensures that the quality range of the items considered is stable over time and that the elementary price index does not suffer from a bias due to quality-driven price changes.

The most extreme special case of this approach is the one where μ takes the price of a single item $o^* \in \mathcal{G}^0 \cap \mathcal{G}^1$ as a representative for the whole elementary aggregate $\mathcal{G}^{\{0,1\}}$. In practice, o^* might be the item that generates the highest turnover or is sold most frequently on the market. Equation (3.3) becomes then just

$$EPI^{0:1}(\mathcal{G}) = \frac{\pi^1(\{o^*\})}{\pi^0(\{o^*\})} = \frac{p^1(o^*)}{p^0(o^*)}\,. \tag{3.7}$$

When the elementary price index is defined as an average of price relatives as in (3.5), the concept itself requires that the range of the items considered is limited to $\mathcal{G}^0 \cap \mathcal{G}^1$, since it is only for these items of the composite elementary aggregate $\mathcal{G}^{\{0,1\}}$ that prices are observable for both time periods. Both elementary price index concepts suffer thus from the restriction that, in order to measure inflation without any bias due to quality change, all the items considered need to be available in both the base and the current time periods. However, the set of items $\mathcal{G}^0 \cap \mathcal{G}^1$ may be small or even empty. The items contained therein may thus be unable to represent well the range of items available in either the base or the current period, i.e. of items of the composite elementary aggregate $\mathcal{G}^{\{0,1\}}$.

In order to relax the mentioned restriction and ideally to be able to average over all the items in $\mathcal{G}^{\{0,1\}}$, i.e. over those being available at least in one of the two periods, it is necessary that prices can be estimated also for items that are not contained in $\mathcal{G}^0 \cap \mathcal{G}^1$. With the aim of providing such price estimates, a quantity of conventional methods of quality adjustments has been developed. We are now going to recall shortly the most important of them.

3.1.4 Conventional methods of quality adjustment

The so-called *direct comparison* or *comparable replacement* method identifies a new item $o_{\text{new}} \in \mathcal{G}^1 \setminus \mathcal{G}^0$ in the current period with a very similar item $o \in \mathcal{G}^0$ in the base period. Any price difference between o and o_{new} is assumed to arise from true inflation and not from quality change. The unobservable price of o_{new} in period 0 is then *defined* as $p^0(o_{\text{new}}) := p^0(o)$, and

the respective ratio representing the price change of this item from period 0 to 1 is calculated by

$$\frac{p^1(o_{\text{new}})}{p^0(o_{\text{new}})} = \frac{p^1(o_{\text{new}})}{p^0(o)}\,. \tag{3.8}$$

A similar approach is taken in the *overlap pricing* method, where, again, a new item $o_{\text{new}} \in \mathcal{G}^1 \setminus \mathcal{G}^0$ is matched with an item $o \in \mathcal{G}^0$. This time, however, the unobservable price $p^0(o_{\text{new}})$ is defined by $p^0(o_{\text{new}}) := k(o, o_{\text{new}})\, p^0(o)$ with the correction factor

$$k(o, o_{\text{new}}) := \frac{p^t(o_{\text{new}})}{p^t(o)}$$

in some intermediate time period $t \in [0, 1]$ where both objects o and o_{new} are available. The respective price relative of o_{new} is then given by

$$\frac{p^1(o_{\text{new}})}{p^0(o_{\text{new}})} = \frac{p^1(o_{\text{new}})}{p^t(o_{\text{new}})}\,\frac{p^t(o)}{p^0(o)}\,. \tag{3.9}$$

In the *explicit quality adjustment* method, a new o_{new} is matched to an old object o just as in the previous methods, but the correction factor $k(o, o_{\text{new}})$ is estimated by other means than a ratio of observed prices. Different alternative methods for estimating $k(o, o_{\text{new}})$, such as expert judgments, quantity adjustments or even hedonic regressions, are described in detail in the literature (ILO et al., 2004, para. 7.72 ff.). 'Hedonic regressions', in this context, must not be confused with proper hedonic elementary price indices as they will be introduced below.

If all of the new items in $\mathcal{G}^1 \setminus \mathcal{G}^0$ are linked in one of these ways to old items in $\mathcal{G}^0$, then a (potentially hypothetical) base period price p^0 is available for any element in $\mathcal{G}^1$. This allows for an inclusion of the prices of all elements of $\mathcal{G}^1$ in (3.5) or (3.6), already. If, additionally, the same methods of quality adjustment are applied to the items that have disappeared between the base and the current period (i.e. the items in $\mathcal{G}^0 \setminus \mathcal{G}^1$), the range over which the averages are taken can potentially be extended to the whole $\mathcal{G}^{\{0,1\}}$.

3.1.5 Summary and critical comments

The elaboration above has shown the main properties and issues of elementary price indices. We discussed that there are basically two approaches for measuring the price evolution within an elementary aggregate. The key lesson is that there needs to be a *constant reference quality* that is present in both the numerator and the denominator of any price ratio used to calculate the elementary price index, be it the ratio of average prices for the whole aggregate

or the price ratio for an individual item. In the raw index concepts introduced above, this reference quality is represented by the set of items from $\mathcal{G}^0 \cap \mathcal{G}^1$ that are considered for the calculations.

For some specific elementary aggregates, especially for those subject to rapid technological progress, the list of items available in both the base and the current periods may be too small to serve as the implementation of reference quality. An elementary price index is expected to measure the price evolution for the whole quality spectrum of items belonging to the elementary aggregate at any time period and not only for those that stay on the market for a sufficiently long period.

The conventional methods of quality adjustment provide tools that allow to impute unobservable prices, in order to augment the admissible set of items used as reference quality range. They are, however, inherently arbitrary, and there is no common methodological approach that covers all of them. They are *ad hoc* solutions that may or may not work for each individual item of an elementary aggregate. Based on the hedonic econometric model introduced in Section 2.3, we are now going to develop a more sophisticated solution to the issues mentioned so far.

3.2 Hedonic elementary price indices

3.2.1 The hedonic econometric model revisited

The hedonic econometric model establishes a relationship between characteristics and prices of the items of an elementary aggregate. Now, it needs to be developed further in order to take account of the time dimension. It will be important for the following analysis that the domain of the hedonic function does not change over time. Therefore, we will assume that the set $\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}}$ of price-relevant characteristics remains the same for all time-periods considered.

Hedonic econometric model (in time). *Let $\mathcal{G}^{\mathcal{T}}$ be a composite elementary aggregate for a given set of time periods $\mathcal{T}$ with distinguishing characteristics $\mathcal{M}_{\mathcal{G}}$. There exists a finite set of characteristics*

$$\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}} = \{m_1, \ldots, m_{K_{\mathcal{G}}}\} \subset \mathcal{M}_{\mathcal{G}}$$

and for each period $t \in \mathcal{T}$ a function $h_{\mathcal{G}}^t : \mathbb{R}^{K_{\mathcal{G}}} \longrightarrow \mathbb{R}_{\geq 0}$, such that the price $p^t(o)$ of any item $o \in \mathcal{G}^t$ available at time t can be written as

$$p^t(o) = h_{\mathcal{G}}^t\big(\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}(o)\big) + \epsilon^t(o) \tag{3.10}$$

with $\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}(o) = (m_1(o), \ldots, m_{K_{\mathcal{G}}}(o))'$. *For all* $t \in \mathcal{T}$, *the residual term* $\epsilon^t(o)$ *is assumed to be stochastic with expectation*

$$\mathrm{E}_{[o]}(\epsilon^t(o)) = 0 \tag{3.11}$$

where $[o]$ *denotes the* $\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}$*-equivalence class to which* o *belongs,* $[o] \in \mathcal{G}^t/\sim_{\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}}$, *and* $\mathrm{E}_{[o]}$ *the expectation built over all items belonging to* $[o]$. *Moreover, for any pair* $o_i, o_j \in \mathcal{G}$ *it is assumed that their residual terms* $\epsilon^t(o_i)$ *and* $\epsilon^t(o_j)$ *satisfy*

$$\mathrm{E}_{[o_i,o_j]}\big(\epsilon^t(o_i)\,\epsilon^t(o_j)\big) = 0 \tag{3.12}$$

for all $t \in \mathcal{T}$, *where* $\mathrm{E}_{[o_i,o_j]}$ *denotes the expectation built over all pairs of items* (o_i', o_j') *with* $o_i' \in [o_i]$ *and* $o_j' \in [o_j]$. ◇

Again, the hedonic function $h_{\mathcal{G}}^t$ for time $t \in \mathcal{T}$ is defined on the quotient set $\mathcal{G}^t/\sim_{\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}}$ of the current elementary aggregate $\mathcal{G}^t$ with respect to the equivalence relation $\sim_{\boldsymbol{m}_{\mathcal{G}}^{\mathrm{pr}}}$. Each equivalence class of items of the same quality is thus mapped at any time t by $h_{\mathcal{G}}^t$ to a certain price value.

The time-dependency of the hedonic function $h_{\mathcal{G}}^t$ is one of the key elements of this reformulation of the hedonic econometric model. The hedonic function attributes a price to each quality, and it is important to recognise that this price may change over time. The quality-adjusted price evolution for an elementary aggregate where the hedonic hypothesis holds is thus fully represented by the evolution of the hedonic function. In other words, an appropriate comparison of the hedonic functions in a base and a current period may thus be seen as an implementation of an elementary price index.

The four key elements of quality-adjusted elementary price indices, namely the items of the elementary aggregates, their qualities, their prices, and time, are summarised in Fig. 3.1. Measurement or quantification leads from level I, the set of all objects, to level II, the set of characteristics vectors. These are then mapped to price values (level III) through hedonic functions that change over time. The triangles and bullets on the time line symbolise the different price distributions in the base and the current period, respectively.

3.2.2 Hedonic elementary price indices

The two elementary price indices concepts, along with the conventional methods for quality adjustment introduced in Section 3.1 incorporate already at least implicitly the notion of quality. Some of the adjustment methods, such as the approach using hedonic regressions for quality adjustments, may even explicitly use information contained in the characteristics of the considered items as a representation of their respective qualities. Incorporating

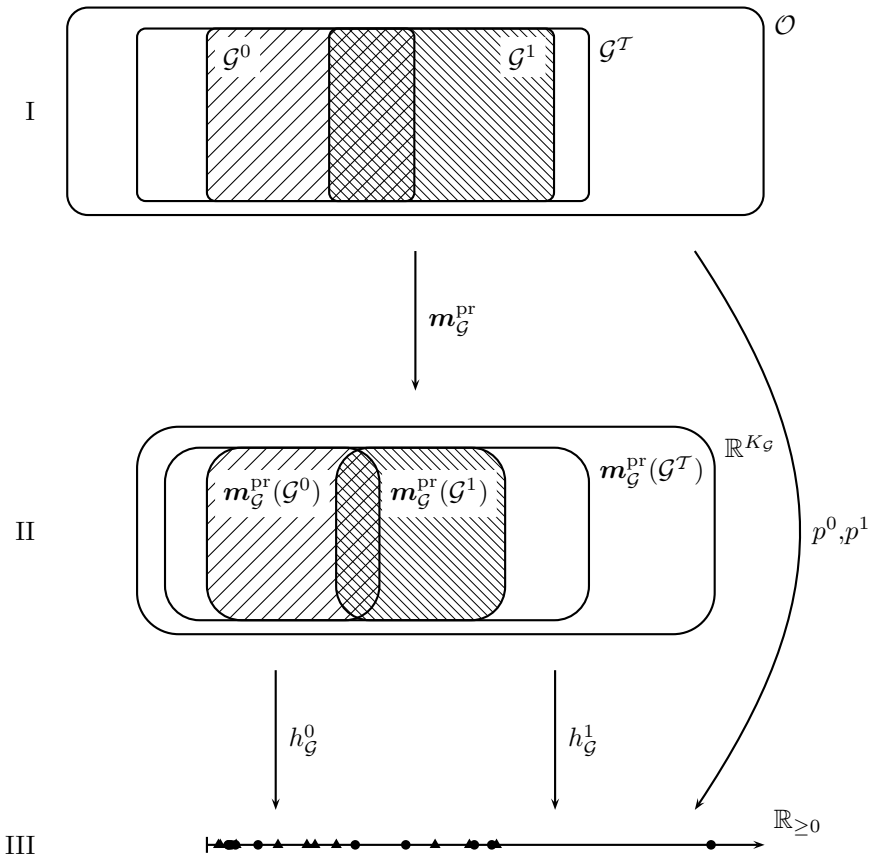


Figure 3.1: **Elementary aggregates, qualities, prices, and time—the four key elements of hedonic elementary price indices**

the space of characteristics as a third dimension besides objects and prices, however, allows for a more precise handling of quality.

The hedonic econometric model allows to define the average price $\pi^t(\mathcal{G}^t)$ of an elementary aggregate $\mathcal{G}^t$ on the basis of an average combination of characteristics. One might specify a representative characteristics vector $\boldsymbol{\mu}^*$, e.g. by

$$\boldsymbol{\mu}^* := \Big(\mu_1\big(m_1(\mathcal{G}^t)\big), \ldots, \mu_{K_\mathcal{G}}\big(m_{K_\mathcal{G}}(\mathcal{G}^t)\big)\Big)$$

where $m_k(\mathcal{G}^t) := \bigcup_{o\in\mathcal{G}^t}\{m_k(o)\}$ and where the functions $\mu_k : \boldsymbol{\mathcal{P}}(\mathbb{R}) \longrightarrow \mathbb{R}$ $(k = 1, \ldots, K_\mathcal{G})$ average individual characteristics over all items of the aggregate. The average price of the elementary aggregate $\mathcal{G}^t$ would consequently be defined as

$$\pi(\mathcal{G}^t) := h^t_\mathcal{G}(\boldsymbol{\mu}^*)\,, \tag{3.13}$$

provided that $h^t_\mathcal{G}$ is defined on the whole range of $\boldsymbol{\mu}^*$. Moreover, a *hedonic elementary price index* in the sense of both (3.3) and (3.5) can be defined by

$$HPI^{0:1}(\mathcal{G}) = \frac{h^1_\mathcal{G}(\boldsymbol{\mu}^*)}{h^0_\mathcal{G}(\boldsymbol{\mu}^*)} \tag{3.14}$$

using the two hedonic functions $h^0_\mathcal{G}$ and $h^1_\mathcal{G}$ in the base and in the current period respectively.

In this approach, the quality of the representative item considered is held constant by introducing the characteristics vector $\boldsymbol{\mu}^*$ in both the numerator and the denominator of the above price ratio. It is thus a generalisation of the index concepts introduced by Brachinger (2002). Here, $\boldsymbol{\mu}^*$ plays the role of the *constant reference quality* discussed in Section 3.2. From a technical point of view, this definition makes sense if $\boldsymbol{\mu}^*$ lies in $\boldsymbol{m}_\mathcal{G}(\mathcal{G}^0) \cap \boldsymbol{m}_\mathcal{G}(\mathcal{G}^1)$, i.e. in the domains of both $h^0_\mathcal{G}$ and $h^1_\mathcal{G}$. If this is not the case, a minimal requirement is that both hedonic functions can be extended to a domain including $\boldsymbol{\mu}^*$.

The disadvantage of (3.14) is that reference quality is represented only through one single characteristics vector. As we have already discussed, it is generally preferable to keep the variety of the considered items of the elementary aggregate intact in taking the whole quality spectrum as a reference for building an elementary price index. In this sense, one might imagine identifying a set $\mathcal{M}^* \subset \boldsymbol{m}_\mathcal{G}(\mathcal{G}^0) \cup \boldsymbol{m}_\mathcal{G}(\mathcal{G}^1)$ of representants from the characteristics space containing more than just one characteristics vector. After transformation through the hedonic functions $h^0_\mathcal{G}$ and $h^1_\mathcal{G}$, an appropriate

measure of location λ can be applied to reduce this variety to one single number, such as

$$HPI^{0:1}(\mathcal{G}) = \frac{\lambda(h^1_{\mathcal{G}}(\mathcal{M}^*))}{\lambda(h^0_{\mathcal{G}}(\mathcal{M}^*))} \tag{3.15}$$

with $h^t_{\mathcal{G}}(\mathcal{M}^*) := \bigcup_{\boldsymbol{m}\in\mathcal{M}^*}\{h^t_{\mathcal{G}}(\boldsymbol{m})\}$ for $t \in \{0,1\}$. The question raised above whether the hedonic functions need to be extended to a larger domain is also relevant here, since $\mathcal{M}^*$ contains in most cases more than $\boldsymbol{m}_{\mathcal{G}}(\mathcal{G}^0)\cap\boldsymbol{m}_{\mathcal{G}}(\mathcal{G}^1)$, i.e. it contains characteristics vectors that could not be observed in both time periods.

Equation (3.15) represents in the sense of (3.3) a ratio of means of imputed prices in two periods. Alternatively, one can also consider a particular average of price ratios as in (3.5). The corresponding hedonic elementary price index is

$$HPI^{0:1}(\mathcal{G}) = \lambda\left(\bigcup_{\boldsymbol{m}\in\mathcal{M}^*}\left\{\frac{h^1_{\mathcal{G}}(\boldsymbol{m})}{h^0_{\mathcal{G}}(\boldsymbol{m})}\right\}\right). \tag{3.16}$$

Obviously, (3.14) is a special case of both (3.15) and (3.16) for the case where $\mathcal{M}^* = \{\boldsymbol{\mu}^*\}$ and λ is just the identity function.

3.3 A stochastic approach to hedonic elementary price indices

3.3.1 The sampling approach It is obvious that in practice, the reference items used for measuring the price evolution of an elementary aggregate cannot cover the complete population (cf. ILO et al., 2004, para. 9.8). Although the number of purchases of objects during a certain time period on a market is finite, it is impossible to observe the whole universe. Any elementary price index, therefore, can only base on *samples of observations*. It needs to be considered as a stochastic variable or, in other words, 'as a function that transforms sample survey data into an index number' (Balk, 2005, p. 676). Moreover, one needs to distinguish between the elementary price index as a meta-parameter of populations of items in different time periods (i.e. the 'population index' in the terms of Dorfman et al., 1999) and its realisation for a specific data set.

Our preferred approach to hedonic elementary price indices will therefore be a stochastic one in which we are going to model the sampling procedure involved. Its philosophy is thus very close to the fairly recent expositions of Dorfman et al. (1999) or Silver and Heravi (2006). We imagine a random draw from all the items of an elementary aggregate $\mathcal{G}^t$ at time t

and denote by $\boldsymbol{M}^t$ the random vector of characteristics and P^t the random variable representing the price of the drawn item. Based on the hedonic econometric model, the relationship between $\boldsymbol{M}^t$ and P^t is given by

$$P^t = h^t_{\mathcal{G}}(\boldsymbol{M}^t) + \epsilon^t \tag{3.17}$$

where the random error ϵ^t has $\mathrm{E}\epsilon^t = 0$ for all $t \in \mathcal{T}$ and is assumed to be independent of $\boldsymbol{M}^t$. Within this additive error model, the hedonic function $h^t_{\mathcal{G}}$ therefore is exactly the conditional mean

$$h^t_{\mathcal{G}}(\boldsymbol{m}) = \mathrm{E}(P^t \,|\, \boldsymbol{M}^t = \boldsymbol{m})\,, \tag{3.18}$$

and the conditional distribution $\mathbb{P}(P^t \,|\, \boldsymbol{M}^t)$ depends on $\boldsymbol{M}^t$ only through $h^t_{\mathcal{G}}$ (see e.g. Hastie et al., 2001, p. 28).

The hedonic elementary price index concepts developed in the last section will now be translated into this new stochastic framework. The formula given in (3.14), for instance, can be written as the ratio

$$HPI^{0:1}(\mathcal{G}) = \frac{h^1_{\mathcal{G}}(\boldsymbol{\mu}^*)}{h^0_{\mathcal{G}}(\boldsymbol{\mu}^*)} = \frac{\mathrm{E}(P^1 \,|\, \boldsymbol{M}^1 = \boldsymbol{\mu}^*)}{\mathrm{E}(P^0 \,|\, \boldsymbol{M}^0 = \boldsymbol{\mu}^*)} \tag{3.19}$$

of the expected prices in the base and current period given that the characteristics equal the constant reference vector $\boldsymbol{\mu}^*$.

This reference characteristics vector can easily be seen as some kind of mean of the characteristics available on a market. If we imagine to have a random vector $\boldsymbol{M}$ representing the outcome of a random draw from a population of reference characteristics, $\boldsymbol{\mu}^*$ can be defined as the expectation of $\boldsymbol{M}$, i.e. $\boldsymbol{\mu}^* := \mathrm{E}\boldsymbol{M}$. In (3.19) as in (3.14), the fixed vector $\boldsymbol{\mu}^*$ plays the role of the constant reference quality within the elementary price index. Again, however, reference quality should embrace the whole quality spectrum of items within the elementary aggregate. This can be achieved in a very general way by identifying the whole distribution of $\boldsymbol{M}$ as reference quality. Any indicator depending only on the distribution of $\boldsymbol{M}$ can then serve for introducing reference quality into a hedonic elementary price index formula.

In the simplest case, where only the expectation of $\boldsymbol{M}$ is taken into consideration, the hedonic elementary price index can be defined by

$$HPI^{0:1}(\mathcal{G}) = \frac{h^1_{\mathcal{G}}(\mathrm{E}\boldsymbol{M})}{h^0_{\mathcal{G}}(\mathrm{E}\boldsymbol{M})} = \frac{\mathrm{E}(P^1 \,|\, \boldsymbol{M}^1 = \mathrm{E}\boldsymbol{M})}{\mathrm{E}(P^0 \,|\, \boldsymbol{M}^0 = \mathrm{E}\boldsymbol{M})}\,. \tag{3.20}$$

The distribution of $\boldsymbol{M}$, as we have just mentioned, represents the probability of drawing a certain characteristics vector from a specified reference

population. It remains an open question, how this reference population and thus the distribution of $\boldsymbol{M}$ should be specified. Supposing that $\boldsymbol{M} \overset{\text{L}}{\sim} \boldsymbol{M}^0$, i.e. when the population observed in the base period is defined as the reference population, we get the index

$$HPI^{0:1}(\mathcal{G}) = \frac{h^1_{\mathcal{G}}(\mathrm{E}\boldsymbol{M}^0)}{h^0_{\mathcal{G}}(\mathrm{E}\boldsymbol{M}^0)}\,,$$

introduced as *true hedonic Laspeyres price index* by Brachinger (2002, p. 6). Conversely, if one takes the population of the current period as reference quality, i.e. $\boldsymbol{M} \overset{\text{L}}{\sim} \boldsymbol{M}^1$, this approach leads to the *true hedonic Paasche price index*

$$HPI^{0:1}(\mathcal{G}) = \frac{h^1_{\mathcal{G}}(\mathrm{E}\boldsymbol{M}^1)}{h^0_{\mathcal{G}}(\mathrm{E}\boldsymbol{M}^1)}\,.$$

From a theoretical point of view, there is, however, no reason why the distribution of $\boldsymbol{M}$ should be restricted to either $\boldsymbol{M}^0$ or $\boldsymbol{M}^1$. One alternative, for instance, would be to define $\boldsymbol{M}$ as a mixture of the other two distributions, i.e.

$$\mathbb{P}_{\boldsymbol{M}} = g\,\mathbb{P}_{\boldsymbol{M}^0} + (1-g)\,\mathbb{P}_{\boldsymbol{M}^1}\,,$$

where $\mathbb{P}_{\boldsymbol{M}}$, $\mathbb{P}_{\boldsymbol{M}^0}$ and $\mathbb{P}_{\boldsymbol{M}^1}$ are the probability measures of $\boldsymbol{M}$, $\boldsymbol{M}^0$ and $\boldsymbol{M}^1$, respectively, and $g \in [0,1]$. Such a concept represents the practice of taking characteristics expressions from the populations in both the base and the current periods for building the population of reference quality. This is, for instance, the case for the *true hedonic adjacent periods price index* as outlined by Brachinger (2002, pp. 6–8).

Instead of implementing the reference quality just by the expectation of $\boldsymbol{M}$, it is preferable to use more information contained in the distribution of $\boldsymbol{M}$ for constructing hedonic elementary price indices. If we translate the definitions (3.15) and (3.16) into the current stochastic framework, we get for instance

$$HPI^{0:1}(\mathcal{G}) = \frac{\mathrm{E}h^1_{\mathcal{G}}(\boldsymbol{M})}{\mathrm{E}h^0_{\mathcal{G}}(\boldsymbol{M})} \tag{3.21}$$

and

$$HPI^{0:1}(\mathcal{G}) = \mathrm{E}\left[\frac{h^1_{\mathcal{G}}(\boldsymbol{M})}{h^0_{\mathcal{G}}(\boldsymbol{M})}\right]\,. \tag{3.22}$$

Again, both of these index definitions include (3.19) as a special case, namely when all the probability mass of $\mathbb{P}_{\boldsymbol{M}}$ is concentrated on $\boldsymbol{\mu}^*$.

It still needs to be explained how the expectations in (3.21) and (3.22) are to be interpreted. Following (3.18), we have

$$\begin{aligned} \mathrm{E}h^t_{\mathcal{G}}(\boldsymbol{M}) &= \mathrm{E}_{\boldsymbol{M}}(\mathrm{E}_{P^t \mid \boldsymbol{M}^t}(P^t \mid \boldsymbol{M})) \\ &= \int_{\mathbb{R}^{K_{\mathcal{G}}}} \left[\int_{\mathbb{R}} p \, \mathrm{d}\mathbb{P}_{P^t \mid \boldsymbol{M}^t}(p \mid \boldsymbol{m}) \right] \mathrm{d}\mathbb{P}_{\boldsymbol{M}}(\boldsymbol{m}) \,. \end{aligned} \tag{3.23}$$

for $t \in \{0,1\}$. Here, $\mathbb{P}_{P^t \mid \boldsymbol{M}^t}$ stands for the probability measure of the conditional distribution of P^t given $\boldsymbol{M}^t$, and $\mathrm{E}_{P^t \mid \boldsymbol{M}^t}$ is the expectation with respect to this probability measure. Moreover, $\mathbb{P}_{\boldsymbol{M}}$ is the probability measure respective to the distribution of $\boldsymbol{M}$, and $\mathrm{E}_{\boldsymbol{M}}$ is its expectation. In the special case where the random variables and vectors considered are continuous, we have

$$\mathrm{E}h^t_{\mathcal{G}}(\boldsymbol{M}) = \int_{\mathbb{R}^{K_{\mathcal{G}}}} \left[\int_{\mathbb{R}} p \, \frac{f_{(P^t, \boldsymbol{M}^t)}(p, \boldsymbol{m})}{f_{\boldsymbol{M}^t}(\boldsymbol{m})} \, \mathrm{d}p \right] f_{\boldsymbol{M}}(\boldsymbol{m}) \, \mathrm{d}\boldsymbol{m} \tag{3.24}$$

with $f_{(P^t, \boldsymbol{M}^t)}$ being the common probability density of P^t and $\boldsymbol{M}^t$, $f_{\boldsymbol{M}^t}$ the marginal density of $\boldsymbol{M}^t$ and, finally, $f_{\boldsymbol{M}}$ the density of $\boldsymbol{M}$.

It can be seen that for this definition to hold, the support of $f_{\boldsymbol{M}}$ needs to be contained in the support of $f_{\boldsymbol{M}^t}$ for $t \in \{0,1\}$. In other words, for each vector $\boldsymbol{m} \in \mathbb{R}^{K_{\mathcal{G}}}$ with $f_{\boldsymbol{M}^t}(\boldsymbol{m}) = 0$, we need to have $f_{\boldsymbol{M}}(\boldsymbol{m}) = 0$. This has to be taken into consideration when the reference quality $\boldsymbol{M}$ is chosen. In particular, $\mathbb{P}_{\boldsymbol{M}}$ must not attribute a positive probability to any set of characteristics vectors that does not have a positive probability with respect to $\mathbb{P}_{\boldsymbol{M}^0}$ and $\mathbb{P}_{\boldsymbol{M}^1}$ as well, i.e. within the populations available in both the base and current period.

In practice, therefore, it is useful to assume that $\mathbb{P}_{\boldsymbol{M}^0}$ and $\mathbb{P}_{\boldsymbol{M}^1}$ attribute a positive probability to any non-discrete set of vectors in the characteristics space, i.e. the cartesian product of the ranges of all price-relevant characteristics. This ensures that out-of-sample-prediction is possible, and thus there are no formal restrictions on the distribution of reference characteristics $\boldsymbol{M}$.

Besides, it can be noted that whenever $\boldsymbol{M} \perp\!\!\!\perp \boldsymbol{M}^t$, by the law of iterated expectations, we have

$$\mathrm{E}_{\boldsymbol{M}}(\mathrm{E}_{P^t \mid \boldsymbol{M}^t}(P^t \mid \boldsymbol{M})) = \mathrm{E}P^t \,.$$

If the reference distribution $\boldsymbol{M}$ is chosen following the Paasche principle, i.e. if $\boldsymbol{M} \perp\!\!\!\perp \boldsymbol{M}^1$, equation (3.21) becomes

$$HPI^{0:1}(\mathcal{G}) = \frac{\mathrm{E}h^1_{\mathcal{G}}(\boldsymbol{M}^1)}{\mathrm{E}h^0_{\mathcal{G}}(\boldsymbol{M}^1)} = \frac{\mathrm{E}P^1}{\mathrm{E}_{\boldsymbol{M}^1}(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M}^1))} \,.$$

On the other hand, if we follow the principle of Laspeyres ($\boldsymbol{M} \overset{\text{L}}{\sim} \boldsymbol{M}^0$) we get the index

$$HPI^{0:1}(\mathcal{G}) = \frac{\mathrm{E}h^1_{\mathcal{G}}(\boldsymbol{M}^0)}{\mathrm{E}h^0_{\mathcal{G}}(\boldsymbol{M}^0)} = \frac{\mathrm{E}_{\boldsymbol{M}^0}(\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M}^0))}{\mathrm{E}P^0}\,.$$

The hedonic index defined in (3.22) can similarly to (3.23) be written as

$$HPI^{0:1}(\mathcal{G}) = \mathrm{E}_{\boldsymbol{M}}\left[\frac{\mathrm{E}_{P^1\,|\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\right]\,.$$

or, in the special case of continuous characteristics variables, as

$$HPI^{0:1}(\mathcal{G}) = \int_{\mathbb{R}^{K_{\mathcal{G}}}} \frac{\int_{\mathbb{R}} p\,\frac{f_{P^1,\boldsymbol{M}^1}(p,\boldsymbol{m})}{f_{\boldsymbol{M}^1}(\boldsymbol{m})}\,\mathrm{d}p}{\int_{\mathbb{R}} p\,\frac{f_{P^0,\boldsymbol{M}^0}(p,\boldsymbol{m})}{f_{\boldsymbol{M}^0}(\boldsymbol{m})}\,\mathrm{d}p}\, f_{\boldsymbol{M}}(\boldsymbol{m})\,\mathrm{d}\boldsymbol{m}\,.$$

The requirements mentioned above concerning the support of $f_{\boldsymbol{M}}$ still apply in this situation.

3.3.2 Universal formulae for hedonic elementary price indices

In the previous paragraphs, we proposed several alternative definitions of a bilateral hedonic elementary price index. We are now going to show that, under certain conditions and with a small modification, the definitions (3.21) and (3.22) coincide with their non-stochastic counterparts (3.15) and (3.16). This will allow us to limit further investigations to the stochastic versions.

Let us thus look at (3.21) first. Assume that $\mathbb{P}_{\boldsymbol{M}}$ is the empirical probability distribution of any finite reference population $\mathcal{M}^*$. In that case, $\mathrm{E}h^t_{\mathcal{G}}(\boldsymbol{M})$ is exactly the arithmetic mean of the elements in $h^t_{\mathcal{G}}(\mathcal{M}^*)$, $t \in \{0,1\}$, proportionally weighing the price values in $h^t_{\mathcal{G}}(\mathcal{M}^*)$ according to the number of elements contained in their preimage under $h^t_{\mathcal{G}}$. As a matter of fact, (3.15) is thus a special case of (3.21) if λ is the arithmetic mean, and the same holds for (3.16) and (3.22).

Admitting only arithmetic means for λ seems, however, far too restrictive in a price index context. The widespread application of geometric means for aggregating price ratios to elementary indices suggests more general index definitions than those given in (3.21) and (3.22).

In this sense, two generalised definitions of a hedonic elementary price index are

$$HPI^{0:1}(\mathcal{G}) = \frac{\varphi^{-1}(\mathrm{E}\varphi(h^1_{\mathcal{G}}(\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}\varphi(h^0_{\mathcal{G}}(\boldsymbol{M})))} \tag{3.25}$$

or

$$HPI^{0:1}(\mathcal{G}) = \varphi^{-1}\left(\mathrm{E}\left[\varphi\left(\frac{h^1_{\mathcal{G}}(\boldsymbol{M})}{h^0_{\mathcal{G}}(\boldsymbol{M})}\right)\right]\right) \tag{3.26}$$

where φ is a continuous and injective function which maps a connected subset of $\mathbb{R}$ to $\mathbb{R}$. Taking $\varphi = \ln$ and for $\mathbb{P}_{\boldsymbol{M}}$ the empirical distribution of $\mathcal{M}^*$ would thus turn (3.16) with λ being the geometric mean into a special case of (3.26), for instance. This same procedure works for λ being the arithmetic or harmonic mean (by taking $\varphi(x) = x$ or $\varphi(x) = 1/x$ respectively), but it is not directly applicable to more complicated location measures λ. The formulae (3.25) and (3.26) therefore are still not universal enough to embrace fully the definitions in (3.15) and (3.16), but they are able to model the most important cases.

It should be noted, finally, that both definitions, (3.25) and (3.26), coincide if $\varphi = \ln$, just as (3.15) and (3.16) coincide if λ is the geometric mean. This is due to the linearity of the expectation and the properties of the natural logarithm.

CHAPTER 4

An Axiomatic Approach to Hedonic Elementary Price Indices

4.1 Bilateral hedonic elementary price indices

In the following paragraphs, the stochastic approach to hedonic elementary price indices introduced in the previous chapter is going to be taken as a base for further investigations. A hedonic elementary price index is assumed to be a certain function of different probability distributions. There are, first of all, the price distributions in both the base and the current periods, represented by the random variables P^0 and P^1, for the composite elementary aggregate $\mathcal{G}^{\{0,1\}}$, which we will denote from now on just by $\mathcal{G}$ for the sake of brevity. Secondly, along with the price distributions, there are the random vectors $\boldsymbol{M}^0$ and $\boldsymbol{M}^1$, representing the respective distributions of characteristics expressions. Finally, there is the random vector $\boldsymbol{M}$ which incorporates the reference quality distribution.

Each of these five random quantities may be seen as a function

$$X \;:\; \begin{array}{ccc} \mathcal{G} & \longrightarrow & \mathbb{R}_{\geq 0} \text{ or } \mathbb{R}^{K_{\mathcal{G}}} \\ o & \longmapsto & X(o) \end{array} .$$

Here, $\mathcal{G}$ itself is the space of events relevant to the experiment being performed as it contains all the hypothetical objects for which prices and characteristics are to be observed. As $\mathcal{G}$ is finite, the power set $\boldsymbol{\mathcal{P}}(\mathcal{G})$ forms a σ-algebra, and a probability space can be established by adding an appropriate probability measure on $\boldsymbol{\mathcal{P}}(\mathcal{G})$. There are, actually, three probability

measures of predominant interest, namely one for the base and one for the current period price and characteristics distribution as well as one that generates the distribution of reference quality.

A general bilateral hedonic elementary price index can now be defined as follows.

Definition 4.1. *A **bilateral hedonic elementary price index** is a functional ϕ of the form*

$$\begin{array}{rcccl} \phi & : & \mathbb{R}_{\geq 0}^{\mathcal{G}} \times \mathbb{R}_{\geq 0}^{\mathcal{G}} \times (\mathbb{R}^{K_{\mathcal{G}}})^{\mathcal{G}} \times (\mathbb{R}^{K_{\mathcal{G}}})^{\mathcal{G}} \times (\mathbb{R}^{K_{\mathcal{G}}})^{\mathcal{G}} & \longrightarrow & \mathbb{R}_{\geq 0} \\ & & (P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) & \longmapsto & \phi(\dots) \end{array} ,$$

where $\mathbb{R}_{\geq 0}^{\mathcal{G}}$ and $(\mathbb{R}^{K_{\mathcal{G}}})^{\mathcal{G}}$ are the sets of all random variables or vectors defined on $\mathcal{G}$ with values in $\mathbb{R}_{\geq 0}$ and $\mathbb{R}^{K_{\mathcal{G}}}$, respectively. ◇

It is important to note that a hedonic index $HPI^{0:1}(\mathcal{G})$ is not a random variable itself but takes probability distributions as arguments and yields a positive real number as a result, just as the mean operator does for individual random variables.

It has been a tradition in the price index literature for a long time to propose desirable properties for index formulae in the form of so-called *tests* or *axioms*. An extensive overview of this axiomatic approach to index number theory was recently given by ILO et al. (2004, Chap. 16). Similar treatises have been written, amongst others, by Balk (1995) as well as Vogt and Barta (1997), for instance. They all refer predominantly to previous works by Fisher (1922), Eichhorn (1976, 1978), Eichhorn and Voeller (1976), Diewert (1993, 1999), and others cited therein.

We are now going to follow an approach similar to the one taken in ILO et al. (2004, para. 20.58 ff.), which itself follows, among others, Eichhorn (1978, pp. 152–60) and Diewert (1995, pp. 5–17), in order to develop axioms for hedonic elementary price indices.

4.2 Axioms for hedonic elementary price indices

One of the most fundamental properties an index should have is its continuity with respect to its arguments. It is, however, one of the most difficult to translate into the current setting, where the arguments of the index are random variables. For sequences of random variables, there are different notions of convergence. We are now going to propose convergence in the $\mathbf{L}_2$ Hilbert space consisting of all equivalence classes of $\mathbb{R}$-valued random variables that have finite second moments (see e.g. Fristedt and Gray,

1997) as a base for the continuity axiom. Since $\mathcal{G}$ is finite, P^0 and P^1 are both elements of $\mathbf{L}_2$.

Test 1 (Continuity). *ϕ is continuous in the two price variables, i.e. if $(P_n^0 : n = 1, 2, \dots)$ and $(P_n^1 : n = 1, 2, \dots)$ are sequences of random variables in $\mathbf{L}_2$ that converge to P^0 and P^1, respectively, then*

$$\lim_{n\to\infty} \phi(P_n^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})$$

and

$$\lim_{n\to\infty} \phi(P^0, P_n^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})$$

for all $\boldsymbol{M}^0$, $\boldsymbol{M}^1$ and $\boldsymbol{M}$. ◇

It remains an open question if continuity of the index should also be required with respect to the distributions of $\boldsymbol{M}^0$, $\boldsymbol{M}^1$, and $\boldsymbol{M}$. Whereas it might be possible to define a sensible notion of convergence also for multivariate random variables, it is unclear how continuity could be proved for indices such as (3.25) or (3.26) where these three distributions of characteristics expressions are essentially used for conditioning expectations. Further research needs to clarify this issue.

A second basic property of any price index is the fact that if neither prices nor characteristics change, the index equals unity:

Test 2 (Identity).

$$\phi(P^t, P^t, \boldsymbol{M}^t, \boldsymbol{M}^t, \boldsymbol{M}) = 1$$

for all P^t, $\boldsymbol{M}^t$ and $\boldsymbol{M}$. ◇

The next two tests represent properties of monotonicity. For this aim, a sensible order relation on $\mathbb{R}^{\mathcal{G}}_{\geq 0}$ and $(\mathbb{R}^{K_{\mathcal{G}}})^{\mathcal{G}}$ needs to be introduced. We thus look first at the following definition:

Definition 4.2. *Let X_0, X_1 be two real-valued random variables with distribution functions F_{X_0} and F_{X_1}, respectively. Then X_1 is called* ***strictly stochastically larger*** *than X_0, noted by $X_1 >_{\text{st}} X_0$, if*

$$F_{X_1}(x) \leq F_{X_0}(x) \quad \forall\, x \in \mathbb{R} \quad \textit{and} \quad \exists\, x_0 \in \mathbb{R} : F_{X_1}(x_0) < F_{X_0}(x_0)\,.$$

Of interest for the current context are, however, conditional price distributions given a certain realisation of the characteristics vector. The following definition extends the notion of stochastic dominance introduced above to conditional distributions.

36 **Definition 4.3.** *Let X_0, X_1, Y be three random variables, the former two having values in $\mathbb{R}$, and let $F_{X_0 \mid y}$ and $F_{X_1 \mid y}$ denote the conditional distribution functions of X_0 and X_1, given $Y = y$, respectively. Then X_1, given $Y = y$, is* ***conditionally strictly stochastically larger*** *than X_0, given $Y = y$, noted by $X_1 \mid y >_{\mathrm{st}} X_0 \mid y$, if*

$$F_{X_1 \mid y}(x) \leq F_{X_0 \mid y}(x) \quad \forall\, x \in \mathbb{R} \quad \textit{and} \quad \exists\, x_0 \in \mathbb{R} : F_{X_1 \mid y}(x_0) < F_{X_0 \mid y}(x_0)\,.$$

This concept now allows the formulation of two axioms on the monotonicity of hedonic elementary price indices. The idea here is that, if the current period prices rise, so does the index:

Test 3 (Monotonicity in current period prices).

$$\phi(P^0, P^{1+}, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) > \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.$$

if $P^{1+} \mid \boldsymbol{m} >_{\mathrm{st}} P^1 \mid \boldsymbol{m}$ for all $\boldsymbol{m} \in \boldsymbol{M}(\mathcal{G})$. ◇

Conversely, if the base period prices increase, the price index decreases:

Test 4 (Monotonicity in base period prices).

$$\phi(P^{0+}, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) < \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.$$

if $P^{0+} \mid \boldsymbol{m} >_{\mathrm{st}} P^0 \mid \boldsymbol{m}$ for all $\boldsymbol{m} \in \boldsymbol{M}(\mathcal{G})$. ◇

Another basic property a hedonic index should possess is the fact that multiplying the current period prices by a positive number λ should lead to a multiplication of the index value by the same factor:

Test 5 (Proportionality in current period prices).

$$\phi(P^0, \lambda P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \lambda \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.$$

for all $\lambda \in \mathbb{R}_{>0}$. ◇

For base period prices, on the other hand, an inverse proportionality is assumed:

Test 6 (Inverse proportionality in base period prices).

$$\phi(\lambda P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \lambda^{-1} \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.$$

if $\lambda \in \mathbb{R}_{>0}$. ◇

The next property cited in the literature represents the desire that an index be some kind of mean value of the observable price ratios. It is, however, not obvious how this so-called *mean value test* can be translated into the current context. Since, for traditional elementary price indices, this test is implied by Tests 1, 2, 3, and 5 (see ILO et al., 2004, p. 363, based on EICHHORN, 1978, p. 155), such a translation is not absolutely necessary.

The *symmetric treatment of outlets axiom* and the *price bouncing test*, as they can be found in ILO et al. (2004, paras. 20.59–60), for instance, are not relevant in this context. They both refer to permutations of price samples and may therefore be replaced later by some invariance properties on the index estimator, but not on the index itself.

The tests presented so far are seen as 'reasonably straightforward and uncontroversial' in the literature (ILO et al., 2004, para. 20.59). The following tests are now refinements that are said not to be 'necessarily accepted by all price statisticians'.

First of all, there is the important time reversal test that may be stated as follows:

Test 7 (Time reversal).

$$\phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = 1/\phi(P^1, P^0, \boldsymbol{M}^1, \boldsymbol{M}^0, \boldsymbol{M}) \qquad \diamond$$

A strengthened version of the time reversal test is the so-called circularity test:

Test 8 (Circularity).

$$\phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) \times \phi(P^1, P^2, \boldsymbol{M}^1, \boldsymbol{M}^2, \boldsymbol{M}) = \phi(P^0, P^2, \boldsymbol{M}^0, \boldsymbol{M}^2, \boldsymbol{M}) \qquad \diamond$$

A common demand on a price index is its invariance to changes in the units of price measurement. This property is reflected by the commensurability axiom or test:

Test 9 (Commensurability).

$$\phi(\lambda P^0, \lambda P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.$$

if $\lambda \in \mathbb{R}_{>0}$. $\diamond$

It should be noted, however, that Test 9 will be satisfied by any index that satisfies Test 5 and Test 6.

Such an invariance under changes of units of measurements could, at first sight, also be proposed for the individual components of the characteristics vectors $\boldsymbol{M}^0$, $\boldsymbol{M}^1$ and $\boldsymbol{M}$. Given that some of the characteristics may be measured at the nominal or ordinal level only, such a rescaling could, however, be senseless under certain circumstances.

All of these ten axioms concentrate on the index' behaviour relating to the two price variables. The characteristics variables are more difficult to tackle. This is mainly due to the fact that the nature and the meaning of the individual characteristics may vary in a large spectrum going from precise metric physical measurements to potentially blurry expert judgments measured on an ordinal or even nominal scale.

One desirable property, however, is that the ordering of the price-relevant characteristics should not be of any importance for hedonic elementary price indices. This is reflected by the fact that permuting the elements of the vectors of characteristics expressions does not change the index value:

Test 10 (Symmetric treatment of characteristics).

$$\phi(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \phi(P^0, P^1, \mathbf{P}\boldsymbol{M}^0, \mathbf{P}\boldsymbol{M}^1, \mathbf{P}\boldsymbol{M})$$

if $\mathbf{P}$ *is a* $K_{\mathcal{G}}$*-dimensional permutation matrix.* ◇

This completes the listing of tests for a bilateral hedonic elementary price index. In the following section, it will be investigated under what circumstances the index concepts introduced in Section 3.3.2 satisfy these axioms.

4.3 Testing the index definitions

In order to check whether (3.25) and (3.26) satisfy the axioms stated above, it seems appropriate to reformulate these formulae such that all the ingredients of Def. 4.1 become visible. In fact, (3.25) and (3.26) can be re-written as

$$\phi_1(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^1 \mid \boldsymbol{M}^1}(P^1 \mid \boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})))} \tag{4.1}$$

and

$$\phi_2(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1 \mid \boldsymbol{M}^1}(P^1 \mid \boldsymbol{M})}{\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})}\right)\right]\right) \tag{4.2}$$

respectively. The mapping φ is, as defined above, a continuous and injective function which maps a connected subset of $\mathbb{R}$ to $\mathbb{R}$.

The two hedonic indices ϕ_1 and ϕ_2 both satisfy the continuity test (Test 1), since the expectation, be it conditional or not, is continuous in the $\mathbf{L}_2$-sense described in the axiom (see e.g. Fristedt and Gray, 1997), and since φ and φ^{-1} are continuous.

It is straightforward to see that ϕ_1 and ϕ_2 satisfy the identity test (Test 2), due to the fact that $\mathrm{E}_{\boldsymbol{M}}$ is linear. In fact, if $P^0 = P^1 = P^t$ and $\boldsymbol{M}^0 = \boldsymbol{M}^1 = \boldsymbol{M}^t$, we have $\phi_1 = 1$ and $\phi_2 = \varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}[\varphi(1)]) = 1$.

More care is necessary to prove under what conditions the indices ϕ_1 and ϕ_2 proposed above satisfy the monotonicity tests (Tests 3 and 4). Theorems on the monotonicity of the mean operator can be found in the literature (see e.g. Lehmann 1986, S. 84, Müller and Stoyan 2002, S. 5, or Ross 1983, S. 252). In the following, we are first going to prove a similar result for strict stochastic dominance and conditional expectations. The central point here is the following lemma:

Lemma 4.4. *Let X_0, X_1 be two real-valued random variables, $X_1 >_{\mathrm{st}} X_0$, and let V be a random variable uniformly distributed on $[0,1]$. Then there exist two non-decreasing functions g_0 and g_1, such that*

1. *$g_1(v) \geq g_0(v)$ for all v,*
2. *$\exists v_0, v_1 \in [0,1]$, $v_0 < v_1$, with $g_1(v) > g_0(v)$ for $v_0 < v < v_1$ and*
3. *$g_0(V) \overset{\text{L}}{\sim} X_0$, $g_1(V) \overset{\text{L}}{\sim} X_1$.* ◇

Proof. Let F_0 and F_1 be the cumulative distribution functions of X_0 and X_1, respectively. As $X_1 >_{\mathrm{st}} X_0$, it follows that $F_1(x) \leq F_0(x)$ for all x, and there exists a value x_0 with $F_1(x_0) < F_0(x_0)$. Let

$$g_i(v) := \inf\{x \mid F_i(x-0) \leq v \leq F_i(x)\} \quad (i = 0, 1).$$

The functions g_0 and g_1 are non-decreasing, because F_0 and F_1 are non-decreasing. Hence,

$$g_i(F_i(x)) \leq x \quad \text{and} \quad F_i(g_i(v)) \geq v \qquad \forall\, x, v \quad (i = 0, 1). \tag{4.3}$$

It follows that $v \leq F_i(x')$ implies $g_i(v) \leq g_i(F_i(x')) \leq x'$ and, conversely, $g_i(v) \leq x'$ implies $F_i(g_i(v)) \leq F_i(x')$. Therefore, $v \leq F_i(x')$, and we have

$$g_i(v) \leq x' \quad \Longleftrightarrow \quad v \leq F_i(x')\,.$$

It follows that

$$\mathbb{P}(g_i(V) \leq x) = \mathbb{P}(V \leq F_i(x)) = F_i(x) \qquad \forall\, x\,,$$

and thus $g_i(V) \overset{\text{L}}{\sim} X_i$, $i = 0, 1$. Moreover,

$$\begin{aligned} g_0(v) &= \inf\{x \,|\, F_0(x-0) \le v \le F_0(x)\} \\ &\le \inf\{x \,|\, F_1(x-0) \le v \le F_1(x)\} = g_1(v) \qquad \forall\, v\,, \end{aligned}$$

as $F_1(x) \le F_0(x) \,\forall\, x$ and because F_0, F_1 are monotonically increasing. To complete the proof, we look at

$$v_\varepsilon := F_1(x_0) + \varepsilon(F_0(x_0) - F_1(x_0)) = F_0(x_0) - (1-\varepsilon)(F_0(x_0) - F_1(x_0))\,.$$

If $0 < \varepsilon < 1$, then, because of $F_0(x_0) - F_1(x_0) > 0$, the monotonicity of g_0 and (4.3), there is

$$g_0(v_\varepsilon) = g_0\big(F_0(x_0) - (1-\varepsilon)(F_0(x_0) - F_1(x_0))\big) \le g_0(F_0(x_0)) \le x_0\,.$$

It remains to be shown that $g_1(v_\varepsilon) > x_0$ for $0 < \varepsilon < 1$, from where we have $g_1(v) > g_0(v)$ for $F_1(x_0) < v < F_0(x_0)$. Assume thus that $g_1(v_\varepsilon) \le x_0$. We then also have $F_1(g_1(v_\varepsilon)) \le F_1(x_0)$. Conversely, (4.3) says that $F_1(g_1(v_\varepsilon)) \ge v_\varepsilon = F_1(x_0) + \varepsilon(F_0(x_0) - F_1(x_0))$. Therefore $F_1(x_0) \ge F_1(x_0) + \delta$ with $\delta = \varepsilon(F_0(x_0) - F_1(x_0)) > 0$, which is a contradiction. ∎

This leads us now to the following proposition on the strict monotonicity of the mean operator:

Corollary 4.5. *Let X_0, X_1 be two real-valued random variables, $X_1 >_{\text{st}} X_0$. Then, $\mathrm{E}(X_1) > \mathrm{E}(X_0)$, if both of the expectations exist.* ◇

Proof. Let V be a random variable with a uniform distribution on $[0, 1]$. It follows from Lemma 4.4 that there are two functions g_0 and g_1 with $g_1(v) \ge g_0(v)$ for all v and $g_i(V) \overset{\text{L}}{\sim} X_i$, $i = 0, 1$. We therefore have

$$\mathrm{E}(X_1) = \mathrm{E}(g_1(V)) \ge \mathrm{E}(g_0(V)) = \mathrm{E}(X_0)$$

due to the monotonicity of the integral. To prove the strict inequality, we define the difference function $g_{\text{d}} := g_1 - g_0$. Hence,

$$\mathrm{E}(X_1) = \mathrm{E}(g_1(V)) = \mathrm{E}\big((g_0+g_{\text{d}})(V)\big) = \mathrm{E}(g_0(V)) + \mathrm{E}(g_{\text{d}}(V)) = \mathrm{E}(X_0) + \mathrm{E}(g_{\text{d}}(V))\,.$$

It remains to be shown that $\mathrm{E}(g_{\text{d}}(V)) > 0$. From Lemma 4.4, we know that $g_{\text{d}}(v) \ge 0$ for all v and that $g_{\text{d}}(v) > 0$ for $v_0 < v < v_1$. Therefore,

$$\mathrm{E}(g_{\text{d}}(V)) = \int_0^1 g_{\text{d}}(v)\,\mathrm{d}v \ge \int_{v_0}^{v_1} g_{\text{d}}(v)\,\mathrm{d}v > 0\,. \qquad ∎$$

Let us now show that this same property also holds for conditional expectations.

Lemma 4.6. *Let Y be a random variable, g_0, g_1 two real-valued measurable functions with $\mathbb{P}((g_1 - g_0)(Y) > 0) = 1$. Then, $g_1(Y) >_{\text{st}} g_0(Y)$.* ◇

Proof. Let F_0 and F_1 be the cumulative distributions functions of $g_0(Y)$ and $g_1(Y)$, respectively. We then have

$$F_1(x) = \mathbb{P}(g_1(Y) \leq x) \leq \mathbb{P}(g_0(Y) \leq x) = F_0(x) \qquad \forall\, x\,.$$

It remains to be shown that there exists a value x_0, such that $F_1(x_0) < F_0(x_0)$. Assume that $F_1(x) = F_0(x)$ for all x. In this case, $\mathbb{P}(g_1(Y) \leq x) = \mathbb{P}(g_0(Y) \leq x)$, i.e.

$$\mathbb{P}(g_1(Y) > x \geq g_0(Y)) = \mathbb{P}(g_0(Y) \leq x) - \mathbb{P}(g_1(Y) \leq x) = 0 \qquad \forall\, x\,,$$

which is a contradiction to the assumption that $\mathbb{P}((g_1 - g_0)(Y) > 0) = 1$. ■

Corollary 4.7. *Let X_0, X_1 and Y be random variables, the former two being real-valued. Assume that $X_1 \,|\, y >_{\text{st}} X_0 \,|\, y$ for all y in the range of Y. Then,*

$$\mathrm{E}(X_1 \,|\, Y) >_{\text{st}} \mathrm{E}(X_0 \,|\, Y)\,,$$

if both of the conditional expectations exist. ◇

Proof. Let $g_i(y) := \mathrm{E}(X_i \,|\, y)$, $i = 0, 1$. From Corollary 4.5, we know that $g_1(y) > g_0(y)$ for all y on the support of the probability (density) function of Y. It follows from Lemma 4.6 that $g_1(Y) >_{\text{st}} g_0(Y)$. ■

Theorem 4.8. *The indices ϕ_1 and ϕ_2 are monotonic in the sense of Tests 3 and 4.* ◇

Proof. We are going to assume that $P^{1+} \,|\, \boldsymbol{m} >_{\text{st}} P^1 \,|\, \boldsymbol{m}$ and $P^{0+} \,|\, \boldsymbol{m} >_{\text{st}} P^0 \,|\, \boldsymbol{m}$ for all $\boldsymbol{m} \in \boldsymbol{M}(\mathcal{G})$. Because φ is continuous and injective, it is also strictly monotonic. It follows from Corollary 4.7 and Lemma 4.6 that $\mathrm{E}_{P^{1+} \,|\, \boldsymbol{M}^1}(P^{1+} \,|\, \boldsymbol{M}) >_{\text{st}} \mathrm{E}_{P^1 \,|\, \boldsymbol{M}^1}(P^1 \,|\, \boldsymbol{M})$ and from Corollary 4.5, we conclude that

$$\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1+} \,|\, \boldsymbol{M}^1}(P^{1+} \,|\, \boldsymbol{M}))) > \varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^1 \,|\, \boldsymbol{M}^1}(P^1 \,|\, \boldsymbol{M})))\,. \tag{4.4}$$

To see this, let

$$g_1(\boldsymbol{m}) := \varphi(\mathrm{E}_{P^{1+} \,|\, \boldsymbol{M}^1}(P^{1+} \,|\, \boldsymbol{m})) \quad \text{and} \quad g_0(\boldsymbol{m}) := \varphi(\mathrm{E}_{P^1 \,|\, \boldsymbol{M}^1}(P^1 \,|\, \boldsymbol{m}))$$

for all admissible $\boldsymbol{m} \in \mathbb{R}^{K_{\mathcal{G}}}$. From Corollary 4.5, we have $g_1(\boldsymbol{m}) > g_0(\boldsymbol{m})$ for all $\boldsymbol{m}$ if φ is increasing, and it follows from Lemma 4.6 that $g_1(\boldsymbol{M}) >_{\text{st}} g_0(\boldsymbol{M})$. Applying Corollary 4.5 a second time leads to $\mathrm{E}_{\boldsymbol{M}}(g_1(\boldsymbol{M})) > \mathrm{E}_{\boldsymbol{M}}(g_0(\boldsymbol{M}))$, and, since φ^{-1} is also strictly monotonically increasing, to the mentioned result.

If, on the other hand, φ is decreasing, we have $g_1(\boldsymbol{m}) < g_0(\boldsymbol{m})$, $g_1(\boldsymbol{M}) <_{\mathrm{st}} g_0(\boldsymbol{M})$ and $\mathrm{E}_{\boldsymbol{M}}(g_1(\boldsymbol{M})) < \mathrm{E}_{\boldsymbol{M}}(g_0(\boldsymbol{M}))$. Since in this case, φ^{-1} is likewise strictly monotonically decreasing, we return to the inequality (4.4).

Therefore,

$$\frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1+}\,|\,\boldsymbol{M}^1}(P^{1+}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))} > \frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))}$$

and, by analogy,

$$\frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0+}\,|\,\boldsymbol{M}^0}(P^{0+}\,|\,\boldsymbol{M})))} < \frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))}\,.$$

This completes the proof of the monotonicity of ϕ_1. The monotonicity of ϕ_2 is proved in a completely analogous way. ■

Let us now turn to the proportionality axioms.

Theorem 4.9. *A sufficient condition for ϕ_1 and ϕ_2 to satisfy the proportionality tests (Tests 5 and 6) is that $\varphi(\lambda\, x) = \varphi(\lambda) + \varphi(x)$ or $\varphi(\lambda\, x) = \varphi(\lambda)\,\varphi(x)$ for all $\lambda, x \in \mathbb{R}$.* ◇

Proof. Let us first assume that

$$\varphi(\lambda\, x) = \varphi(\lambda) + \varphi(x) \tag{4.5}$$

for all $\lambda, x \in \mathbb{R}$. We then have $\varphi^{-1}(\mu + y) = \varphi^{-1}(\mu)\varphi^{-1}(y)$ for all $\mu, y \in \mathbb{R}$. To see this, we define $\mu^* = \varphi^{-1}(\mu)$ and $y^* = \varphi^{-1}(y)$. Thus, $\varphi^{-1}(\mu)\,\varphi^{-1}(y) = \varphi^{-1}(\varphi(\mu^*))\,\varphi^{-1}(\varphi(y^*)) = \mu^*\, y^* = \varphi^{-1}(\varphi(\mu^*\, y^*))$. Given (4.5), this last term is equal to $\varphi^{-1}(\varphi(\mu^*) + \varphi(y^*)) = \varphi^{-1}(\mu + y)$.

Due to the property just introduced and the linearity of the conditional expectations, we have now

$$\begin{aligned}
\phi_1(P^0, \lambda P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) &= \frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(\lambda P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))} \\
&= \frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\lambda \mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))} \\
&= \frac{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}[\varphi(\lambda) + \varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M}))])}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))} \\
&= \frac{\varphi^{-1}(\varphi(\lambda) + \mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))} \\
&= \frac{\varphi^{-1}(\varphi(\lambda))\,\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{1}\,|\,\boldsymbol{M}^1}(P^{1}\,|\,\boldsymbol{M})))}{\varphi^{-1}(\mathrm{E}_{\boldsymbol{M}}\varphi(\mathrm{E}_{P^{0}\,|\,\boldsymbol{M}^0}(P^{0}\,|\,\boldsymbol{M})))} \\
&= \lambda\phi_1(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.
\end{aligned}$$

This proves the satisfaction of test 5 by ϕ_1 if φ satisfies (4.5). If $\varphi(\lambda\, x) = \varphi(\lambda)\,\varphi(x)$, the same proof holds when all the additions are replaced by multiplications. Moreover, an analogous proof can be given for ϕ_2 or test 6. ∎

By simplification, ϕ_1 always satisfies the circularity test (Test 8) and thus automatically the time reversal test (Test 7). For ϕ_2, we have one sufficient condition given in the following theorem.

Theorem 4.10. *A sufficient condition for ϕ_2 to satisfy the circularity axiom (Test 8) is that $\varphi(x\,y) = \varphi(x) + \varphi(y)$ for all $x, y \in \mathbb{R}$.* ◇

Proof. From the condition $\varphi(x\,y) = \varphi(x) + \varphi(y)$, it follows immediately that $\varphi^{-1}(x+y) = \varphi^{-1}(x)\,\varphi^{-1}(y)$ for all $x, y \in \varphi(\mathbb{R})$ (see proof of Theorem 4.9). Therefore

$$
\begin{aligned}
&\phi_2(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) \times \phi_2(P^1, P^2, \boldsymbol{M}^1, \boldsymbol{M}^2, \boldsymbol{M}) \\
&= \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\right)\right]\right)\ \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^2\,|\,\boldsymbol{M}^2}(P^2\,|\,\boldsymbol{M})}{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}\right)\right]\right) \\
&\quad = \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\right)\right] + \mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^2\,|\,\boldsymbol{M}^2}(P^2\,|\,\boldsymbol{M})}{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}\right)\right]\right) \\
&\quad\quad = \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\right) + \varphi\left(\frac{\mathrm{E}_{P^2\,|\,\boldsymbol{M}^2}(P^2\,|\,\boldsymbol{M})}{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}\right)\right]\right) \\
&\quad\quad\quad = \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\,\frac{\mathrm{E}_{P^2\,|\,\boldsymbol{M}^2}(P^2\,|\,\boldsymbol{M})}{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}\right)\right]\right) \\
&\quad\quad\quad\quad = \phi_2(P^0, P^2, \boldsymbol{M}^0, \boldsymbol{M}^2, \boldsymbol{M})\,.
\end{aligned}
$$
∎

The commensurability axiom (Test 9) is always satisfied by ϕ_2 since

$$
\begin{aligned}
\phi_2(\lambda P^0, \lambda P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) &= \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(\lambda P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(\lambda P^0\,|\,\boldsymbol{M})}\right)\right]\right) \\
&= \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\lambda\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\lambda\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\right)\right]\right) \\
&= \varphi^{-1}\left(\mathrm{E}_{\boldsymbol{M}}\left[\varphi\left(\frac{\mathrm{E}_{P^1\,|\,\boldsymbol{M}^1}(P^1\,|\,\boldsymbol{M})}{\mathrm{E}_{P^0\,|\,\boldsymbol{M}^0}(P^0\,|\,\boldsymbol{M})}\right)\right]\right) \\
&= \phi_2(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M})\,.
\end{aligned}
$$

Moreover, it is satisfied by ϕ_1 under the condition given in Theorem 4.9. This is due to the fact that any index satisfying Tests 5 and 6 also satisfies Test 9.

Table 4.1: **Satisfaction of the tests by the two indices for different transformation functions φ**

Index	$\varphi(x)$	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
ϕ_1	x	•	•	•	•	•	•	•	•	•	•
	$\ln x$	•	•	•	•	•	•	•	•	•	•
	x^{-1}	•	•	•	•	•	•	•	•	•	•
	any	•	•	•	•	–	–	•	•	–	•
ϕ_2	x	•	•	•	•	•	•	–	–	•	•
	$\ln x$	•	•	•	•	•	•	•	•	•	•
	x^{-1}	•	•	•	•	•	•	–	–	•	•
	any	•	•	•	•	–	–	–	–	•	•

•: satisfied, –: not satisfied in general

Finally, ϕ_1 and ϕ_2 both satisfy Test 10 since

$$\mathrm{E}(P \,|\, \boldsymbol{M}) = \mathrm{E}(P \,|\, \mathbf{P}\boldsymbol{M})$$

for any permutation matrix $\mathbf{P}$.

In Section 3.3.2, we identified three important candidates for φ, namely $\varphi(x) = x$, $\varphi(x) = \ln x$ and $\varphi(x) = x^{-1}$. Table 4.1 lists the tests which are satisfied by the indices ϕ_1 and ϕ_2 in general or depending on how φ is chosen among these candidates.

The fact that Test 7 and Test 8 are not generally satisfied by ϕ_2 and $\varphi(x) = x$ or $\varphi(x) = x^{-1}$ can be seen as follows: If $\varphi(x) = x$ and $\mathbb{P}_{\boldsymbol{M}}$ is the empirical distribution of any reference set $\mathcal{M}^*$, ϕ_2 has the form

$$\frac{1}{\mathrm{card}\mathcal{M}^*} \sum_{\boldsymbol{m} \in \mathcal{M}^*} \frac{h^1_{\mathcal{G}}(\boldsymbol{m})}{h^0_{\mathcal{G}}(\boldsymbol{m})}\,, \tag{4.6}$$

which is exactly the formula of the Carli elementary price index. If, under the same assumption concerning $\mathbb{P}_{\boldsymbol{M}}$, $\varphi(x)$ equals x^{-1}, then (4.6) becomes the harmonic mean of the price relatives. It is, however, known from the literature that both of these index formulae generally fail the time reversal and the circularity test (see e.g. ILO et al., 2004, p. 364).

To see that ϕ_1 may fail Test 5 if φ does not satisfy one of the conditions of Theorem 4.9, let us consider the following setting. Assume that $\mathbb{P}_{\boldsymbol{M}}$ is a two-point distribution with values $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ each having the probability 1/2. Assume further that

$$p^1_1 := \mathrm{E}_{P^1 \,|\, \boldsymbol{M}^1}(P^1 \,|\, \boldsymbol{M} = \boldsymbol{m}_1)) \neq \mathrm{E}_{P^1 \,|\, \boldsymbol{M}^1}(P^1 \,|\, \boldsymbol{M} = \boldsymbol{m}_2)) =: p^1_2\,.$$

If we now take $\varphi = \exp$ (and therefore $\varphi^{-1} = \ln$), we have

$$\begin{aligned}\phi_1(P^0, \lambda P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) &= \frac{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\mathrm{E}_{P^1 \mid \boldsymbol{M}^1}(\lambda P^1 \mid \boldsymbol{M})))}{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})))} \\ &= \frac{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\lambda \mathrm{E}_{P^1 \mid \boldsymbol{M}^1}(P^1 \mid \boldsymbol{M})))}{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})))} \\ &= \frac{\ln(\exp(\lambda p_1^1)\frac{1}{2} + \exp(\lambda p_2^1)\frac{1}{2})}{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})))} \\ &= \frac{-\ln 2 + \ln((\exp \lambda p_1^1) + (\exp \lambda p_2^1))}{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})))},\end{aligned}$$

and this last expression is different from

$$\lambda\, \phi_1(P^0, P^1, \boldsymbol{M}^0, \boldsymbol{M}^1, \boldsymbol{M}) = \lambda\, \frac{-\ln 2 + \ln((\exp p_1^1) + (\exp p_2^1))}{\ln(\mathrm{E}_{\boldsymbol{M}} \exp(\mathrm{E}_{P^0 \mid \boldsymbol{M}^0}(P^0 \mid \boldsymbol{M})))}$$

in general. This counterexample also holds for ϕ_2 and Test 5, for Test 6 as well as for ϕ_1 and Test 9.

4.4 Conclusion

From an axiomatic point of view, the choice between ϕ_1 and ϕ_2 is biased towards the former, at least if one is willing to accept the additional restriction that $\varphi(\lambda\, x) = \varphi(\lambda) + \varphi(x)$ or $\varphi(\lambda\, x) = \varphi(\lambda)\, \varphi(x)$ for all $\lambda, x \in \mathbb{R}$. This is due to the fact that, in this case, all the axioms presented in Section 4.2 are satisfied. It is true that this also holds for ϕ_2 together with $\varphi = \ln$, but then ϕ_1 and ϕ_2 coincide, such that this special case is already embraced by ϕ_1.

The only argument that speaks for ϕ_2 based on the results obtained above is that ϕ_2 always satisfies the commensurability axiom (Test 9) whereas ϕ_1 does not in every case. This, however, is only relevant if the above restrictions on φ do not apply. The index ϕ_2, finally, remains also a candidate if one is prepared to disregard the time reversal and the circularity test.

Unfortunately, there is no apparent proof that the above restrictions on φ are not only sufficient but also necessary for ϕ_1 to satisfy all the tests. The family of admissible transformation functions φ therefore cannot be specified further here.

As a matter of fact, based on what has been presented so far, there is no single response to the question of how a hedonic elementary price index

should be defined. Several interpretations are possible, and, just as in the case of conventional (i.e. non-hedonic) elementary indices, other criteria need to be taken into account in order to decide between the alternatives. These criteria presumably do not depend on the fact that quality adjustments are undertaken but resemble those presented, for instance, in ILO et al. (2004, Chap. 20).

The results we obtained in the current part are still innovative in the way that the task of defining a hedonic index was completely separated from the estimation. It is important to see that a quality-adjusted or, more specifically, a hedonic elementary price index is an economic parameter that first needs to be properly defined. It is a non-stochastic parameter of the price and quality distributions realised on the market, and it depends on the concept of a reference quality spectrum that is held constant for comparing two time periods.

Now that such candidate definitions are identified, we can turn our attention to the question of how these parameters may be estimated. This is going to be attempted in Chapter 6.

PART II

Estimation Techniques for Hedonic Elementary Price Indices

CHAPTER 5

Estimating the Hedonic Function

5.1 General considerations

The issue of estimating the hedonic function

$$h^t_{\mathcal{G}}(\boldsymbol{m}) = \mathrm{E}(P^t \,|\, \boldsymbol{M}^t = \boldsymbol{m})$$

for a given elementary aggregate $\mathcal{G}$ at a specific point t in time and for any meaningful characteristics vector $\boldsymbol{m} \in \mathbb{R}^{K_{\mathcal{G}}}$ is fundamental for the subsequent estimation of a hedonic elementary price index. Since the focus will always be directed to one specific elementary aggregate, the index $\mathcal{G}$ will be dropped in the further analysis.

The starting point for a hedonic regression is always a sample of different representatives of the elementary aggregate under review at time t. More specifically, the estimation of h^t is based on a random sample, i.e. on N^t independent and identically distributed (i.i.d.) random variables

$$(P^t_1, \boldsymbol{M}^t_1), \ldots, (P^t_{N^t}, \boldsymbol{M}^t_{N^t}), \tag{5.1}$$

where $(P^t_n, \boldsymbol{M}^t_n) \overset{\perp}{\sim} (P^t, \boldsymbol{M}^t)$ for all $n \in \{1, \ldots, N^t\}$.

If $\mathcal{H} := \{h : \mathbb{R}^K \longrightarrow \mathbb{R}_{\geq 0}\}$ denotes the potential hedonic functions applicable to the given elementary aggregate, any estimate $\hat{h}^t$ of the hedonic function h^t results then from a mapping

$$\begin{array}{rccc} \mathfrak{h} & : & \mathbb{R}^{N^t}_{\geq 0} \times \mathbb{R}^{N^t \times K} & \longrightarrow & \mathcal{H} \\ & & (\boldsymbol{P}^t, \mathbf{M}^t) & \longmapsto & \hat{h}^t := \mathfrak{h}[\boldsymbol{P}^t, \mathbf{M}^t] \end{array} \tag{5.2}$$

where $\boldsymbol{P}^t = (P_1^t, \ldots, P_{N^t}^t)'$ denotes the random vector of sampled prices and $\mathbf{M}^t$ is the respective $N \times K$ random matrix $(\boldsymbol{M}_1^t, \ldots, \boldsymbol{M}_{N^t}^t)$ of the price-relevant characteristics. As a consequence, $\hat{h}^t$ is actually a random variable with values in $\mathcal{H}$.

In order to simplify the notation, we are not going to distinguish $\hat{h}^t$ as an estimator (i.e. a random variable with value in $\mathcal{H}$) or as an estimation (i.e. an element of $\mathcal{H}$) of h^t. It will be clear from the context which of the two meanings is applicable.

Given any specific representative of the elementary aggregate with characteristics vector $\boldsymbol{m}$, its price predictor is then defined by

$$\hat{P}^t := \hat{h}^t(\boldsymbol{m}) = \mathfrak{h}[\boldsymbol{P}^t, \mathbf{M}^t](\boldsymbol{m})\,.$$

In this framework, the vector $\boldsymbol{m}$ is not random but fixed in advance, and the randomness of $\hat{P}^t$ stems from the randomness of the estimator $\hat{h}^t$.

5.2 Model selection

As we have just discussed, the estimation of a hedonic function for a specific period t is a mapping of the data $(\boldsymbol{P}^t, \mathbf{M}^t)$ to the set $\mathcal{H}$ of all possible hedonic functions. In practice, however, only very restricted subsets of $\mathcal{H}$ are commonly regarded as candidates for $\hat{h}^t$. In specifying the so-called *functional form* of a hedonic regression, one implicitly limits the range of possible outcomes of $\mathfrak{h}$ to a well-defined family of functions.

Figure 5.1 describes, inspired by Hastie et al. (2001, p. 199), the relationship between the true and estimated hedonic functions as well as the possible outcomes of $\mathfrak{h}$ in $\mathcal{H}$. The black dot labelled 'True h^t' indicates the true hedonic function as it is assumed to exist by the hedonic econometric model. It is this point in $\mathcal{H}$ one would like to estimate as accurately as possible. However, as there is a model residual ϵ^t which represents quality-independent price components, the actual realisation of h^t in a given sample will almost surely differ from the true hedonic function. The set of possible realisations is depicted by the gray shaded area containing an 'Observed h^t' as an example.

When specifying the functional form of the hedonic regression model, one explicitly restricts the outcomes of $\mathfrak{h}$ to the subset of $\mathcal{H}$ containing all the

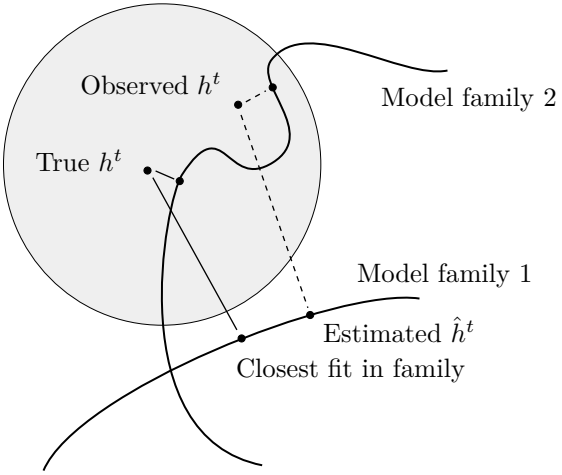


Figure 5.1: **Schematic outline of the set $\mathcal{H}$ of all possible hedonic functions**

elements of the proposed form. If a linear regression approach is chosen, for instance, only functions of the form

$$h(\boldsymbol{m}) = \beta_0 + \sum_{k=1}^{K} \beta_k \, m_k \tag{5.3}$$

with $\beta_0, \ldots, \beta_K$ being any real numbers are allotted as candidates for choosing the hedonic function estimate $\hat{h}^t$. Two such subsets of $\mathcal{H}$ are depicted by the curved lines labelled 'Model family 1' and 'Model family 2' in Fig. 5.1. It is obvious that the intersection of two or more such subsets is not necessarily empty due to the fact that any given hedonic function may be contained in several model families.

In general, the estimation of a hedonic function is done by determining or searching for an element in such a model family which is—in a certain sense—closest to the observation. The distance to be minimised is measured by an appropriate metric such as the least squares or least absolute deviations criteria. This procedure, however, does not yield the hedonic function within the family that is closest to the truth but the function that is closest to the observation.

The degree of flexibility of a regression approach to closely fit the data—or, in other words, the degrees of freedom involved—are connected in some

sense with the rigidity of the model families depicted in Fig. 5.1. If the functional form is too rigid, it may occur that both the estimated $\hat{h}^t$ and the closest fit in family are far away from the true h^t. This is potentially the case for family one in the drawing. On the other hand, however, the actual estimate $\hat{h}^t$ in this first example is near to the closest fit in family, and it has about the same distance from either the observed or the true hedonic function. Another extreme is the second model family in the drawing which tends to over-fit the data. Here, the estimated $\hat{h}^t$ is far away both from the closest fit in family and from the true hedonic function. Nevertheless, the estimate within the second family is nearer to the truth than the estimate within the first family. In this case and for this realisation, the second family would thus be preferable, even though the over-fitting is important.

These reflections describe the difficulty of choosing an appropriate functional form in practice. The question of how to compare different model families with respect to their goodness of fit is going to be treated in detail in Section 5.4. There has, moreover, been a lot of discussion in the literature on whether one or the other regression approach are valid or preferable from an economic point of view for estimating hedonic functions. Diewert (2003a), for example, states that the linear regression approach or, equivalently, the family of functions (5.3), is 'unlikely to be consistent with microeconomic theory'. Triplett (2004) discusses in detail, however, that most of these studies—including the one by Diewert—implicitly or explicitly treat the special case where all the buyers have the same preferences. This assumption, however, is empirically inappropriate. Largely based on Rosen (1974), Triplett concludes not only that the 'hedonic functional form is purely an empirical issue' but also that 'Any empirical form that fits the data is consistent with the theory'. He shares thus the opinion expressed already by Griliches (1971a, p. 58) that 'There is no a priori reason to expect price and quality to be related in any particular fixed fashion'.

Looked at in this way, the choice of a functional form for a hedonic regression is a purely statistical problem. Any regression technique that can be used to describe the relationship between characteristics and price is therefore a candidate for further investigation. A few such approaches are now going to be presented.

5.3.1 Linear regression Linear regression methods are undoubtedly the most commonly used technique for estimating the hedonic function. Triplett (2004, pp. 180 ff.) discusses the history and the common practice for choosing functional forms in hedonic regressions. According to him, the two most widely used functional families are the semi-log and the double-log approach.

In the *semi-log* functional form, the underlying regression equation is

$$\ln P^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t \, M_k^t + \eta^t \tag{5.4}$$

and the corresponding hedonic function has the form

$$\mathfrak{h}[\boldsymbol{P}^t, \mathbf{M}^t](\boldsymbol{m}) = \exp\left(\hat{\beta}_0^t[\boldsymbol{P}^t, \mathbf{M}^t] + \sum_{k=1}^{K} \hat{\beta}_k^t[\boldsymbol{P}^t, \mathbf{M}^t] m_k \right) .$$

Here, the $\hat{\beta}_k^t[\boldsymbol{P}^t, \mathbf{M}^t]$ are ordinary least squares estimates of the parameters β_k^t, $k \in \{0, \ldots, K\}$. Due to the log-transformation of the prices, the random error term η^t is not directly comparable to ϵ^t in (3.17).

It needs to be noted that the above specification holds if categorical variables only have two levels. If, however, a categorical variable M_k^t has $L_k > 2$ levels, it first needs to be recoded into $L_k - 1$ dichotomous variables $M_{k,1}^t, \ldots, M_{k,L_k-1}^t$ and, for each of them, a coefficient $\hat{\beta}_{k,l}^t[\boldsymbol{P}^t, \mathbf{M}^t]$ needs to be estimated. For the sake of simplicity, we are going to stick to the notation introduced above, and this convention is assumed to be tacitly applied where necessary.

As a second model family, the *double-log* approach starts from the equation

$$\ln P^t = \beta_0 + \sum_{k \in \mathcal{K}_{\text{con}}} \beta_k^t \, \ln M_k^t + \sum_{k \in \mathcal{K}_{\text{cat}}} \beta_k^t M_k^t + \eta^t \tag{5.5}$$

and yields the hedonic function

$$\mathfrak{h}[\boldsymbol{P}^t, \mathbf{M}^t](\boldsymbol{m}) = \exp\left(\hat{\beta}_0^t[\boldsymbol{P}^t, \mathbf{M}^t] + \sum_{k \in \mathcal{K}_{\text{con}}} \hat{\beta}_k^t[\boldsymbol{P}^t, \mathbf{M}^t] \ln m_k + \sum_{k \in \mathcal{K}_{\text{cat}}} \hat{\beta}_k^t[\boldsymbol{P}^t, \mathbf{M}^t] m_k \right)$$

Here, $\mathcal{K}_{\text{con}}$ and $\mathcal{K}_{\text{cat}}$ assemble the indices of the continuous and the categorical exogenous variables respectively.

A convenience of these two 'log linear' approaches is that the resulting hedonic function estimates always yield positive prices. With the choice of either the semi-log or the double-log approach, one implicitly assumes that the variance of the error term ϵ^t is proportional to $(\mathrm{E}[P^t])^2$ (see e.g. Montgomery and Peck, 1992, p. 98) while η^t, if at all, is homoscedastic. The predominance of these two approaches in the hedonic price index literature indicates that this assumption might often be adequate in practice. Interestingly, the double-log approach is the 'overwhelming favourite' for IT equipment studies, while the semi-log form is the most widely used for 'noncomputer products' (Triplett, 2004, p. 180), such as automobiles.

The linear regression approach

$$P^t = \beta_0 + \sum_{k=1}^{K} \beta_k \, M_k^t + \epsilon^t \tag{5.6}$$

together with the hedonic function estimate

$$\mathfrak{h}[\boldsymbol{P}^t, \mathbf{M}^t](\boldsymbol{m}) = \hat{\beta}_0^t[\boldsymbol{P}^t, \mathbf{M}^t] + \sum_{k=1}^{K} \hat{\beta}_k^t[\boldsymbol{P}^t, \mathbf{M}^t] m_k \tag{5.7}$$

is less frequently used in practice, presumably because it does not fit the data well in most of the cases (see e.g. Murray and Sarantis, 1999 or Yu, 2003). Moreover, the linear hedonic function (5.7) is likely to produce negative prices, especially if $\boldsymbol{m}$ is dissimilar to the vectors contained in $\mathbf{M}^t$. Such a situation might be observed when $\boldsymbol{m}$ is the characteristics vector of an item unavailable in period t. It is one of the main aims of the hedonic price index theory to be able to handle such settings, so it would be a severe drawback if exactly in these cases, the estimated prices were negative and thus impossible to interpret.

The question of whether to introduce the variables themselves or to use their logarithms can be tested statistically using the Box-Cox or Box-Tidwell transformations (see e.g. Berndt, 1991, p. 128). The idea here is to transform a variable X to $X^{(\lambda)}$ by

$$X^{(\lambda)} = \frac{X^\lambda - 1}{\lambda}$$

where λ is a parameter to be estimated. If $\lambda = 1$ for all variables (P^t and M_k^t, $k \in \{1, \ldots, K\}$), the functional form is linear. Conversely, if $\lambda \to 0$

then $X^{(\lambda)} \to \ln X$ or, in other words, if λ tends to 0 for all the continuous variables, the functional form approaches to the double-log specification. The difference between the Box-Cox and the Box-Tidwell transformations is that the latter allows for different λs for the individual exogenous variables. Testing one of the three functional forms described above means thus testing if the λs have the appropriate values, 0 or 1 respectively.

Box-Cox models have been used in empirical studies by Bitros and Panas (1988) or Yu (2003), for instance. According to Triplett (2004, p. 182), however, 'the Box-Cox test quite commonly results in rejection of all the functional forms offered up to it.' This holds at least for the three approaches mentioned above and is one of the reasons why Dickie et al. (1997) argue for testing any hedonic price equation also against alternatives that are not nested within the Box-Cox form. Anglin and Gençay (1996, pp. 633–4) as well as Curry et al. (2001, p. 661) summarise the discussion in the literature about the use of linear, semi-log, double-log or Box-Cox models for hedonic regressions. On this basis, both of them conclude that more flexible functional forms need to be investigated. The alternatives they propose are a semi-parametric approach and the use of neural networks respectively. Other techniques such as boosted regression trees seem to be promising as well (see e.g. van Wezel et al., 2005).

In the present piece of work, we will content ourselves with one further approach that goes beyond the linear or log-linear models presented above, namely the partial least squares (or PLS) regression.

5.3.2 PLS regression Partial least squares regression is particularly useful in a setting where the independent variables are numerous and correlated. Similar to principal components regression (PCR), the main idea of this method is to first project the vector of covariates to a few latent variables which are then used as regressors for one or more dependent variables. In contrast to PCR which concentrates on the independent variables only, PLS regression takes into consideration the dependent variables as well for determining the latent variables. In other words, partial least squares 'seeks directions that have high variance *and* have high correlation with the response' (Hastie et al., 2001, p. 67).

The ability of reducing a large number of correlated regressors to a smaller dimension makes the PLS method particularly suitable for hedonic regressions. This is especially true if categorical characteristics with many levels are involved (e.g. the makes or models of cars). PLS regressions have been used in hedonic studies for example by Tenenhaus et al. (2004) for modelling expert judgments of products based on their characteristics. A funda-

mental overview on PLS regression is given by Abdi (2004). Comprehensive presentations are contained in Martens and Næs (1989) or Tenenhaus (1998).

Let $\mathbf{X}$ denote the design matrix built upon $\mathbf{M}^t$, i.e. the matrix having as columns the N^t realisations of M_k^t (or $M_{k,l}^t$ if the kth variable is categorical with more than two levels, see above), $k \in \{1, \ldots, K\}$. The matrix $\mathbf{X}$ has thus N^t rows and

$$\operatorname{card} \mathcal{K}_{\text{con}} + \sum_{k \in \mathcal{K}_{\text{cat}}} (L_k - 1)$$

columns, where L_k is the number of levels of the categorical variable M_k^t. Unlike in linear regressions with an intercept term, there is no column in the design matrix containing only ones. Let $\boldsymbol{Y}$ be the random vector of response variables, e.g. $\boldsymbol{Y} = \ln \boldsymbol{P}^t$ where the natural logarithm is applied component-wise. Moreover, $\mathbf{X}_{\text{c}}$ and $\boldsymbol{Y}_{\text{c}}$ denote centered copies of $\mathbf{X}$ and $\boldsymbol{Y}$, respectively, where each component is diminished by the respective column mean.

The idea of PLS regression is now to project the centered design matrix $\mathbf{X}_{\text{c}}$ onto A latent variables $\hat{\mathbf{T}} = (\hat{\boldsymbol{T}}_1, \ldots, \hat{\boldsymbol{T}}_A)$ by a matrix $\hat{\mathbf{V}}$, i.e. $\hat{\mathbf{T}} = \mathbf{X}_{\text{c}}\hat{\mathbf{V}}$. The matrix $\hat{\mathbf{V}}$ is thereby determined iteratively such that, among other conditions, the scaled covariances between $\boldsymbol{Y}$ and the $\hat{\boldsymbol{T}}_a$ are maximal. Details are to be found in the literature mentioned above. A least squares fit of a linear regression model with $\hat{\mathbf{T}}$ as the design matrix and $\boldsymbol{Y}_{\text{c}}$ as the response vector is finally estimated, yielding the so-called vector of loadings

$$\hat{\boldsymbol{Q}} = (\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}\boldsymbol{Y}_{\text{c}}\,.$$

The hedonic function can finally be written as

$$\mathfrak{h}[\boldsymbol{P}^t, \mathbf{M}^t](\boldsymbol{m}) = \exp(\bar{Y} + \boldsymbol{x}'\hat{\mathbf{V}}\hat{\boldsymbol{Q}}) \tag{5.8}$$

where $\bar{Y}$ is the mean of $\boldsymbol{Y}$, and $\boldsymbol{x}'$ is a row vector built upon $\boldsymbol{m}$ in exactly the same manner as the centered design matrix $\mathbf{X}_{\text{c}}$ is built upon $\mathbf{M}^t$, i.e. having the same structure and being centered exactly as any row in $\mathbf{X}_{\text{c}}$. The exponential function in (5.8) corresponds to the logarithm applied to the prices in the definition of $\boldsymbol{Y}$. If the prices are transformed in another manner (or not at all), the respective inverse function needs to be introduced here.

The optimal number A of factors in a PLS regression model is usually determined by cross-validation on the original data set. Several algorithms have been developed to carry out PLS regression. The calculations in Chapter 8 are going to be done with the 'modified kernel algorithm' proposed

by Dayal and MacGregor (1997) and implemented by Wehrens and Mevik (2006).

5.4 Model assessment

Starting from the idea that 'the functional form for hedonic functions should depend on the data, and not on some a priori reasoning' (Triplett, 2004, p. 188), there is a need for an overall measure of the predictive power of a model in order to find the one which best fits the data.

Davison and Hinkley (1997) present an approach of how to use bootstrap techniques to estimate such an *aggregate prediction error*. Their approach will now be applied to the context of hedonic functions. Let us assume that the prediction error of a single prediction $\hat{P}$ is measured by a *cost function* $c(P, \hat{P})$, usually an increasing function of $|P - \hat{P}|$. Overall prediction accuracy is then measured by

$$D(h, \hat{h}) = \mathrm{E}_{(P_+, \boldsymbol{M}_+)}\left[c(P_+, \hat{h}(\boldsymbol{M}_+))\right]$$

where $\hat{h} = \mathfrak{h}[\boldsymbol{p}, \mathbf{M}]$ is a fixed estimate of the hedonic function, and $(P_+, \boldsymbol{M}_+)$ is a new observation that stems from the same distribution as the original data. Clearly, as the true hedonic function h, i.e. the distribution of $(P, \boldsymbol{M})$, is unknown, D or, more precisely,

$$\Delta(\mathfrak{h}) = \mathrm{E}_{\hat{h}}(D(h, \hat{h})) = \mathrm{E}_{(P, \boldsymbol{M})}\left[\mathrm{E}_{(P_+, \boldsymbol{M}_+)}\left[c(P_+, \mathfrak{h}[\boldsymbol{P}, \mathbf{M}](\boldsymbol{M}_+))\right]\right]$$

needs to be estimated in practice. One is thus interested in the average prediction accuracy of models learnt by $\mathfrak{h}$ over all possible data sets of size N sampled from $(P, \boldsymbol{M})$.

A first approach to obtain such an estimate would be to use the same data set $(\boldsymbol{p}, \mathbf{M})$ as it was used to estimate the hedonic function. This leads to the *apparent error*

$$\hat{\Delta}_{\mathrm{app}} = D(\hat{h}, \hat{h}) = \frac{1}{N} \sum_{n=1}^{N} c(p_n, \hat{h}(\boldsymbol{m}_n))$$

which, however, tends to underestimate Δ. This is due to the fact that the same data set is used for the model fitting and the predictions. The mean difference between the true aggregate prediction error and the apparent error will be denoted by

$$e(\mathfrak{h}) = \mathrm{E}_{\hat{h}}[D(h, \hat{h}) - D(\hat{h}, \hat{h})] = \Delta(\mathfrak{h}) - E_{\hat{h}}[D(\hat{h}, \hat{h})]\,,$$

the *expected excess error*. The idea now is to build a bootstrap estimate $\hat{e}$ of $e(\mathfrak{h})$ in order to modify the apparent error into a reasonable estimate of Δ.

If one assumes that $\hat{h}_{\star r} = \mathfrak{h}[\boldsymbol{p}_{\star r}, \mathbf{M}_{\star r}]$ $(r = 1, \ldots, R)$ are bootstrap replications of $\hat{h}$, a bootstrap estimate $\hat{e}_{\mathrm{B}}$ of the expected excess error can be calculated by

$$\hat{e}_{\mathrm{B}} = \frac{1}{R} \sum_{r=1}^{R} \left[\frac{1}{N} \sum_{n=1}^{N} c(p_n, \hat{h}_{\star r}(\boldsymbol{m}_n)) - \frac{1}{N} \sum_{n=1}^{N} c(p_{\star rn}, \hat{h}_{\star r}(\boldsymbol{m}_{\star rn})) \right] . \quad (5.9)$$

Finally, a bootstrap estimate $\hat{\Delta}_{\mathrm{B}}$ of the aggregate prediction error is obtained by

$$\hat{\Delta}_{\mathrm{B}} = \hat{e}_{\mathrm{B}} + \hat{\Delta}_{\mathrm{app}} . \quad (5.10)$$

If different functional forms of hedonic functions or, to say it in the terminology used by Hastie et al. (2001), different learners $\mathfrak{h}$ are to be compared, it is necessary to define an ordering structure on the universe of all the possible learning algorithms. One such ordering could be obtained by defining a learner $\mathfrak{h}_1$ to be preferable to $\mathfrak{h}_2$, denoted by $\mathfrak{h}_1 \succ \mathfrak{h}_2$, if and only if $\Delta(\mathfrak{h}_1) < \Delta(\mathfrak{h}_2)$, i.e. if the aggregate prediction error of the resulting hedonic function estimate $\hat{h}$ is lower for $\mathfrak{h}_1$ than for $\mathfrak{h}_2$.

In other words, one would want to test the null hypothesis $\Delta(\mathfrak{h}_1) < \Delta(\mathfrak{h}_2)$ based on estimations of $\Delta(\mathfrak{h}_1)$ and $\Delta(\mathfrak{h}_2)$ in order to reject a learner $\mathfrak{h}_1$ in favour of $\mathfrak{h}_2$. This null hypothesis can alternatively be written as

$$\Delta(\mathfrak{h}_1) - \Delta(\mathfrak{h}_2) < 0 . \quad (5.11)$$

Having calculated $\hat{\Delta}^t_{\mathrm{B}}(\mathfrak{h}_1) - \hat{\Delta}^t_{\mathrm{B}}(\mathfrak{h}_2)$ for a set of independent time periods $t \in \{1, \ldots, T\}$, we are going to propose using a Wilcoxon signed rank test for testing nonparametrically the null hypothesis that the median of these estimated differences is less than zero. Rejecting this null hypothesis would then support the assumption that $\mathfrak{h}_2 \succ \mathfrak{h}_1$.

Empirical results using these methods are going to be presented in Section 8.4.

5.5 Further econometric considerations

A careful investigation of important econometric issues such as variable selection, outlier detection, the impact of unmeasured characteristics, heteroscedasticity of the random error terms and multicollinearity of the independent variables that are inherent in every regression problem would go

beyond the scope of this thesis. For estimating the hedonic functions in Part III, these issues are going to be dealt with on an *ad hoc* basis. No comprehensive treatment of these questions is therefore given here, but the applied methods will briefly be presented in Chapter 8, where applicable. For theoretical treatises of some of these aspects, the reader is referred to Triplett (2004) or to the more specific articles by Arguea and Hsiao (1993), Dickie et al. (1997), Andersson (2000), or Benkard and Bajari (2003).

CHAPTER 6

Estimating Hedonic Elementary Price Indices

6.1 Index estimators

In the previous chapter, we studied techniques for estimating the hedonic function $h^t(\boldsymbol{m}) = \mathrm{E}(P^t \mid \boldsymbol{M}^t = \boldsymbol{m})$ for a given elementary aggregate at a specific point t in time and for any characteristics vector $\boldsymbol{m}$. These estimates are now key elements for constructing estimators for hedonic elementary price indices.

Bilateral hedonic elementary price indices, as they were defined in (3.25) and (3.26), are characterised by the fact that they describe the price evolution between a base and a comparison period, 0 and 1, while holding the quality of the observed items constant. Any estimator of a specified hedonic elementary price index will therefore base on essentially three constituents, namely the two estimated hedonic functions $\hat{h}^0$ and $\hat{h}^1$ representing the quality-price relationship for both the base and the comparison periods, and the distribution $\mathbb{P}_{\boldsymbol{M}}$ of reference quality. Depending on the transformation function φ that is taken for the definition of the index, the estimator is going to have a different form.

6.1.1 Reference quality

It has already been discussed in Section 3.3 that, conforming to the way the reference quality distribution $\mathbb{P}_{\boldsymbol{M}}$ is specified, different index concepts can be generated. Among these are,

for instance, the different 'true' hedonic indices as outlined by BRACHINGER (2002).

The definitions (3.25) and (3.26) postulate that a hedonic elementary price index is built upon expectations of specific random variables with respect to $\mathbb{P}_{\boldsymbol{M}}$. Estimators for these expectations depend thus immediately on how this reference quality distribution is specified.

There are essentially two different approaches of specifying $\mathbb{P}_{\boldsymbol{M}}$ that need to be distinguished. The first of them consists of choosing deliberately one or several explicit items of an elementary aggregate for representing the reference quality. These items—let us denote them by $o_1, \ldots, o_N$—are then fixed over time, and only these are used for defining and estimating a quality-adjusted price index. In other words, the distribution of $\boldsymbol{M}$ is chosen such that exclusively the characteristics vectors $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$ where $\boldsymbol{m}_n := (m_1(o_n), \ldots, m_K(o_n))'$ for all $n \in \{1, \ldots, N\}$ have a non-zero probability. It remains open whether the same probability is attributed to all of these characteristics vectors, i.e. $\mathbb{P}(\boldsymbol{M} = \boldsymbol{m}_n) = 1/N$ for all n, or whether there is an individual weighting of each of these items. It is possible, for instance, to weigh each representant according to expenditure shares or sales volumes by choosing an appropriate probability value.

In this first approach, the elementary price index definitions (3.25) and (3.26) can be rewritten explicitly as

$$HPI^{0:1} = \frac{\varphi^{-1}\left(\sum_{n=1}^{N} \varphi(h^1(\boldsymbol{m}_n)) \times \mathbb{P}(\boldsymbol{M} = \boldsymbol{m}_n)\right)}{\varphi^{-1}\left(\sum_{n=1}^{N} \varphi(h^0(\boldsymbol{m}_n)) \times \mathbb{P}(\boldsymbol{M} = \boldsymbol{m}_n)\right)} \tag{6.1}$$

and

$$HPI^{0:1} = \varphi^{-1}\left(\sum_{n=1}^{N} \varphi\left(\frac{h^1(\boldsymbol{m}_n)}{h^0(\boldsymbol{m}_n)}\right) \times \mathbb{P}(\boldsymbol{M} = \boldsymbol{m}_n)\right). \tag{6.2}$$

Estimators for both of these indices are obtained straightaway by replacing the hedonic functions h^0 and h^1 by estimators $\hat{h}^0$ and $\hat{h}^1$ respectively in the above formulae.

The second fundamental approach for specifying $\mathbb{P}_{\boldsymbol{M}}$ is to interpret $\boldsymbol{M}$ as being the outcome of a random draw from a certain population of quality vectors. This population—we call it $\mathcal{M}^*$—might consist of all characteristics vectors that exist on the market at a specific point in time or over a certain time period. In this case, the expectation with respect to $\mathbb{P}_{\boldsymbol{M}}$ cannot be calculated explicitly like in the first approach, since the whole population

$\mathcal{M}^*$ is not known to the observer in general. This expectation needs to be estimated.

Due to the fact that the individual characteristics $m_1, \ldots, m_K$ may be very diverse in their nature—they may be measured on any thinkable scale—and that there may be interactions, it seems difficult to define and estimate a parametric probability model for $\mathbb{P}_{\boldsymbol{M}}$, from where the expectation could be extracted. A more practicable approach is therefore to estimate $\mathbb{P}_{\boldsymbol{M}}$ empirically on the basis of a random sample of elements of $\mathcal{M}^*$.

Assume that such an i.i.d. sample is given by $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N$ where $\boldsymbol{M}_n \overset{\text{L}}{\sim} \boldsymbol{M}$ for all $n \in \{1, \ldots, N\}$. The hedonic elementary price indices (3.25) and (3.26) can then be estimated empirically by

$$\widehat{HPI}^{0:1} = \frac{\varphi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(\hat{h}^1(\boldsymbol{M}_n))\right)}{\varphi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(\hat{h}^0(\boldsymbol{M}_n))\right)} \tag{6.3}$$

and

$$\widehat{HPI}^{0:1} = \varphi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\varphi\left(\frac{\hat{h}^1(\boldsymbol{M}_n)}{\hat{h}^0(\boldsymbol{M}_n)}\right)\right), \tag{6.4}$$

respectively. These formulae are almost identical to estimators of (6.1) and (6.2) for the case where all the representants $\boldsymbol{m}_1, \ldots \boldsymbol{m}_N$ are equiprobable. The only difference is that, now, there is a third source of randomness, apart from the random variables $\hat{h}^0$ and $\hat{h}^1$ already present in the estimators of (6.1) and (6.2). This third source is the sample $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N$ which, in contrast to the first approach, is not fixed in advance.

If these two approaches for specifying the reference quality are compared, it is difficult to identify a favourite. The first of them, where the reference characteristics vectors are explicitly chosen, has the advantage that the reference quality is known exactly. In the second approach, contrarily, reference quality is not known completely and needs to be estimated. The latter might, however, be more representative for the effective quality spectrum available on the market than the first approach, where only a restricted set of representants is chosen. The price of this potential gain in representativeness is an increased variability of the price index estimator. This is due to the fact that the reference quality needs to be estimated based on a random sample.

Since for hedonic indices, an extensive database of product characteristics is in general available from the estimation of the hedonic functions, the

additional cost for estimating $\mathbb{P}_{\boldsymbol{M}}$ following the idea of the second approach should be negligible. Certainly, a straightforward approach is to estimate the reference quality distribution by some or even all of the items in this database. Moreover, assuming that this data builds a representative sample of the transactions for a certain elementary aggregate in a specific market over a certain time period, an implicit weighting of the individual items according to their sales volumes would even automatically take place.

6.1.2 Bilateral hedonic elementary price index estimators

In the estimators presented in the previous section, the transformation function φ remained unspecified. Depending on the form of this function, different candidate index formulae can now be outlined. In Chapter 4, we discussed that, in order to satisfy some of the axioms, φ should be chosen such that $\varphi(\lambda\, x) = \varphi(\lambda)+\varphi(x)$ or $\varphi(\lambda\, x) = \varphi(\lambda)\,\varphi(x)$ for all $\lambda, x \in \mathbb{R}$. Moreover, it was shown that $\varphi(x) = x$, $\varphi(x) = \ln x$, and $\varphi(x) = x^{-1}$ are interesting candidates of such transformation functions.

Table 6.1 displays the bilateral hedonic price index estimators that are generated when using the index estimators (6.3) and (6.4) together with all three proposed transformation functions φ. Most of the elementary index formulae used in practice (see e.g. ILO et al., 2004, pp. 360–1) can be reproduced like that, namely those attributed to Dutot, Jevons, Carli, and the one that is called 'Harmonic Carli' here. Moreover, we find a 'Harmonic Dutot' index estimator which does not appear in the mentioned literature.

From the axiomatic point of view adopted in Chapter 4, it is the Dutot, Jevons and 'Harmonic Dutot' index estimators that are serious candidates for being used in practice, since for the others, the underlying indices do not generally satisfy the time reversal and circularity tests. Jevons is certainly a particularly attractive candidate since it evolves from both index estimators, (6.3) and (6.4). Moreover, there is further theoretical support for this formula which might also be adaptable in this context (see ILO et al., 2004, p. 371). Yu (2003) provides empirical support for the choice of the Jevons formula in a hedonic price index framework and, more generally, Silver and Heravi (2006) show that the difference between the Dutot and the Jevons indices in a matched-model framework depends mainly on the change in the dispersion of log prices over time.

6.1.3 Double imputation

In the estimators proposed in Table 6.1, only imputed (estimated) prices are used even though observed prices might be available for certain characteristics vectors $\boldsymbol{m}_n$ in one or the other time period. These formulae follow thus the 'double imputation'

Table 6.1: **Bilateral hedonic elementary price index estimators**

Formula	Transformation	Resulting estimator	Index type
(6.3)	$\varphi(x) = x$	$\hat{I}_{\mathrm{D}}^{0:1} = \frac{\sum_{n=1}^{N} \hat{h}^1(\boldsymbol{M}_n)}{\sum_{n=1}^{N} \hat{h}^0(\boldsymbol{M}_n)}$	Dutot
	$\varphi(x) = \ln x$	$\hat{I}_{\mathrm{J}}^{0:1} = \sqrt[N]{\prod_{n=1}^{N} \frac{\hat{h}^1(\boldsymbol{M}_n)}{\hat{h}^0(\boldsymbol{M}_n)}}$	Jevons
	$\varphi(x) = x^{-1}$	$\hat{I}_{\mathrm{HD}}^{0:1} = \frac{\left(\sum_{n=1}^{N} \left(\hat{h}^1(\boldsymbol{M}_n)\right)^{-1}\right)^{-1}}{\left(\sum_{n=1}^{N} \left(\hat{h}^0(\boldsymbol{M}_n)\right)^{-1}\right)^{-1}}$	'Harmonic Dutot'
(6.4)	$\varphi(x) = x$	$\hat{I}_{\mathrm{C}}^{0:1} = \frac{1}{N} \sum_{n=1}^{N} \frac{\hat{h}^1(\boldsymbol{M}_n)}{\hat{h}^0(\boldsymbol{M}_n)}$	Carli
	$\varphi(x) = \ln x$	$\hat{I}_{\mathrm{J}}^{0:1} = \sqrt[N]{\prod_{n=1}^{N} \frac{\hat{h}^1(\boldsymbol{M}_n)}{\hat{h}^0(\boldsymbol{M}_n)}}$	Jevons
	$\varphi(x) = x^{-1}$	$\hat{I}_{\mathrm{HC}}^{0:1} = \left(\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{h}^1(\boldsymbol{M}_n)}{\hat{h}^0(\boldsymbol{M}_n)}\right)^{-1}\right)^{-1}$	'Harmonic Carli'

method as outlined in detail by TRIPLETT (2004, p. 70 ff.). Moreover, in this aspect, they contradict the practice of most of the hedonic index studies, where prices are imputed only for items entering or exiting the market during the considered time period, whereas observed prices are used wherever possible ('single imputation').

The question of whether or not such a double imputation is admissible, has been discussed with scepticism in the mentioned literature. TRIPLETT concludes that this matter 'is not settled, and depends ... on one's interpretation of hedonic residuals'. PAKES (2003, p. 1589) empirically compares the resulting index values in a similar context under both regimes and discovers that they are 'virtually identical'.

The reason for the appearance of this double imputation approach in the present context is based on the manner the estimators were developed. The adoption of the hedonic econometric model led to equation (3.17) where, at any time t, the price P^t of a randomly chosen item of an elementary aggregate was written as a function $h^t(\boldsymbol{M}^t)$ of its price-relevant characteristics expressions plus a quality-independent component ϵ^t having mean zero. For being able to control for quality differences, hedonic elementary price indices were then defined as ratios of conditional expectations of P^t given a reference quality distribution $\mathbb{P}_{\boldsymbol{M}}$. With the introduction of estimators of the indices, estimations of all these conditional expectations were implemented as predictions from the estimated hedonic functions as they were developed in Chapter 5.

From this point of view, double imputation perfectly conforms to the theory. If observed prices were used in the index estimators, one would introduce unsystematic price residuals into the calculations. Using the estimators in Table 6.1 allows, in contrast, to get rid of these unsystematic price components, provided—and this might be the essential problem in practice—that the hedonic functions are correctly specified. In fact, whether or not there is a large difference between indices estimated by single or double imputation estimators depends largely on the variance of the quality-independent price component ϵ^t and on the prediction error of the hedonic functions (cf. PAKES, 2005, p. 22).

HULTEN's (2003, p. 9) belief that 'a large part of the problem reflects a lower degree of confidence in data that are imputed using regression analysis' seems to point out exactly the reason why double imputation is rarely used in practice and vehemently criticised in theory, e.g. by TRIPLETT (2004). 'Price estimates', HULTEN states, 'collected directly from an underlying population are generally regarded as "facts".' If, in contrast, 'the price is inferred using

regression techniques, it becomes a "processed" fact subject to researcher discretion'.

Assuming, however, that the hedonic hypothesis holds and that hedonic functions can be estimated sufficiently well, the theory developed here nevertheless supports the use of double imputation techniques for estimating hedonic elementary price indices. Moreover, it is important to note that the formulae proposed here go further than most of the propositions in the literature in the sense that prices are imputed for *all* the representative characteristics vectors. In the literature, hedonic price imputations (single or double) are usually only done, if a matching of raw prices between the base and the current time periods is not possible, while for the items where a matching is possible, conventional matched-model indices are calculated. Examples of such 'hybrid' indices are proposed, discussed, and applied by, e.g., Pakes (2003), de Haan (2004), and Van Reenen (2005).

It is also from this last point of view that Triplett (2004, p. 70 ff.) argues that, 'If the hedonic function is correctly specified, then it seems incontrovertible that double imputation creates error', since, among other reasons, 'the regression line does not provide the best estimate of the price when the price ... is actually observed'. Silver and Heravi (2004a), in contrast, propose sales weighted hedonic indices where prices are imputed for both matched and unmatched items. According to Silver's report on the seventh meeting of the 'Ottawa' International Working Group on Price Indices (2003), 'There remains the issue of whether imputations for the whole sample are preferable to imputations for only the unmatched sample using the double imputation method'. It seems that this question needs to be discussed further.

6.1.4 The time dummy variable method

An important special case of a quality-adjusted price index estimator is the time dummy variable method (see e.g. Griliches, 1971a, p. 59, Silver and Heravi, 2003, pp. 280–1, or Triplett, 2004, p. 48–55). There, the bilateral hedonic elementary price index to be estimated is directly a parameter of the hedonic function, which is estimated based on the pooled observations of both the reference and the current period.

It is interesting to see that this approach still is a special case of the price index concept presented here. Assume that $\mathfrak{h}$ is a 'learner', which maps the space of the price and characteristics observations to the set $\mathcal{H}$ of all potential hedonic functions as presented in Chapter 5. The hedonic function

estimator $\hat{h}^t$ for time $t \in \{0, 1\}$ may then be defined as

$$\hat{h}^t(\boldsymbol{m}) = \mathfrak{h}[\boldsymbol{P}^{0:1}, \mathbf{M}^{0:1}]((\boldsymbol{m}', t)')$$

where $\boldsymbol{P}^{0:1} = (\boldsymbol{P}^{0\prime}, \boldsymbol{P}^{1\prime})'$ is the vector of the pooled price observations for periods 0 and 1 and

$$\mathbf{M}^{0:1} = \begin{pmatrix} \mathbf{M}^0 & \mathbf{0} \\ \mathbf{M}^1 & \mathbf{1} \end{pmatrix}$$

is the matrix of the pooled characteristics expressions for periods 0 and 1, augmented by a dummy variable indicating the period an observation stems from. Hereby, $\mathbf{0}$ and $\mathbf{1}$ stand for appropriate vectors containing only zeros and ones respectively.

If, for the hedonic function, the semi-logarithmic functional form (5.4) is adopted and $\beta_{K+1}^{0:1}$ denotes the coefficient of the time dummy variable, it is evident that

$$\hat{h}^1(\boldsymbol{m}) = \exp(\hat{\beta}_{K+1}^{0:1}) \times \hat{h}^0(\boldsymbol{m})\,.$$

Due to this relation, all the estimators given in Table 6.1 can, in this case, be simplified to $\exp(\hat{\beta}_{K+1}^{0:1})$ and are thus completely independent of the reference characteristics vectors $\boldsymbol{M}_n$, $n \in \{1, \ldots, N\}$ (compare Brachinger, 2002, p. 12).

The time dummy variable method, however, has been criticised in the literature, mainly because the coefficients of the characteristics variables are in this approach implicitly assumed to remain constant over the two pooled time periods (see e.g. Silver and Heravi, 2003, 2004a, and 2004b).

6.2 Bootstrap replications of hedonic elementary price indices

6.2.1 Estimation error and confidence intervals

The index formulae given in Table 6.1 are all point estimators of the respective bilateral hedonic elementary price indices (6.3) and (6.4). Given that the hedonic price predictors $\hat{h}^0$ and $\hat{h}^1$ as well as, in general, the sample of reference characteristics $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N$ are random variables, this randomness propagates to any estimator $\hat{I}^{0:1}$. However, trying to deduce the probability distribution of $\hat{I}^{0:1}$ analytically from those of its constituents seems ambitious given the potential complexity of the hedonic functions and the generally unknown form of $\mathbb{P}_{\boldsymbol{M}}$. Yet, it is possible to employ Monte-Carlo simulations to estimate this distribution empirically.

Let us define the random *estimation error* of the hedonic elementary price index estimator $\hat{I}^{0:1}$ by

$$\zeta^{0:1} := \hat{I}^{0:1} - I^{0:1} .$$

A sensible method for qualifying the estimator $\hat{I}^{0:1}$ is the estimation of confidence intervals for the underlying index. An equitailed $(1 - 2\alpha)$ confidence interval for $I^{0:1}$ is given by

$$\left[\hat{I}^{0:1} - \zeta^{0:1}_{1-\alpha}, \hat{I}^{0:1} - \zeta^{0:1}_{\alpha}\right] \tag{6.5}$$

where ζ^t_α and $\zeta^{0:1}_{1-\alpha}$ are the α and $(1 - \alpha)$ quantiles of $\zeta^{0:1}$ respectively.

In order to determine these quantiles, some knowledge on the probability distribution of $\zeta^{0:1}$ is needed. Bootstrap resampling methods, as they are extensively described, e.g., by Davison and Hinkley (1997), provide one manner of acquiring such knowledge. Their main idea is to use computer simulations for generating an empirical approximation of the distribution of interest based on new arrangements of the input data. These arrangements are essentially random samples drawn with replacement from the original data set.

In this sense, the distribution of $\zeta^{0:1}$ may be estimated by the empirical distribution of

$$\zeta^{0:1}_\star := \hat{I}^{0:1}_\star - \hat{I}^{0:1} . \tag{6.6}$$

Here, $\hat{I}^{0:1}_\star$ are bootstrap replications of the hedonic price index, while $\hat{I}^{0:1}$ remains fixed. The following section is going to discuss how such replications of $\hat{I}^{0:1}_\star$ can be acquired.

6.2.2 Resampling methods The process of generating the replications $\hat{I}^{0:1}_\star$ can reasonably be split into two sub-problems. Any estimator $\hat{I}^{0:1}$ is a function of $\hat{h}^0$, $\hat{h}^1$, and of the realisations $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$ of the reference characteristics sample. We can thus independently generate first replications $\hat{h}^0_\star$ and $\hat{h}^1_\star$ of $\hat{h}^0$ and $\hat{h}^1$ respectively, and then replications $\boldsymbol{m}_{\star 1}, \ldots, \boldsymbol{m}_{\star N}$ of $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$. The second step will be left out, if $\mathbb{P}_{\boldsymbol{M}}$ puts positive probability on a discrete, fixed and known set of reference characteristics vectors only (cf. Section 6.1.1). We are now going to present three different approaches for tackling the first sub-problem, i.e. the generation of replications of the hedonic function estimators.

A first and most generally applicable bootstrap approach is the so-called *case-based resampling* procedure (see Davison and Hinkley, 1997, Section 6.2.4). For both periods $t \in \{0, 1\}$, a sample of N^t price-characteristics combinations is drawn with replacement from the N^t original data points at

period t. This new sample is then used as an input to the regression algorithm $\mathfrak{h}$ from where simulated values of $\hat{h}^0_\star$ and $\hat{h}^1_\star$ are obtained. Repeating this procedure R times leads to R new estimates of the hedonic function in both time periods, from where R replications of $\zeta^{0:1}_\star$ can be deduced.

Algorithm 6.1 (Case-based resampling).
For $r = 1, \ldots, R$, for $t \in \{0, 1\}$,

1. sample $\nu^t_{\star 1}, \ldots, \nu^t_{\star N^t}$ randomly with replacement from $\{1, \ldots, N^t\}$
2. for $n = 1, \ldots, N^t$, set $p^t_{\star n} = p^t_{\nu^t_{\star n}}$, $\boldsymbol{m}^t_{\star n} = \boldsymbol{m}^t_{\nu^t_{\star n}}$;
3. compute $\hat{h}^t_{\star r} := \mathfrak{h}[\boldsymbol{p}^t_\star, \mathbf{M}^t_\star]$, where $\boldsymbol{p}^t_\star = (p^t_{\star 1}, \ldots, p^t_{\star N^t})$ and $\mathbf{M}^t_\star = (\boldsymbol{m}^t_{\star 1}, \ldots, \boldsymbol{m}^t_{\star N^t})$. ◇

Pakes (2003) and Van Reenen (2005) apparently used such a case-based bootstrap procedure for the estimation of standard errors of hedonic elementary price indices.

A second approach is the *model-based resampling*, where the resampling takes place on the residuals of the original model (see Davison and Hinkley, 1997, Sect. 6.2.3). The underlying idea here is that, given the model (3.17), the regression residuals are estimates of the random errors ϵ^t. A simulated price for a certain characteristics vector $\boldsymbol{m}^t_n$ can thus be obtained by adding such a residual to the regression fit $\hat{h}^t(\boldsymbol{m}^t_n)$.

An important condition of the model-based approach is that the residuals involved need to be suitable for simulating the distribution of the random errors ϵ^t. In other words, the raw residuals $e^t_n = p^t_n - \hat{h}^t(\boldsymbol{m}^t_n)$ need in general to be modified in such a manner that they have the same variance as ϵ^t before they can be used for a model-based resampling. In the case of linear regression, i.e. if (3.17) is specified by $\boldsymbol{P}^t = \mathbf{X}^t\boldsymbol{\beta}^t + \boldsymbol{\epsilon}^t$ where $\mathbf{X}^t$ is the design matrix containing the data $\mathbf{M}^t$ plus a constant, the vector of raw residuals $\boldsymbol{e}^t = (e^t_1, \ldots, e^t_{N^t})$ can be written as $\boldsymbol{e}^t = (\mathbf{I} - \mathbf{H}^t)\boldsymbol{\epsilon}^t$ where $\mathbf{H}^t = \mathbf{X}^t(\mathbf{X}^{t\prime}\mathbf{X}^t)^{-1}\mathbf{X}^{t\prime}$ is the *hat matrix* of the regression model at time t. Therefore, Davison and Hinkley (1997, p. 261) recommend to work with the modified residuals

$$r^t_n = \frac{p^t_n - \hat{h}^t(\boldsymbol{m}^t_n)}{(1 - h^t_n)^{1/2}} \tag{6.7}$$

where h^t_n is the nth diagonal element of $\mathbf{H}^t$, because their variances match with those of ϵ^t. This makes sense if the standard assumption of homoscedasticity of the ϵ^t is tenable. In order to get proper estimates of ϵ^t having mean zero, the values of r^t_n are finally going to be re-centered by individually subtracting their average $\bar{r}^t$.

Algorithm 6.2 (Model-based resampling).
For $r = 1, \ldots, R$,

1. for $t \in \{0, 1\}$, for $n = 1, \ldots, N^t$,
   a) sample $\epsilon^t_{\star n}$ from $r^t_1 - \bar{r}^t, \ldots, r^t_{N^t} - \bar{r}^t$;
   b) compute the simulated response $p^t_{\star rn} = \hat{h}^t(\boldsymbol{m}^t_n) + \epsilon^t_{\star n}$;
2. compute $\hat{h}^t_{\star r} := \mathfrak{h}[\boldsymbol{p}^t_{\star r}, \mathbf{M}^t]$, where $\boldsymbol{p}^t_{\star r} = (p^t_{\star r1}, \ldots, p^t_{\star rN^t})$. ◇

Inconvenient in the model-based resampling is the fact that properly modified, i.e. variance-adjusted residuals may not be easy to acquire, especially when the regression approach $\mathfrak{h}$ is complicated. In nonlinear regression models, for instance, the so-called leverages h^t_n are not available straight away. Moreover, the assumption of homoscedasticity of the random errors ϵ^t may not be justified. This is particularly true if one assumes that the model residuals ϵ^t are not just random noise but economically significant. (See Reis and Santos Silva, 2002 or Triplett, 2004, p. 186 ff. for some comments on the issue of heteroscedastic error terms in hedonic regressions.)

Case-based resampling, in contrast, is always applicable and does not depend on any assumption about ϵ^t. Yet, Davison and Hinkley (1997, p. 264 ff.) identify two disadvantages of case-based compared to model-based resampling. First, they state that case-based estimations might be inefficient if the constant-variance model is correct, and, secondly, they argue that case-based simulations lead to simulated samples with different designs, because the vectors $\boldsymbol{m}^t_{\star 1}, \ldots, \boldsymbol{m}^t_{\star N^t}$ are randomly sampled. The design matrix of a regression model, however, 'fixes the information content of a sample, and in principle our inference should be specific to the information in our data'.

A third approach which overcomes the disability of the model-based approach to cope with heteroscedastic error terms is the so-called *wild bootstrap* originally proposed by Liu (1988) and developed further by Davidson and Flachaire (2001). Here, the error terms $\epsilon^t_1, \ldots, \epsilon^t_{N^t}$ are still assumed to be mutually independent and to have a common mean of zero, but they may be heteroscedastic with $\mathrm{E}[(\epsilon^t_n)^2] = (\sigma^t_n)^2$. The error term, in this case, may be written as $\epsilon^t_n = \sigma^t_n \, v^t_n$ where $\mathrm{E}[v^t_n] = 0$ and $\mathrm{E}[(v^t_n)^2] = 1$. Correspondingly, the simulated prices $p^t_{\star rn}$ are no longer obtained by adding to $\hat{h}^t(\boldsymbol{m}^t_n)$ any $\epsilon^t_{\star n}$ sampled from the centered modified residuals, but by adding the corresponding modified residual r^t_n multiplied by a random number $\epsilon^t_{\star n}$ drawn from a completely independent auxiliary distribution having mean zero and variance one. For the case of hypothesis testing, both Davidson and Flachaire (2001) and MacKinnon (2002) advise to draw the numbers

$\epsilon^t_{\star n}$ from the Rademacher distribution, i.e. from a discrete random variable having either the value -1 or 1 with probability ½ each. For the estimation of confidence intervals, there is weaker evidence for this specific choice and other alternatives such as the one proposed by Mammen (1993) might also be appropriate. Nevertheless, the Rademacher distribution is going to be proposed in the following third resampling approach.

Algorithm 6.3 (Wild bootstrap).
For $r = 1, \ldots, R$,

1. for $t \in \{0, 1\}$, for $n = 1, \ldots, N^t$,
   a) sample $\epsilon^t_{\star rn}$ from a discrete random variable having either the value -1 or 1 with probability 1/2 each;
   b) compute the simulated response $p^t_{\star rn} = \hat{h}^t(\boldsymbol{m}^t_n) + r^t_n \, \epsilon^t_{\star rn}$;
2. compute $\hat{h}^t_{\star r} := \mathfrak{h}[\boldsymbol{p}^t_{\star r}, \mathbf{M}^t]$, where $\boldsymbol{p}^t_{\star r} = (p^t_{\star r1}, \ldots, p^t_{\star r N^t})$. ◇

The main advantage of the wild bootstrap compared to the model-based approach is its ability to incorporate heteroscedastic error terms. Moreover, it shares the property of the model-based approach concerning the unmodified regression design. However, it is an open question whether its efficiency is still better than the one of the case-based approach.

The three approaches presented so far provide solutions for what has been called the first sub-problem in generating replications $\hat{I}^{0:1}_\star$ of hedonic elementary price indices. The second issue is the generation of replications of the reference characteristics vectors. Here, however, only case-based resampling is applicable. In other words, the replications $\boldsymbol{m}_{\star 1}, \ldots, \boldsymbol{m}_{\star N}$ need to be sampled individually with replacement from $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$.

Regardless of the approach that is chosen for the simulation of $\hat{h}^0_\star$ and $\hat{h}^1_\star$, a sample of simulated estimation errors $\zeta^{0:1}_{\star 1}, \ldots, \zeta^{0:1}_{\star R}$ is finally calculated by (6.6). If, for instance, the Jevons formula is used for estimating the respective indices, $\zeta^{0:1}_{\star R}$ can be calculated by

$$\zeta^{0:1}_{\star R} := \sqrt[N]{\prod_{n=1}^{N} \frac{\hat{h}^1_\star(\boldsymbol{m}_{\star n})}{\hat{h}^0_\star(\boldsymbol{m}_{\star n})}} - \sqrt[N]{\prod_{n=1}^{N} \frac{\hat{h}^1(\boldsymbol{m}_n)}{\hat{h}^0(\boldsymbol{m}_n)}} \,. \tag{6.8}$$

Furthermore, if the increasingly ordered values of $\zeta^{0:1}_{\star r}$ $(r = 1, \ldots, R)$ are denoted by $\zeta^{0:1}_{\star[1]}, \ldots, \zeta^{0:1}_{\star[R]}$, an estimate of the confidence interval (6.5) is given by

$$\left[\hat{I}^{0:1} - \zeta^{0:1}_{\star[(R+1)(1-\alpha)]}, \hat{I}^{0:1} - \zeta^{0:1}_{\star[(R+1)\alpha]}\right] . \tag{6.9}$$

For this reason, the number R of bootstrap replications has to be chosen such that $(R+1)\alpha$ is an integer.

It should be noted that, at this stage, the regression approach $\mathfrak{h}$ remains completely unspecified. Moreover, these resampling approaches do not depend on the specification of the hedonic price index by, e.g., the Jevons formula. They are similarly applicable to all bilateral hedonic price index estimators of the form (6.3) or (6.4).

PART III

Hedonic Elementary Price Indices for Used Cars

CHAPTER 7

Data Source: The AutoScout24 Marketplace

7.1 Why used cars?

The decision to work with used cars data in the empirical part of the present thesis was largely influenced by the availability of an appropriate data set. The demands on such data are high if they shall be used for hedonic regressions. The hedonic econometric model assumes that there is a finite set of price-relevant characteristics which are perfectly sufficient for modelling the quality-price relationship. No indication is given, however, how these characteristics can be determined in practice.

'Getting the characteristics right is not just the first step in estimating a hedonic function, it is the most important step', says Triplett (2004, p. 140). While a data-driven variable selection using econometric techniques may take place routinely when searching for the price-relevant characteristics of an elementary aggregate, it is even more important to first get measurements of a sufficiently extensive set of variables to choose from.

When statistical offices start the estimation of a hedonic elementary price index, they need to have the (financial) means for investigating together with experts which variables are potentially price-relevant for a specific elementary aggregate. Only then, they may develop a strategy of how these variables can be measured for a representative sample of the population.

Such an approach was not feasible for the present piece of work. The empirical analysis here rather had to base on measurements of variables

that were already available in a properly designed database. Such a database could be found for the market of used cars in Switzerland. It was a great opportunity that the private company AutoScout24, maintaining an internet platform for trading cars (see www.autoscout24.ch), accepted to open its database for the purpose of price index research.

According to AutoScout24, more than half of the 700 000 used cars traded per year in Switzerland find their new owner through their internet platform (see Settele, 2005). Due to their cooperation with about 3000 car merchants, the cars offered on the platform do not only stem from private persons but to a large part from institutional dealers.

Cars—either new or used—have frequently been analysed in hedonic studies. Beginning with Court (1939), one of the very first articles on hedonic price indices, automobile prices attracted the most attention in hedonic analyses up to Griliches (1971b, p. 3). The list of examples mentioned therein extends to the present with, e.g., Griliches (1971a), Bitros and Panas (1988), Gordon (1990), Arguea and Hsiao (1993), Murray and Sarantis (1999), Reis and Santos Silva (2002), or van Dalen and Bode (2004), ranging from theoretical to applied studies, e.g. for specific countries or markets. In general, hedonic price equations are not only estimated with the aim of building a quality-adjusted price index but also for studying the economic behaviour of consumers with regard to certain characteristics of a product. Boulding and Purohit (1996) is one such example.

Several national statistical agencies currently use hedonic approaches for estimating their elementary price indices for used cars. Amongst these are New Zealand, Germany, the Netherlands, Finland, and Sweden (see Nair, 2004, German Federal Statistical Office, 2003, and Ahnert and Kenny, 2004). In the US, according to the Boskin report (Boskin et al., 1996), 'The CPI index for used cars has long been known to be upward biased, simply because no quality adjustments were applied to this category at all'. This changed in 1987 where the Bureau of Labor Statistics (BLS) began adjusting the used cars elementary index for quality changes, although not with hedonic methods. This practice was reviewed by Pashigian (2001), but he did not mention any explicit need for changing towards a hedonic elementary price index for used cars.

Even though there may be goods such as personal computers and their operating systems, where the change in quality is more rapid and more prevalent (cf. White et al., 2004, p. 19), applying hedonic techniques for the used cars elementary price index seems to be an acceptable choice. Moreover, used cars are currently the tenth most highly weighted elementary aggregate in the Swiss CPI (see Table 7.1). With a weight of 1.253 percent, the

Table 7.1: **Current weights in percent and historical comparison of the 15 most important elementary aggregates in the Swiss CPI**

Elementary aggregate	Current weight	Historical weights		
	2006	2000	2002	2004
Rentals for flats	18.724	20.143	20.093	19.637
Hospital services	6.418	6.287	5.56	6.049
Medical services	3.906	2.897	3.497	3.746
Meals in restaurants and cafés	3.627	4.185	4.141	3.659
Pharmaceutical products	2.907	1.848	2.407	2.735
New motor cars	2.610	1.816	2.148	2.509
Petrol	2.557	–	–	–
Dental services	1.763	1.457	1.596	1.670
Fuel oil	1.683	1.281	0.926	1.181
Used motor cars	**1.253**	**0.950**	**1.080**	**0.871**
Package holidays	1.181	1.342	1.475	1.239
Mobile telephone services	1.109	–	–	–
Passenger transport services (direct)	1.106	1.094	1.083	1.046
Fixed-line telephone services	1.089	–	–	–
Other outpatient services	1.063	0.550	0.954	1.041

Source: Swiss Federal Statistical Office. The elementary aggregates 'Petrol', 'Mobile telephone services', and 'Fixed-line telephone services' were contained in higher level aggregates until 2005.

impact of the elementary index for used cars on the whole CPI is comparably important.

7.2 Data retrieval, storage, structure, and analysis

7.2.1 Data import Fig. 7.1 displays the infrastructure that has been installed for data retrieval, storage, and analysis. As AutoScout24 did not keep a history of the published advertisements over a period that was sufficiently long for estimating price indices, we opted for a daily transfer of the currently advertised cars. For this purpose, AutoScout24 installed an export script on their server that nightly stored the published advertisements along with some related tables to comma-separated ASCII files.

Table 7.2 lists the data fields contained therein for each advertisement. They consist of 51 characteristics variables of the advertised car (including the 34 equipment dummies referenced by `cars_equipment` and listed in Table 7.3). Furthermore, there are four variables describing the publication

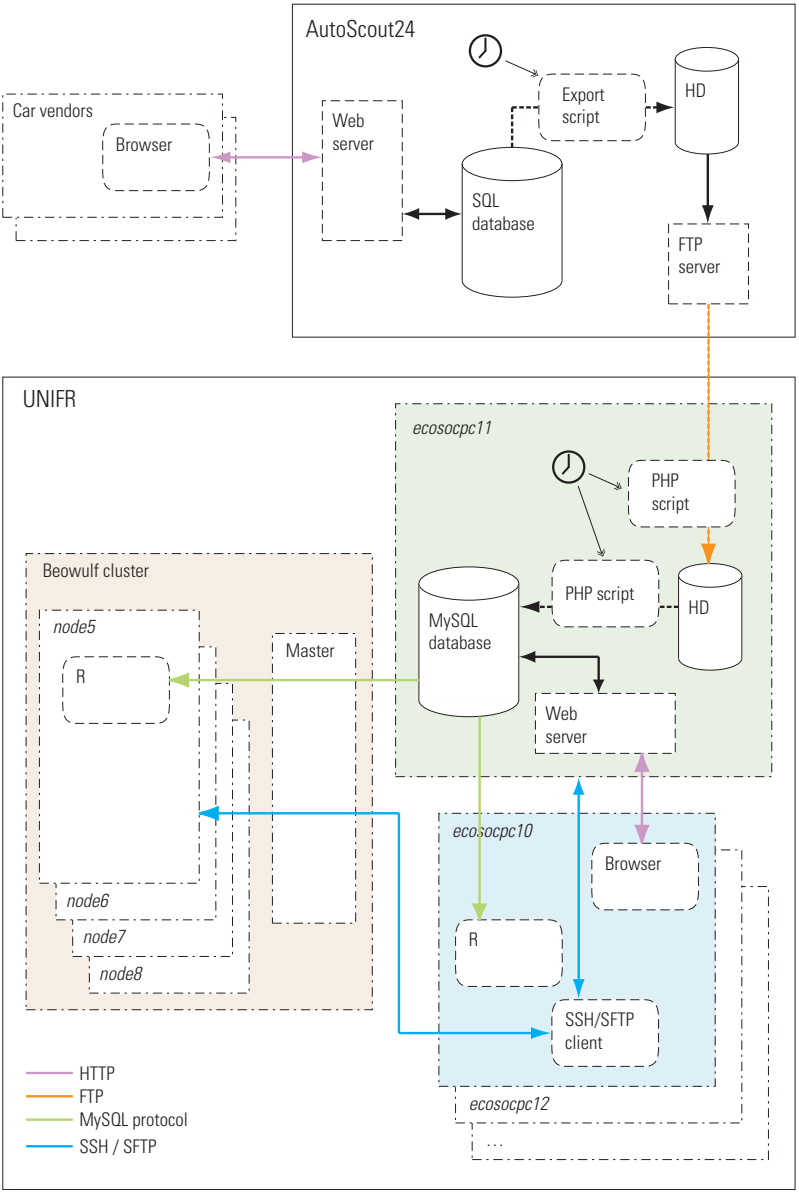


Figure 7.1: **Network layout for data retrieval and analysis**

date, the date of the last update, the type of advertisement (`cars_adtype`) and the number of page impressions for each advertisement. Finally, there is a unique identifier (`cars_car_id`) as well as the postcode of the advertiser.

At the Seminar of Statistics, a personal computer (`ecosocpc11` in Fig. 7.1) was installed for the purpose of retrieving, managing, and supplying the data for further analysis. Each night, a scheduled PHP script used to fetch by FTP from AutoScout24 the exported files (about 20 to 30 megabytes of data each time) and to store them to a local hard disk. Great care had to be directed to the handling of errors and exceptions during the file transfer. Each snapshot provided by AutoScout24 had to be fetched within 24 hours before it was overwritten with new data and thus lost for any further analysis.

Once the data was stored locally, another PHP script managed the integration of the new data into a MySQL database. In a first step, all the auxiliary reference tables were updated with the newly arrived content, and the list of advertisements was read into a temporary table named `cars`.

In order to keep a complete history of the advertised cars, the content of `cars` was then integrated into an all-embracing table named `carhist` containing all the advertisements that had existed at any previous date. In addition to the fields already available in `cars`, the table `carhist` contained two fields `from_dt` and `to_dt` giving the start and the end date of each entry. Moreover, there was a field named `car_revision` that numbered the different versions of an advertisement over time.

The integration of the newly arrived advertisements (`cars`) into `carhist` was done in four steps. First, all the entries in `carhist` that were no longer available in `cars` were identified, and their field `to_dt` was set to the date and time when the current snapshot was exported by AutoScout24. This value was determined through the time of the last modification of the respective file on their FTP server.

Secondly, all the advertisements in `cars` that did not yet figure in `carhist` were imported. For each of them, a new entry was made in `carhist` inheriting from the original all the characteristics of the advertised car. The start date `from_dt` was set to the value given in `cars_changed` and `to_dt` to a virtual date in the indefinite future (31.12.9999 23:59:59). The revision number of a newly appeared item was set to 0.

The third step dealt with advertisements in `carhist` that were still available in `cars` but had been changed in between. For this purpose, the date and time contained in `cars_changed` was compared to `from_dt` and `to_dt` of the entries in `carhist` corresponding to each advertisement. If `cars_changed` happened to lie between the start and the end date of an entry in `carhist`, its

Table 7.2: **Data fields received from AutoScout24**

Data field	Description
cars__car__id	Identification number of advertisement (primary key)
cars__cartype__id	Car type code (new, used, demonstration, old-timer)
cars__make__id	Make code (e.g. VW, Opel, BMW, Audi, ...)
cars__model__id	Model code (e.g. Golf, Astra, A4, Passat, ...)
cars__model__full	Full model description (e.g. '2.2dCi Authentique')
cars__body__id	Body code (e.g. cabriolet, limousine, coupé, ...)
cars__doors	Number of doors
cars__seats	Number of seats
cars__bodycol__id	Body colour code (e.g. blue, black, silver, ...)
cars__bodycol__met	Metallic paint dummy
cars__intcol__id	Interior colour code (e.g. grey, anthracite, ...)
cars__fuel__id	Fuel type code (e.g. petrol, diesel, electricity, ...)
cars__cylinders	Number of cylinders
cars__ccm	Cubic capacity in cm^3
cars__hp	Horsepower
cars__gear__id	Gear code (manual, automatic)
cars__firstreg__year	Year of first registration
cars__firstreg__mth	Month of first registration
cars__mileage	Mileage in km
cars__price	Advertised price
cars__pricenew	Original price when new
cars__cur	Currency of the displayed prices
cars__warranty	Remaining warranty (number of months)
cars__comments	Textual description of features, damages etc.
cars__accident	Dummy indicating if car has had an accident
cars__equipment	List of applicable equipment codes (see Table 7.3)
cars__created	Date and time of creation of advertisement
cars__changed	Date and time of last update of advertisement
cars__adtype	Advertisement type code
addresses__zip	Postcode of the advertiser
cars__hits	Number of page impressions of the advertisement

Table 7.3: **Equipment dummies contained in the database**

Variable name	Description
Airbag	Driver airbag
Airbag.Beifahrer	Passenger airbag
ABS	Anti-lock brakes
Elektr.Fensterheber	Electric front windows
Elektr.Schiebedach	Electric sunroof
Klima	Air conditioning
Klima.Automatik	Climate control
Ledersitze	Leather seats
Radio.CD	Radio with compact disc player
Radio.Tonband	Radio with compact audio cassette player
Schiebedach	Sunroof
Sitzheizung	Seat heating
Tempomat	Cruise control
Zentralverriegelung	Central locking
X8fach.bereift	Two sets of tyres
Anhängerkupplung	Trailer coupling
Standheizung	Stand heating
Elektr.Parkhilfe	Electronic parking aid
Xenon.Scheinwerfer	Xenon headlights
Sportsitze	Sports seats
Sidebags	Side airbag
Faltdach	Folding roof
Servo	Power steering
Alufelgen	Aluminium wheels
Sperrdiff.	Locking differential
Elektr..Sitze	Electric front seats
Alarmanlage	Anti-theft alarm
Nebelscheinwerfer	Fog headlights
X4x4	Four-wheel drive
Sonnendach	Moonroof
Wegfahrsperre	Engine immobiliser
Hardtop	Hardtop
Navigationssystem	Guidance system
Partikelfilter	Particulate filter

end date was set to the value of `cars_changed`. Then, a new entry with a new revision number was created in `carhist` inheriting the end date from its predecessor, the start date from `cars_changed` and all the car characteristics from the respective entry in `cars`.

Finally, one had to deal with advertisements in `cars` that had disappeared for a moment but reappeared in the latest export file. In this case, a new entry was created in `carhist`. Its revision number was set to one more than the highest revision number already existing for the respective advertisement. Furthermore, the start and end dates of the new entry were determined as for completely new advertisements.

The entire import procedure will be illustrated in the following example. Assume that the table `carhist` contains the values given at the top of Table 7.4 and that the table `cars` given in the middle needs to be imported. Therein, the dates and times are such that $t_0 < t_1 < t_6$ and $t_2 < t_3 < t_7$ (see Fig. 7.2). The four steps described above are now as follows:

1. Handling of disappeared items: Since `car_id` number 3 is no longer available in `cars`, its value of `to_dt` is set to t_{10}, i.e. the time of modification of the file on the FTP server.

2. Handling of new items: The `cars_car_id` number 5 is not yet contained in `carhist`. It is thus added with revision number 0.

3. Handling of changed items: Since $t_6 \in]t_1, \infty[$, the advertisement number 1 has apparently changed. A new entry is thus added to `carhist`.

4. Handling of reappeared items: With $t_7 > t_3$, the advertisement number 2 has reappeared. It is integrated into `carhist` with a new revision number.

At the end, the table `carhist` has the form given at the bottom of Table 7.4. The values printed in bold are those that have been updated or added. Note that the advertisement number 4 contained in `cars` has not been integrated since its modification date t_5 is identical to the one already contained in `carhist`.

Fig. 7.2 shows graphically the five advertisements described in the example above. The thick horizontal lines stand for the advertisements and their different revisions (indicated as small numbers above). The two vertical dashed lines represent two subsequent export times of the database by AutoScout24. The top table in Table 7.4 contains all the information up to t_9, the table at the bottom all the information up to t_{10}.

Table 7.4: **Integration of new advertisements into the database**

`carhist` (before import)

car_id	car_revision	from_dt	to_dt	…
1	0	t_0	t_1	…
1	1	t_1	∞	…
2	0	t_2	t_3	…
3	0	t_4	∞	…
4	0	t_5	∞	…

`cars` (to be imported, date of modification on FTP server: t_{10})

cars_car_id	cars_changed	…
1	t_6	…
2	t_7	…
4	t_5	…
5	t_8	…

`carhist` (after import)

car_id	car_revision	from_dt	to_dt	…
1	0	t_0	t_1	…
1	1	t_1	$\boldsymbol{t_6}$	…
2	0	t_2	t_3	…
3	0	t_4	$\boldsymbol{t_{10}}$	…
4	0	t_5	∞	…
5	**0**	$\boldsymbol{t_8}$	∞	**…**
1	**2**	$\boldsymbol{t_6}$	∞	**…**
2	**1**	$\boldsymbol{t_7}$	∞	**…**

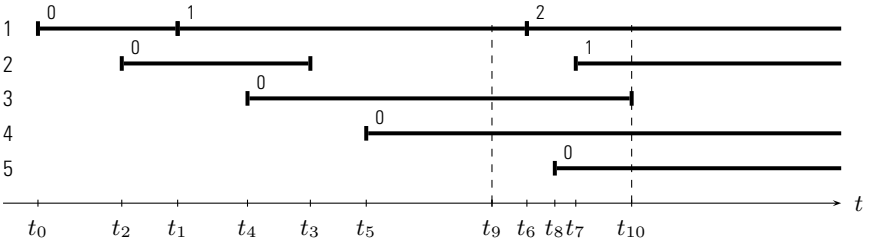


Figure 7.2: **Timeline of the advertisements described in Table 7.4**

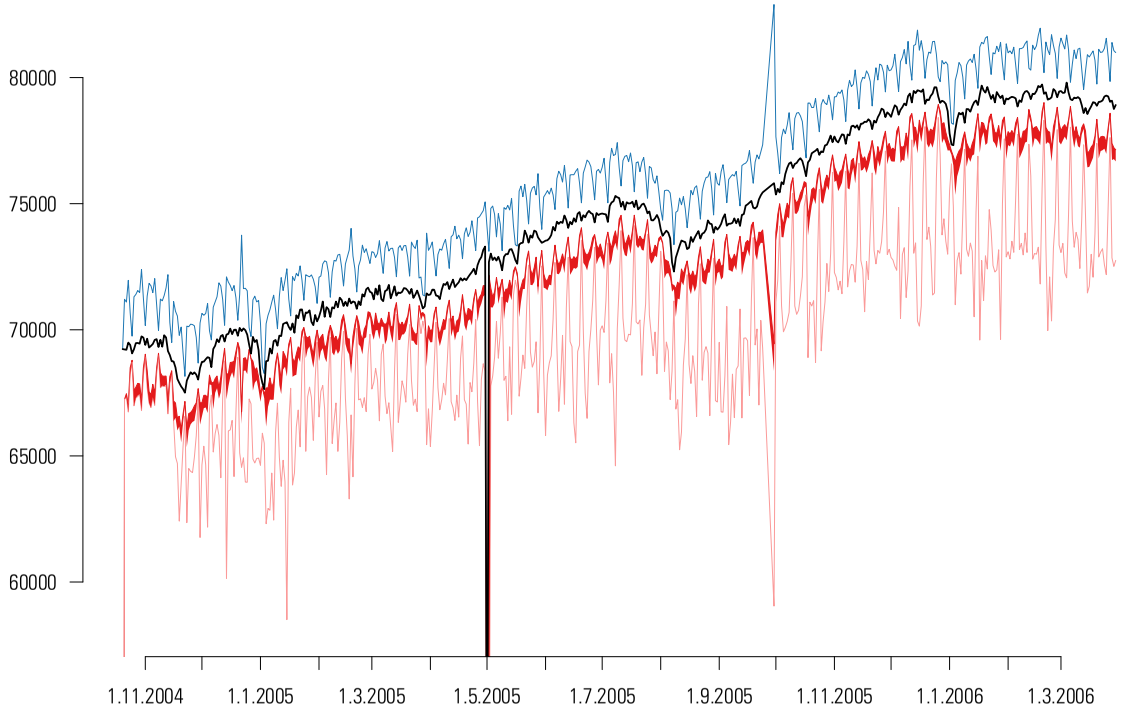


Figure 7.3: **Number of advertisements retrieved from AutoScout24**

The absolute number of advertisements contained in the database is depicted in Fig. 7.3. The black line indicates the number of entries in `cars` as they have been retrieved from AutoScout24 from 20 October 2004 to 31 March 2006. It ranges from 67 506 (22 November 2004) to 79 801 (4 March 2006) advertisements, except for one outlier on 1 May 2005, where only 48 258 entries were transmitted, presumably due to a technical failure. The distance between the black line and the dotted blue line at the top represents the number of advertisements that had disappeared in the file fetched at each specific date compared to the day before. The height of the solid red-coloured area below the black line reflects the number of advertisements that had reappeared whereas the white space above it shows the number of new advertisements. Finally, the vertical distance between the red area and the dotted red line at the bottom represents the number of advertisements that had been changed from the previous to each current date.

It can be seen from Fig. 7.3 that the number of changing, appearing, reappearing, or disappearing advertisements was generally lower on week-ends than on working days. The spike at the end of September 2005 is due to the fact that from 25 to 29 September, no data was fetched from AutoScout24 due to an unnoticed breakdown of `ecosocpc11` caused by a power failure in the city of Fribourg. For this reason, the statistic for 30 September shows more advertisements than usual that had been changed in the meantime.

7.2.2 Preparing the data for further analysis

The structure of the table `carhist` just described is convenient for the extraction of data for a certain point in time. Retrieving all the cars available at any time t can be done by a simple SQL statement selecting all the entries of `carhist` where t is between `from_dt` and `to_dt`.

It is somewhat more difficult to retrieve all the advertisements for a whole period $[t_0, t_1]$ instead of one single time t. The solution adopted here returned all those entries in `carhist` where `from_dt` $\leq t_1$ and `to_dt` $\geq t_0$. If, however, more than one revision of an advertisement satisfied these criteria, only the latest of them was retained. In the example given in Fig. 7.2, a query for all the advertisements of the period $[t_9, t_{10}]$ would thus yield the entries number 1 (revision 2), 2 (revision 1), 3, 4, and 5 (each revision 0). If one is interested in only those advertisements that were updated during the given period, the additional criterion `from_dt` $\geq t_0$ needs to be applied. For the example above, in this case, the query yields only the advertisements number 1 (revision 2), 2 (revision 1) and 5 (revision 0).

Whenever data was extracted from the database for statistical analysis, several other criteria were applied in order to retain only relevant entries. First, only advertisements having `cars_firstreg_year` > 0 were selected in order to allow for a proper specification of the car's age. Secondly, only advertisements where the `price` was given and strictly positive were extracted. A third criterion based on `cars_cartype_id` ensured that only advertisements for used and not for new, demonstration, or old-timer cars were extracted. Finally, a last rule was applied sorting out any entry where the price values were not given in Swiss Francs or where the currency had not been specified. This was the case for about 0.2‰ of the data.

For almost all variables contained in `carhist`, the raw import data were used for any further analysis. There are, however, two exceptions to this rule. First, for every advertisement, a new variable `age` has been calculated yielding the number of months between the year and month of first registration and `from_dt`. Where no month of first registration was given, `age` was defined to be the difference in months between the beginning of the year of first registration and `from_dt`. Secondly, the value 'not available' was introduced instead of any value 0 in the fields `doors`, `seats`, `cylinders`, `ccm`, `hp`, `mileage`, and `pricenew` as well as for all the equipment dummies in the case where none of them was contained in `equipment`.

7.2.3 Infrastructure for statistical analysis

Statistically analysing such huge amounts of data is very demanding both for the software and for the hardware involved. The choice of a software environment for statistical computing fell on R (R Development Core Team, 2006), mainly because of its unrestricted availability and extensibility. This choice turned out to be a very lucky one, as during the whole project, there never was a situation where any idea or wish could not be realised with R. Thanks to the availability of the extension packages `DBI` and `RMySQL` (R Special Interest Group on Databases (R-SIG-DB), 2006; James and DebRoy, 2006), it was straightforward to develop an import function that communicated with the MySQL database. The packages `Rmpi` and `snow` (Yu, 2004; Tierney et al., 2004) provided a comfortable interface to parallel computing, mainly used for the bootstrapping algorithms to be presented later.

On the hardware side, we soon discovered that our desktop PCs (Pentium 4, 1.5 GHz, 768 MB RAM, Windows XP Professional—`ecosocpc11` etc. in Fig. 7.1) were too weak for efficiently estimating either the hedonic functions or the hedonic price indices. The reason for this failure was mainly the limitation of the RAM size. Dealing with up to 70 000 observations and

more than fifty variables (or even more, if categorical variables were transformed to dummies) for the hedonic regressions was not feasible with the indicated memory size. Analogous restrictions applied for the estimation of the hedonic indices based on up to 200 000 reference characteristics vectors. A solution to these hardware requirements could be found by outsourcing all the calculations to a Beowulf class cluster (see `www.beowulf.org`), where R 2.3.1 was installed on four nodes (Dual Opteron, 2 GHz, 3 GB RAM, Linux Fedora Core 2).

7.3 Descriptive statistics and data filtering

In the following paragraphs, we present some descriptive statistics for the advertisements available during the year 2005. For this purpose, all the data for this period were fetched from the database as described above (Section 7.2.2). This yielded 476 274 observations.

Table 7.5 lists the most important quantitative and qualitative variables along with some summary statistics. For the quantitative variables, the minimum, the maximum, the arithmetic mean, the median as well as the first and third quartile (Q1 and Q3, respectively) are given. Moreover, the share of missing values (marked as 'NA', i.e. 'not available') is indicated. For the qualitative variables, the shares of the five most important categories are displayed.

Fig. 7.4 illustrates the relative importance of the dummy variables for equipment, accident, and metallic paint. The larger a black rectangle is, the more advertisements have the value one in the variable concerned. Results are given for the twenty most frequent car models as well as for the whole data set (red bar in the background). The widths of the rectangles are proportional to the frequency of the different car models indicated on the horizontal axis of the figure. The figure shows that characteristics such as, e.g., anti-lock brakes, airbags, engine immobilisers, or power steering were rather common whereas, e.g., folding roofs, stand heatings, or two sets of tyres occurred only very rarely.

A histogram of the price distribution is depicted in Fig. 7.5 on a logarithmic scale. At first sight, it seems that the logarithm of the price variable is almost normally distributed. An adequate normal probability plot of ln(`price`) is given in the left frame of Fig. 7.6. The quantiles of the standard normal distribution are drawn on the horizontal axis while the vertical axis contains the sample quantiles of ln(`price`). It is apparent that there are a number of price values at both tails of the distribution that show a

Table 7.5: **Summary statistics of selected variables in the 2005 data**

	price	pricenew	age	mileage	doors	seats
Min.	1	1	0	1	1	1
Q1	9800	33000	31	32700	3	5
Median	16900	45660	56	68000	5	5
Mean	23180	58360	69.34	77450	4.115	4.785
Q3	27800	69190	94	110000	5	5
Max	10000000	1000000	1198	1000000	9	56
NA	0.00 %	75.27 %	0.00 %	0.11 %	3.05 %	6.13 %

	cylinders	ccm	hp
Min.	1	1	1
Q1	4	1762	110
Median	4	1997	140
Mean	4.803	2266	156
Q3	6	2598	192
Max	90	60010	999
NA	4.11 %	2.40 %	3.37 %

make		body		bodycol	
VW	11.82 %	Limousine	46.35 %	blau	20.85 %
OPEL	8.96 %	Kombi	21.39 %	schwarz	18.98 %
BMW	8.20 %	Geländewagen	7.73 %	silber	18.90 %
AUDI	8.17 %	Cabrio	7.27 %	grau	10.55 %
MERCEDES-BENZ	7.55 %	Minivan	6.78 %	grün	8.48 %
(Other)	55.30 %	(Other)	9.57 %	(Other)	22.00 %
NA	0.00 %	NA	0.90 %	NA	0.24 %

gear		fuel		intcol	
Handschaltung	65.15 %	Benzin	86.13 %	schwarz	20.37 %
Automat	33.81 %	Diesel	12.92 %	grau	17.57 %
		Elektro	0.01 %	anthrazit	12.11 %
		Erdgas	0.00 %	beige	6.98 %
		Erdgas/Benzin	0.02 %	blau	3.28 %
				(Other)	2.67 %
NA	1.04 %	NA	0.93 %	NA	37.02 %

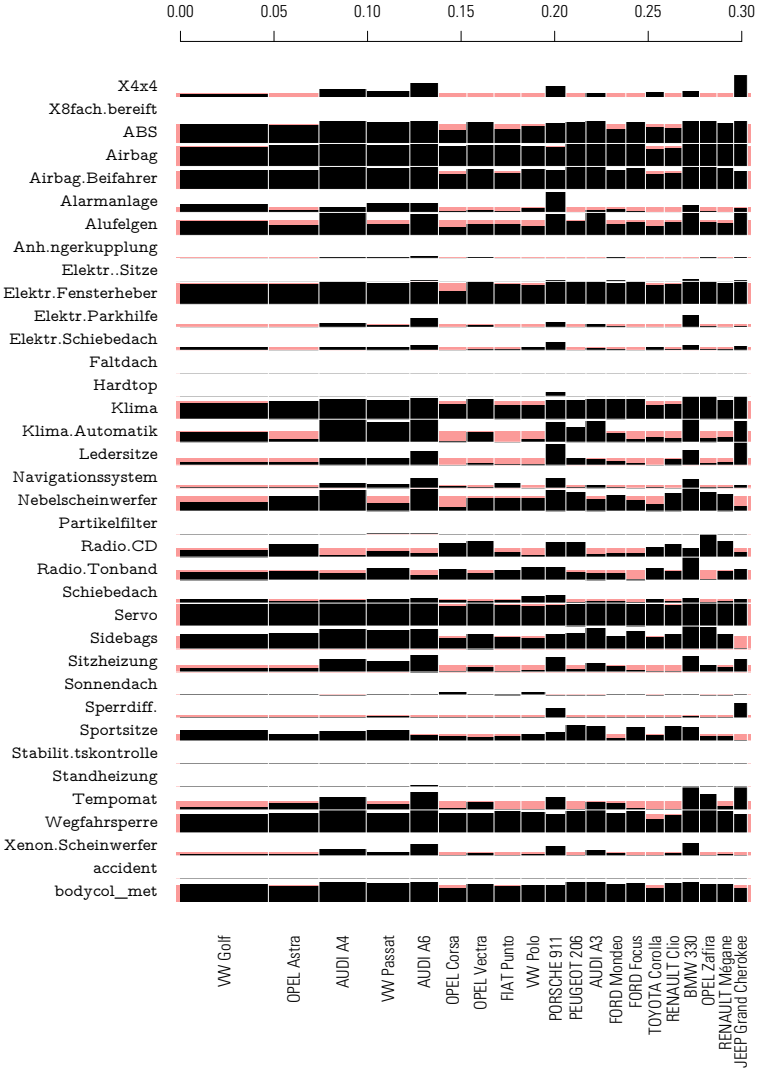


Figure 7.4: **Relative importance of the equipment dummies for the twenty most common car models in the database**

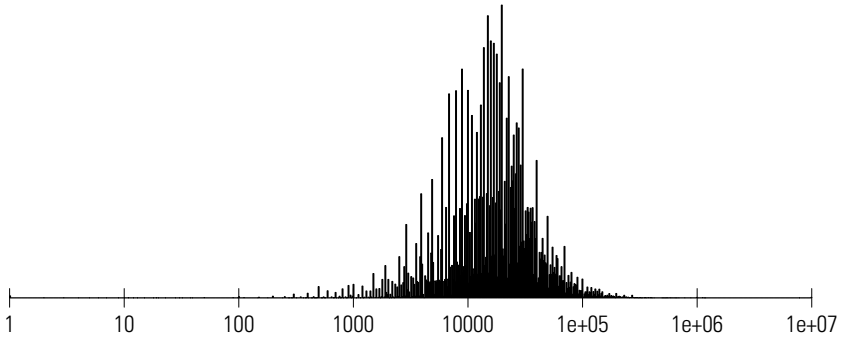


Figure 7.5: **Histogram of the observed prices (in CHF)**

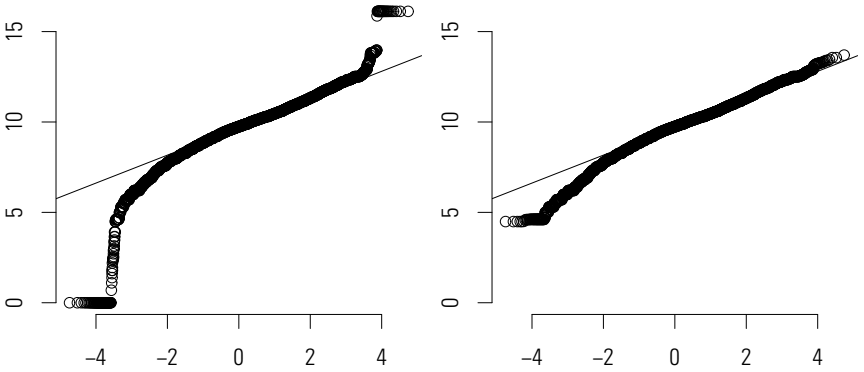


Figure 7.6: **Normal probability plot of the natural logarithm of the price for the original and the truncated data set**

special behaviour. Based on a more detailed analysis, we recognised that most of these advertisements embodied erroneous information. Even though the advertising rules of AutoScout24 clearly request that the quoted prices must be end-consumer prices containing the value-added tax as well as any other additional charges, it could be observed that virtual prices such as 1 or 999 999 CHF were entered by vendors several times. This behaviour was often accompanied by a comment (in `cars_comments`) saying that the vendor was waiting for bids by phone or that the advertised car was not offered but sought by the vendor. Since AutoScout24 did not regularly check the compliance of the advertisements entered by their customers with the rules, there are several such entries in the database.

For any further analysis, therefore, all the advertisements that did not have a price higher than 52 and lower than 999 999 CHF were filtered out. The lower bound of 52 CHF, on one hand, was chosen to be the same as the minimal price for an advertisement on the AutoScout24 platform. The idea upon this choice is based is that the vendor would probably want to get at least the price of the advertisement back when he sells the car. On the other hand, eliminating all the cars with a price higher than a million did not seem to be too restrictive since the meaningful advertisements falling in this range were very sparse. An *ad hoc* analysis for the advertisements being excluded like that (0.04 % of the 2005 data) showed that the large majority of them really contained obviously inappropriate information. A normal probability plot of the truncated data is shown in the right frame of Fig. 7.6 showing that the truncation leads to a distribution of ln(`price`) that is closer to normality.

The truncation of the allowed price range as described above is admittedly somewhat unsatisfactory. On one hand, there *are* advertisements having prices lower than 52 or higher than 999 999 CHF which do not contain erroneous information. On the other hand, and this might be even more important, errors and unwanted entries are also contained among the large majority of advertisements that fall into the admitted price range. In a production environment, e.g. when an official CPI should be estimated using these data, more care would be needed in order to 'clean' the database. With appropriate data mining techniques, one would probably be able to find and filter out more effectively the unwanted advertisements. One could also imagine applying such techniques for identifying candidates for a subsequent manual analysis.

Finally, the geographical distribution of the advertised cars is displayed in Fig. 7.7. This map is based on the postcodes of the dealers as they are available in the database. A circle is drawn for every postcode having its

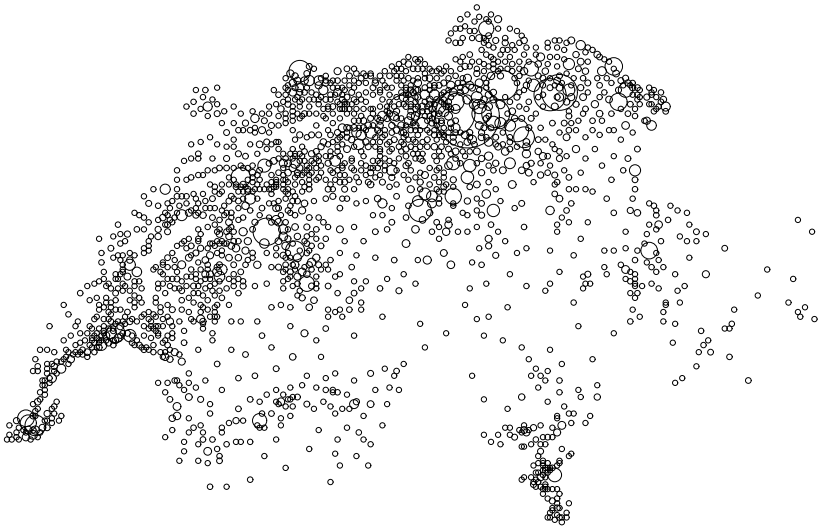


Figure 7.7: **Geographical distribution and relative number of the advertisements**

center approximatively at the place where the corresponding town or village is situated. The diameter of each circle is a positively monotonic function of the number of advertisements having the corresponding postcode.

The picture shows that the geographic structure of Switzerland seems to be astonishingly well covered by the data set. Moreover, the observed advertisements seem to correspond closely to the density of the population. More detailed analyses, however, would be needed to confirm these hypotheses. An interesting question for further research might be to compare differences and changes in prices of used cars for the different regions of Switzerland.

7.4 Conceptual limitations of the data set

At the beginning of the current chapter, we noted that the data provided by AutoScout24 seemed to be particulary suitable for a hedonic analysis. After having described the nature of these data, the question whether this is really true needs now to be reconsidered.

The advantages of this data set are obvious: It contains a large number of standardised characteristics which can be used directly for estimating

hedonic functions. Moreover, it was discussed that quality-adjusted price indices for cars in general or particularly for used cars have often been estimated with hedonic techniques in the literature and increasingly also in the practice of statistical offices.

There are, however, at least two downsides that have to be mentioned with regard to the AutoScout24 data. One of them is the quality of the data in general and the second is related to the validity of the price values in particular.

It has been discussed in the previous section that a number of advertisements containing obviously erroneous information have been detected. While the share of such 'dirty' data points seems to be particularly high at both tails of the price distribution, it is unknown how precise and adequate the data in general are. While the advertisements stemming from large institutional dealers probably are quite accurate, the same is not true for those entered by private individuals. Since the validity of the data entered cannot be checked for each individual car, errors are to be expected. This is particularly true for the equipment dummies, where no distinction is currently made for items that are not present and others where one does not know whether they are present or not. In other words, a value of zero cannot be distinguished from 'not available' for the equipment dummies. It was described above that these variables were all attributed the value 'NA' whenever all of them contained zeros, but this might be inappropriate. On the other hand, it seems plausible that some individuals do not know whether their car possesses one or the other characteristic or not. So they should have the possibility to indicate their uncertainty.

As a general rule, it might be an interesting question for further research whether some kind of plausibility rating for each vendor (or category of vendors) could be introduced in order to downweigh advertisements that are less plausible in favour of those that stem from a more credible source. This could be done by AutoScout24 directly by introducing some sort of vendor rating mechanism. This would give the purchaser of a car the possibility to indicate to what degree the information contained in the advertisement was appropriate, and this information could then be displayed on the web site whenever the same vendor sells another car afterwards. Similar rating systems can currently be found in online auctions for instance.

The second important downside is related to the fact that the prices indicated with the advertisements, assuming that they are correct, are offering prices instead of transaction prices. This is a problem for estimating price indices because for these, methodically, one is interested in the prices actually paid by the consumers along with the precise quality they buy for a

certain price rather than in the price labels being attached by the vendors. The problem is similar to the use of list prices for new cars as discussed, for instance, by GORDON (1990, pp. 355 ff.). There, it is explained how age-adjusted prices of used cars can be used as a proxy for the transaction prices of new cars. This assumes, however, that the prices available for used cars are real transaction prices or are very close to them.

There is evidence for the assumption that the prices published on platforms such as the one provided by AutoScout24 are upper limits for the prices effectively paid. The probable existence of consumers bargaining with vendors for a lower price or for additional equipment (such as an additional set of tyres, for instance) should not be neglected. Another issue is that vendors might publish prices which are higher than the market is willing to pay. As a result, the respective car would be advertised on the platform but not sold for the indicated price.

There are at least two approaches that could be suitable for tackling this last issue. The first consists in looking exclusively at cars that are advertised for a comparatively short time period only. Therebehind lies the idea that cars offered at or below the market price would be sold quickly whereas the others would stay on the platform until the market price changes. Another quite similar approach would be to observe how the advertised price of a car evolves over time until a purchaser is found. An economically plausible assumption would probably be that the last price observed is nearest to the market and to the realised transaction price.

The second, probably less promising approach is to produce an estimate of the discounts vendors are giving to their customers. Trying to ask the vendors to indicate the real transaction price when they unpublish an advertisement seems to be a difficult task. Many of them would probably want to keep such information confidential. Asking the same question to the purchasers would even be more difficult because they are and cannot be known in general. A feasible alternative might be to try and estimate an overall discount share, possibly subdivided into different car types and price classes. Assuming that these shares remain constant over time, such an overall discount factor, however, would not change most of the hedonic price index estimates, at least not those satisfying the commensurability axiom (Test 9).

This last remark provides evidence for the assumption that discarding such discounts and working with the advertised prices might not influence the resulting index estimates dramatically. There is a need, however, for testing this hypothesis empirically before such data is used in a production environment. In our project, the main concern is to study methods for in-

dex estimation and not to interpret the resulting index estimates. It seems thus admissible to ignore this concern and to take the data as if they contained transaction prices. Doing so might change the estimates but not the methodology applied.

CHAPTER 8

Estimating the Hedonic Function for Used Cars

This chapter is going to present the estimation procedure for the hedonic functions with different functional forms or model types. For the current analysis, hedonic functions have been estimated on a monthly basis. In the sequel, February 2005 on some occasions will be taken as an example to illustrate the estimation process. The explanations, however, are applicable to any other month as well.

8.1 Linear regression

8.1.1 Simple semi-log model (SSL)

The first model type to be presented is the standard version of a semi-log model as it has been discussed in Section 5.3.1. In a first step, we will now present the final algorithm developed for estimating this type of hedonic function. Several econometric questions are deliberately left open at this stage. This will allow us to describe the algorithm in a compact manner. The mentioned questions, however, are afterwards reconsidered in the critical comments that follow in Section 8.1.2.

The algorithm starts with fetching the car advertisements that had been created or updated during, e.g., February 2005 from the database as described in Section 7.2.2. This yields 64 951 advertisements as displayed in the first column of Table 8.1. The analysis is restricted to new and updated advertisements in order to work with up-to-date data; it has been discussed

Table 8.1: **Sample sizes considered for the hedonic function estimation**

Month	Original data	SSL	ESL	SDL	FDL	SPLS	EPLS	SSL/m
Oct 2004	37356	66.8%	66.9%	66.3%	66.6%	68.4%	67.2%	67.0%
Nov 2004	57647	77.0%	77.1%	76.7%	76.7%	78.7%	77.9%	77.3%
Dec 2004	68743	81.2%	81.3%	80.9%	80.9%	83.2%	82.5%	81.6%
Jan 2005	66138	82.2%	82.3%	81.9%	81.9%	84.1%	83.3%	82.5%
Feb 2005	64951	84.0%	84.1%	83.8%	83.8%	85.8%	84.8%	84.3%
Mar 2005	64716	85.2%	85.4%	84.9%	84.9%	87.1%	86.2%	85.6%
Apr 2005	66221	84.2%	84.3%	83.9%	83.9%	86.0%	85.0%	84.5%
May 2005	77978	83.9%	84.0%	83.7%	83.6%	85.6%	84.8%	84.3%
Jun 2005	68785	83.6%	83.8%	83.3%	83.3%	85.3%	84.4%	84.0%
Jul 2005	66545	84.5%	84.6%	84.0%	84.2%	86.3%	85.5%	84.9%
Aug 2005	70692	84.6%	84.7%	84.2%	84.3%	86.4%	85.5%	84.9%
Sep 2005	70509	85.2%	85.3%	84.9%	84.9%	87.2%	86.4%	85.6%
Oct 2005	68310	85.2%	85.4%	84.8%	84.9%	87.3%	86.5%	85.7%
Nov 2005	67853	84.8%	84.9%	84.4%	84.5%	86.8%	85.9%	85.2%
Dec 2005	68830	85.6%	85.8%	85.2%	85.3%	87.7%	87.0%	86.1%
Jan 2006	72596	85.5%	85.7%	85.1%	85.3%	87.6%	86.7%	85.9%
Feb 2006	69712	85.2%	85.3%	84.9%	84.9%	87.3%	86.2%	85.6%
Mar 2006	74408	83.4%	83.5%	83.0%	83.1%	85.3%	84.2%	83.8%

that advertisements remaining on the platform for a long time might indicate that the published price is higher than the consumers are willing to pay.

The second step of the algorithm consists of defining a maximal and a minimal regression model for the data. We are going to explain below what roles these two models play. The maximal formula for the simple semi-log model is given by

```
log(price) ~ make + body + bodycol + bodycol_met + fuel + gear +
 doors + seats + cylinders + ccm + hp + age * mileage + warranty +
 accident + X4x4 + X8fach.bereift + ABS + Airbag + Airbag.Beifahrer +
 Alarmanlage + Alufelgen + Anh.ngerkupplung + Elektr..Sitze +
 Elektr.Fensterheber + Elektr.Parkhilfe + Elektr.Schiebedach +
 Faltdach + Hardtop + Klima + Klima.Automatik + Ledersitze +
 Navigationssystem + Nebelscheinwerfer + Partikelfilter + Radio.CD +
 Radio.Tonband + Schiebedach + Servo + Sidebags + Sitzheizung +
 Sonnendach + Sperrdiff. + Sportsitze + Standheizung + Tempomat +
 Wegfahrsperre + Xenon.Scheinwerfer + Zentralverriegelung
```

while the minimal formula just contains

The notation of these formulae corresponds to their standard representation in the S language (see CHAMBERS, 1998, p. 166 or CHAMBERS and HASTIE, 1993, Chap. 2). The symbol '~' is a special operator which stands for 'is fitted to' or 'is modelled as' while '+' means inclusion of the corresponding variable. The '*' operator denotes factor crossing, i.e. 'age * mileage' expands to 'age + mileage + age:mileage' where 'age:mileage' is interpreted as the interaction of 'age' and 'mileage'. The response term 'log(price)' stands for the natural logarithm of the price variable.

When a linear model is fit using such a formula, the information contained therein is used to build an appropriate design matrix $\mathbf{X}$. Unless the constant factor '1' is removed explicitly by adding '- 1' to the formula, the first entry in every row of the design matrix will contain the value 1 corresponding to a constant regressor term. Quantitative variables are included as single columns whereas categorical variables with L levels are split into $L - 1$ columns containing only the values 0 and 1 where appropriate.

Once the formulae are defined, the simple semi-log model algorithm continues with removing from the original data set all the observations that contain missing values in at least one of the variables included in the maximal model formula. Then a linear model is fit to the remaining data using the maximal model formula as described above. In order to detect outliers or, more precisely, influential observations, DFFITS statistics are then calculated based on the model residuals for every observation contained in the data set (see BELSLEY et al., 1980; VELLEMAN and WELSCH, 1981; BRACHINGER, 1990; HARRELL, 2001). Observations having

$$\text{DFFITS}_i > 2\sqrt{\frac{K}{N^t}}$$

with K and N^t being the number of columns and rows of the design matrix, respectively, are marked as influential and deleted from the data set. This criterion corresponds to the recommendations given in the mentioned literature. The share of data points left over after the exclusion of incomplete and influential observations is 84.0% for February 2005. Similar values for all the other months are displayed in Table 8.1.

The next and final step consists of performing a step-wise model selection based on the Akaike information criterion (AIC). This is done with the function stepAIC in the R package MASS (see VENABLES and RIPLEY, 2002, p. 175 ff.). Starting from a linear model fitted using the maximal formula, the algorithm searches iteratively for the subset of predictor variables that

minimises AIC by removing or adding terms on the right-hand side of the model equation as needed. The range of models searched consists of all those having at least the terms given in the minimal formula and at most the terms in the maximal formula.

Table A.1 in the appendix displays the terms that remained in the final model for the months considered. The adjusted R^2 statistics of these simple semi-log models are all located between 0.9436 and 0.9488. A plot of the observed (horizontal axis) versus predicted (vertical axis) price values for this model for February 2005 is displayed in the top-left frame of Fig. 8.1 on page 113. There, the data points used for the estimation of the final model are represented by black circles while the coloured crosses show the observations that were tagged as outliers and not used for the estimation of the model.

8.1.2 Critical comments The algorithm just displayed provides a comparatively simple approach for estimating a hedonic function. A semi-log model is fit to the data where outliers have been removed. The variables to be included are chosen automatically according to the AIC criterion, and the accuracy of the model fit measured by the adjusted R^2 values is respectable. Several important issues, however, are neglected in this procedure. Some of them shall now be commented.

One critical point is that the estimation does not use the whole information contained in the data. Several variables, such as `pricenew`, `intcol`, and `model`, as well as the postcode of the dealer have not been considered. The reasons for adopting this practice were diverse for the individual variables. `pricenew` and `intcol` were discarded because they were missing for from about one third up to three quarters of the observations (see the 'NA' shares displayed in Table 7.5 for the 2005 data). Since all the observations containing missing values for the relevant variables were thrown out of the data set, the loss of information by doing so while keeping these two variables in the model would have been important. Since `pricenew`, however, turned out to be an important predictor when analysed for the data where it was available, a special treatment of this variable led to the enhanced semi-log model that is going to be displayed in Section 8.1.3.

The reason for excluding the `model` variable was its huge number of different categories (1270 for the 2005 data). Including it would have meant adding almost as many (exactly one less) additional columns to the design matrix leading to an important increase in computing time and memory size requirement for performing estimations and predictions. We are going to present a technique of how to include the model variable in a simple

semi-log model in Section 8.3. Taking into account the information on the geographic location of the vendors, finally, would necessitate a more complex spatial price model. Including just cartesian coordinates as predictors would probably not be meaningful enough in such an essentially linear hedonic regression model. Building a more sophisticated spatial model, however, would be an extensive research question of its own.

Strongly connected with the choice of the analysed variables is the second piece of criticism that is directed towards the handling of missing values in the data. It has just been mentioned that several variables were excluded because they contained too many missing values. This approach seems to be adoptable since the number of complete observations is large. On the other hand, one could also argue that some of these observations might not be missing at random but due to the way the data are generated. One could imagine, for instance, that an institutional dealer does not report one or the other variable simply because he does not survey it. This would lead to the exclusion of all these observations even though they might contain valuable information for the other predictors. Throwing away all the incomplete observations is thus just a first approximation. Further research would be needed to analyse imputation mechanisms for missing values.

A third critical point is related to the model specification. Except for the term '`age:mileage`', the models taken into consideration here do not contain any second-order or even higher-order interactions between the variables. It seems plausible, however, that some characteristics or bundles of characteristics might be correlated with each other or with, e.g., the make of the car. Neglecting such dependencies in the model building process might thus be inappropriate. On the other hand, introducing second-order terms for these more than fifty variables would, again, increase dramatically the number of covariates. More sophisticated techniques are thus needed to reduce the number of dimensions while taking into account the existing interactions—PLS regression is going to be one such candidate.

Finally, an important weakness of this simple semi-log approach is the way it handles influential observations. Even though there is theoretical evidence for the applied criterion based on the DFFITS estimations, especially if one is interested in appropriate predictions of the endogenous variable (see Brachinger, 1990, p. 209), it seems unsatisfactory, as it does not distinguish erroneous from somewhat aberrant but still meaningful observations. Due to this fact, the goodness of fit of the model presented here tends to be over-estimated since many 'good' data points are just thrown away. Other techniques, e.g. those stemming from the data mining literature, which are able to discover real outliers, might therefore be preferred. It seems im-

portant that any hedonic function estimator is robust against outliers to a certain extent, especially in our case where the data do contain erroneous information. The reader is referred to DIEWERT (2003b, pp. 28–9) for a short comment on whether or not to remove influential observations is appropriate in hedonic regressions in general.

Based on the criticism just presented, we are now going to study several alternatives to this simple semi-log model in order to confront some of these imperfections. First, an enhanced semi-log model will be presented where the `pricenew` variable is included. Then, we are going to study first a double-log model and then a mixture of both a semi-log and double-log model. Section 8.2 is going to present two models which try to deal with interactions between the exogenous variables, while the last approach to be presented in Section 8.3 starts from the idea of estimating an individual regression model for every car model separately.

8.1.3 Enhanced semi-log model (ESL)

It has been noted in the previous subsection that the original price of a car—represented by the variable `pricenew`—might have an important explanatory power on the price of a used car. This is due to the fact that the original price of a new car reflects its quality as it is perceived by the consumers. People are willing to pay more for a car that offers, in a certain sense, a higher quality.

If one observes a car over time, there are only a few characteristics that change while most of them—at least those representing the equipment of the car—remain constant despite the usage. In the present context, those that do or may change are `age`, `mileage`, `warranty`, and `accident`. One could thus imagine that the price of a used car could well be explained by these four variables combined with `pricenew`. In other words, one could expect that `price` depends on the time-independent equipment variables only through `pricenew`.

Since the ageing properties of a car, however, are different for different makes or models, the variable `pricenew` alone might not be sufficient for replacing the whole set of equipment variables. Nevertheless, it is interesting to investigate whether the original price is able to replace at least some of them. For this purpose, the simple semi-log model will now be extended to incorporate `pricenew` whenever it is available.

The enhanced semi-log hedonic function estimator starts with exactly the same procedure as the simple semi-log algorithm for estimating a hedonic function without taking into account the `pricenew` variable. In an additional step, however, this same procedure is repeated for the sub-population of the

data where the original price is available. In this second step, the term 'log(pricenew)' is added to the right-hand side of both the maximal and minimal formulae. Adding the natural logarithm of the original price and not the original price itself contradicts in a sense the philosophy of the semi-log compared to the double-log models. Since the current price, however, is incorporated through the logarithm, it would be incoherent not to do the same with the original price.

The hedonic function generated by this algorithm contains thus information on two fitted semi-log models. Every time it is to be applied on a characteristics vector in order to predict a price value, it determines first whether the variable pricenew is contained in the vector. If this is the case, the price is predicted using the model with the pricenew variable, if not, it is predicted using the simple semi-log model.

Table A.2 displays the terms that remained in the final models for the months considered. The bullets to the left of the vertical bars stand for variables that were included in the models without pricenew. These are the same as those contained in Table A.1. The bullets to the right of the vertical bars refer to the models where log(pricenew) is included. It turns out that less equipment variables remain in the models with pricenew than in those without it, which is what had been expected. The adjusted R^2 statistics of the models with pricenew are located between 0.953 and 0.9627 whereas those of the models without pricenew correspond to the values given in Section 8.1.1.

A plot of the observed versus predicted price values for this enhanced semi-log model for February 2005 is displayed in the top-right frame of Fig. 8.1.

8.1.4 Simple double-log model (SDL)

The third family of models to be implemented are based on the double-log approach (see Section 5.3.1) which is, according to Triplett (2004, p. 180), the 'overwhelming favourite' for IT equipment studies. There, all the metric variables (price, doors, seats, cylinders, ccm, hp, age, and mileage) are introduced into the formulae of the simple semi-log model through natural logarithms. Here, the variable pricenew is, again, disregarded.

The terms that remained in the final models are displayed in Table A.3. Here, the adjusted R^2 statistics range from 0.9289 to 0.9369. They are thus lower than in the simple semi-log model.

8.1.5 Flexible double-log model (FDL)

In order to let the data speak on whether or not to include a metric variable in its original form or in its form transformed by the natural logarithm (or both), the following 'flexible' double-log model algorithm specifies the maximal formula as

```
log(price) ~ make + body + bodycol + bodycol_met + fuel + gear +
 log(doors) + log(seats) + log(cylinders) + log(ccm) + log(hp) +
 log(age) * log(mileage) + doors + seats + cylinders + ccm + hp +
 age * mileage + warranty + accident + X4x4 + X8fach.bereift + ABS +
 Airbag + Airbag.Beifahrer + Alarmanlage + Alufelgen +
 Anh.ngerkupplung + Elektr..Sitze + Elektr.Fensterheber +
 Elektr.Parkhilfe + Elektr.Schiebedach + Faltdach + Hardtop + Klima +
 Klima.Automatik + Ledersitze + Navigationssystem +
 Nebelscheinwerfer + Partikelfilter + Radio.CD + Radio.Tonband +
 Schiebedach + Servo + Sidebags + Sitzheizung + Sonnendach +
 Sperrdiff. + Sportsitze + Standheizung + Tempomat + Wegfahrsperre +
 Xenon.Scheinwerfer + Zentralverriegelung
```

and the minimal formula as

```
log(price) ~ log(age) + log(mileage)
```

The effect of these settings is that, during AIC minimisation, both the logarithmic and the linear terms are taken into account while requiring at least `log(age)` and `log(mileage)` to be present.

The terms that remained in the final models are displayed in Table A.4. Interestingly, almost all the metric variables are present in the final model both in their original and in their logarithmic form. This might indicate that neither the pure semi-log nor the pure double-log models are good enough for fitting the dependence structure in the data. Here, the adjusted R^2 statistics range from 0.9451 to 0.9505, which is slightly higher than for both the simple semi-log and the simple double-log model. A more sophisticated comparison of these models, however, will be discussed in Section 8.4.

8.2 PLS regression

8.2.1 Simple PLS model (SPLS)

A completely different family of models are those based on the partial least squares algorithm (see Section 5.3.2). For the implementation within the present project, two different settings have been used: a 'simple' one where the number A of latent variables is fixed in advance and an

'enhanced' one where A is determined by ten-fold adjusted cross-validation on each individual data set.

In contrast to the semi-log and double-log models presented so far, it was possible here to introduce the `model` variable as a regressor into the model formula. Moreover, the approach presented in the enhanced semi-log algorithm of estimating two different models for the subsets of the data with and without taking into account the variable `pricenew` has been followed for these two PLS models by default. The PLS model formula is thus given by

```
log(price) ~ model + body + bodycol + bodycol_met + fuel + gear +
 doors + seats + cylinders + ccm + hp + age * mileage + warranty +
 accident + X4x4 + X8fach.bereift + ABS + Airbag + Airbag.Beifahrer +
 Alarmanlage + Alufelgen + Anh.ngerkupplung + Elektr..Sitze +
 Elektr.Fensterheber + Elektr.Parkhilfe + Elektr.Schiebedach +
 Faltdach + Hardtop + Klima + Klima.Automatik + Ledersitze +
 Navigationssystem + Nebelscheinwerfer + Partikelfilter + Radio.CD +
 Radio.Tonband + Schiebedach + Servo + Sidebags + Sitzheizung +
 Sonnendach + Sperrdiff. + Sportsitze + Standheizung + Tempomat +
 Wegfahrsperre + Xenon.Scheinwerfer + Zentralverriegelung
```

in the case without `pricenew`. In the other case, `log(pricenew)` has been added to the right-hand side of this expression. Depending on the number of different `model` categories that were present in each month's data, the design matrices $\mathbf{X}$ consisted of up to more than 1300 columns.

In order to detect and remove outliers, a mechanism analogous to the one for the semi-log and double-log models has been implemented. It relies on the *leverage* of each individual observation i which, in a PLS regression framework, can be defined as

$$h_i = \frac{1}{N^t} + \hat{\boldsymbol{t}}_i'(\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\boldsymbol{t}}_i$$

where $\hat{\mathbf{T}}$ is the factor score matrix as it has been described in Section 5.3.2 and $\hat{\boldsymbol{t}}_i'$ is the i-th row of $\hat{\mathbf{T}}$. The details are to be found in Martens and Næs (1989, pp. 276–7). Following the advice presented there, an observation i has been tagged as influential if

$$h_i > \frac{2(1+A)}{N^t}\,.$$

This was done after having estimated a first PLS model with $A = 50$ latent variables. Then, the same model was fit again using only the non-influential data. This last model would then be used by the resulting hedonic function for its price predictions.

Table 8.2: **Preferred number A_{opt} of latent variables determined by cross-validation for the different enhanced PLS models**

	Oct 2004	Nov 2004	Dec 2004	Jan 2005	Feb 2005	Mar 2005
w. pricenew	31	42	87	90	57	96
w/o pricenew	80	74	81	77	82	95

	Apr 2005	May 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005
w. pricenew	100	93	100	95	100	100
w/o pricenew	83	95	91	83	98	80

	Oct 2005	Nov 2005	Dec 2005	Jan 2006	Feb 2006	Mar 2006
w. pricenew	100	96	99	100	100	100
w/o pricenew	85	79	92	81	100	83

The PLS model estimations in R were calculated using the package `pls` (Wehrens and Mevik, 2006) and the algorithm `kernelpls.fit` therein. This algorithm was developed by Dayal and MacGregor (1997) and is particularly efficient if the number of observations exceeds largely the number of variables, which is true in the present context.

8.2.2 Enhanced PLS model (EPLS) The enhanced PLS model deviates from the simple one in the way the number A of latent variables is determined. For the former, A is fixed to 50 in advance whereas, for the latter, A is chosen between 1 and 100 such that the mean squared error of prediction (MSEP) of the resulting PLS regression model is minimal. The MSEP is estimated by ten-fold adjusted cross-validation on the learning data set (see Mevik and Cederkvist, 2004). This is done for both the model without and with the variable `pricenew` individually yielding for each of them a preferred number A_{opt} of latent variables to include. Table 8.2 shows the realised values of A_{opt} for the different EPLS models. Since many of them are close or even equal to 100, it might have been judicious to allow an even larger number of latent variables to be chosen. Such a practice, however, would have increased further the memory requirements and the calculation time. The predictions in the resulting hedonic function were thus based on the preferred number of components displayed in Table 8.2.

Whether or not either the simple or the enhanced PLS models perform

better than the semi-log and double-log models presented above is going to be discussed in Section 8.4. It is apparent, however, that several fundamental issues inherent in the estimation of a hedonic function for used cars could not be solved just by switching to this other regression technique. The most prominent of them is outlier detection since the mechanism for doing so is just as rudimentary here as it was for the linear regression models. A potentially interesting approach of a robust method of estimating PLS models based on the BACON algorithm (see Billor et al., 2000) has been developed by Kondylis and Hadi (2005).

8.3 Per-model simple semi-log model (SSL/m)

With the enhanced semi-log and the two PLS models, we already presented an approach where different models were fit to two subsets of the original population: one model, excluding the `pricenew` variable, on the whole population and the other, including the `pricenew` variable, on the part of the population where this information was given. The idea of fitting different models within a family of models to different subsets of the training data is now going to be developed one step further.

One major drawback of the simple semi-log approach presented in Section 8.1.1 is its disability of integrating the car model variable into the regression model. It seems realistic to assume that there are characteristics variables which interact with the car model, since some models may have accessories as a standard fitting that others do not. Accessories may thus not be price-relevant for all the models of a certain make. For this reason, one could imagine fitting a semi-logarithmic regression model of the form given above not to the data set as a whole, but to each car model individually. Interactions between the car model and the other characteristics variables are, in this manner, implicitly taken care of. When a prediction is to be made for a given characteristics vector, this is done using the estimated parameter vector for the corresponding car model.

An important inconvenience related to this pure 'per-model' simple semi-log approach—as it was implemented in Beer (2007), for instance—is that the observations for the individual months are not numerous enough to admit of fitting a regression model for every individual car model. It only makes sense to carry out a regression in a specific month for car models where at least about one hundred observations are available. This is twice as much as there are exogenous variables. If regression models were only fitted for these most frequent car models, however, a certain part of the

Table 8.3: **Indicators of the per-model simple semi-log hedonic function estimates**

Month	Model count	Model share	Data share	Overall adj. R^2	Partial adj. R^2
Oct 2004	69	10.6%	59.1%	0.958	0.967
Nov 2004	117	16.9%	73.6%	0.961	0.967
Dec 2004	133	18.6%	76.7%	0.962	0.967
Jan 2005	132	18.2%	76.1%	0.961	0.967
Feb 2005	135	18.9%	76.0%	0.963	0.969
Mar 2005	135	18.4%	76.0%	0.962	0.969
Apr 2005	127	17.3%	75.1%	0.962	0.968
May 2005	151	20.1%	78.1%	0.961	0.966
Jun 2005	136	18.4%	76.5%	0.961	0.967
Jul 2005	135	18.2%	76.3%	0.961	0.967
Aug 2005	139	18.6%	76.7%	0.960	0.966
Sep 2005	137	18.3%	75.9%	0.960	0.965
Oct 2005	139	18.3%	76.2%	0.961	0.967
Nov 2005	135	18.1%	75.1%	0.961	0.966
Dec 2005	137	18.2%	76.1%	0.960	0.966
Jan 2006	144	19.1%	76.8%	0.960	0.965
Feb 2006	141	18.9%	76.2%	0.961	0.967
Mar 2006	146	19.2%	76.9%	0.959	0.964

training data would be neglected completely. Moreover, predictions would not be feasible for every complete characteristics vector in the training data but only for the car models where a regression was carried out.

In order to overcome this weakness, we implemented here the idea of predicting prices of rare car models, i.e. where less than one hundred observations were available in the training data, using a simple semi-log model fitted to the data set containing all of the advertisements in the current time period. The resulting *per-model simple semi-log approach* is therefore an amended version of the original simple semi-log model yielding more precise price predictions for the most frequent car models. The formulae of the regression models for the individual car models are almost the same as those of the simple semi-log model. The only difference is that in each analysed subset of the data, variables showing no variance (e.g. `make`) are removed from the formula. Outlier detection and variable selection are performed for every car model individually.

Table 8.3 presents some summary statistics for the hedonic functions estimated for the different months under consideration. The column entitled 'Model count' gives the number of car models for which individual regres-

sions were performed. 'Model share' is 'Model count' divided by the total number of different car models available in the respective learning data set. 'Data share' represents the share of the observations (after outlier deletion) that belong to these car models receiving special attention.

Calculating adjusted R^2 statistics for this per-model approach requires special care, as the number of included covariates varies between the different car models. Starting from the unadjusted R^2 values, we chose to do the adjustment as if all models had the maximal number of covariates over all regressions performed in a month. This yielded the values given in the column entitled 'Overall adj. R^2' in Table 8.3 showing an important increase compared to the simple semi-log approach. The 'Partial adj. R^2' values are adjusted R^2 statistics calculated for the 'Data share' of the observations only where special per-model regression fits are available. They indicate that the fit is somewhat better for observations of car models receiving individualised treatment—which is what could be expected.

One final aspect that calls for some attention here is the existence of high *multicollinearity* between the independent variables. Parameter estimates of linear regression models are known to become unstable and potentially misleading if several covariates are highly correlated. This is usually not a problem if the purpose of a regression analysis is to make predictions of the response variable, 'provided that the values of the independent variables for which inferences are to be made follow the same multicollinearity pattern as the data on which the regression model is based' (Neter et al., 1985, p. 393). Multicollinearity may be an issue, however, if out-of-the-sample predictions are made—which is, to a certain extent, the case for hedonic elementary price indices.

The six previous modelling approaches of the hedonic function were all based on a large and varied data set, so that multicollinearity never seemed to be an issue. This, unfortunately, is not the case for the per-model simple semi-log approach which showed to be much more susceptible to problems related to multicollinearity. Restricting the analysis to individual car models resulted, for instance, in low variance in certain characteristics for certain models. As a result, it could be observed that hedonic functions of this second type yielded highly unrealistic price predictions in the order of less or more than 10^{-10} or 10^{10} Swiss francs respectively for certain reference characteristics vectors. This issue showed to be even more severe if no variable selection was performed.

For this reason, we decided to constrain all hedonic functions introduced in this chapter to return only price predictions that lie in a plausible range between one and one million Swiss francs. For the current per-model simple

semi-log type hedonic functions, moreover, predictions made using one of the model-specific regression equations are only accepted and returned if the resulting price is contained in the price range of the original data. If this is not the case, the price is automatically predicted using the overall simple semi-log model fitted to the entire learning data set.

8.4 Overall examination of the different models

After having estimated these seven types of hedonic functions, one would like to answer the question which of them performs best. If we compare the plots displayed in Fig. 8.1 of the observed versus predicted price values for the different models, it is interesting to see that the PLS models seem to give better predictions in absolute terms for prices higher than about 50 000 francs. For lower price values, the semi-log and double-log models seem to perform better—which is, however, mainly due to the fact that, at the lower end of the price range, many observations are tagged as outliers (marked as coloured crosses) by the linear models. Since the prices are log-transformed in all the models, a fixed absolute difference between untransformed prices has more weight at a lower than at a higher price level. In other words, the least squares criterion is less influenced by an absolute difference between the observed and the predicted price of a car the higher both prices are. Interestingly, this phenomenon apparently has a weaker effect for the PLS than for the linear models.

In Section 5.4, an overall measure $\Delta(\mathfrak{h})$ of the predictive power of a model type $\mathfrak{h}$ has been presented. We are now going to investigate bootstrap estimates $\hat{\Delta}_{\mathrm{B}}$ of the aggregate prediction error for each hedonic function. These are calculated using two different cost functions, namely the *squared-log cost function*

$$c_{\mathrm{ql}}(P, \hat{P}) = (\ln P - \ln \hat{P})^2$$

and the *absolute cost function*

$$c_{\mathrm{a}}(P, \hat{P}) = |P - \hat{P}| \,.$$

The motivation for the squared-log cost function is the fact that it corresponds exactly to the least squares criterion being minimised when estimating the semi-log or double-log models by the ordinary least squares procedure. The absolute cost function is an alternative where the original prices—and not their logarithms—are compared to each other.

Table 8.5 gives bootstrap estimates $\hat{\Delta}_{\mathrm{B}}$ of the aggregate prediction error Δ for each hedonic function using both cost functions introduced above. They

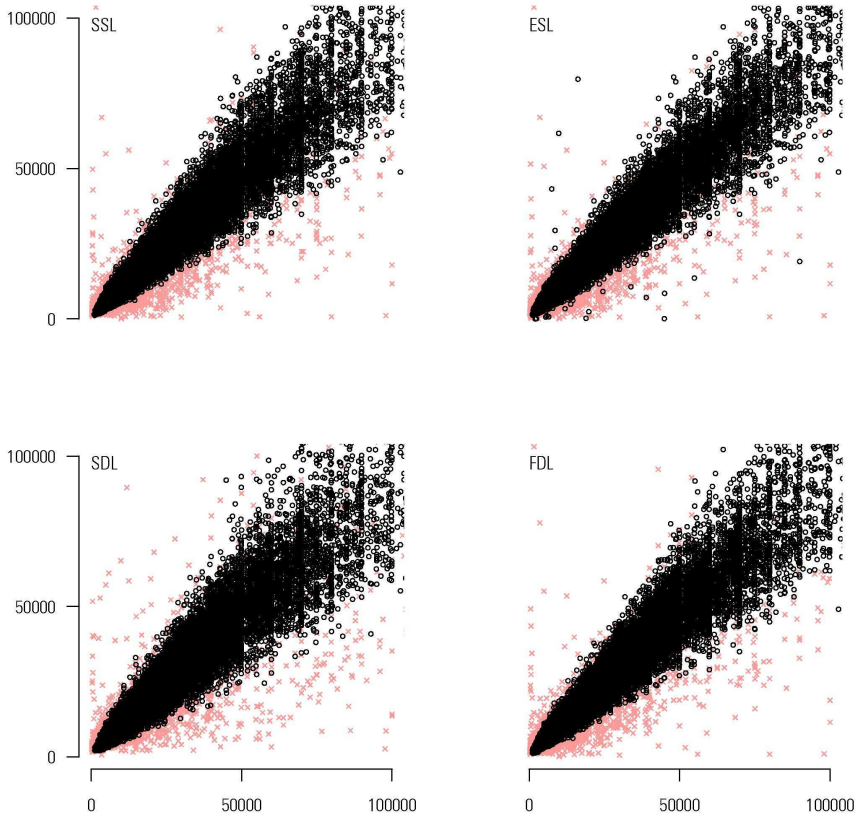
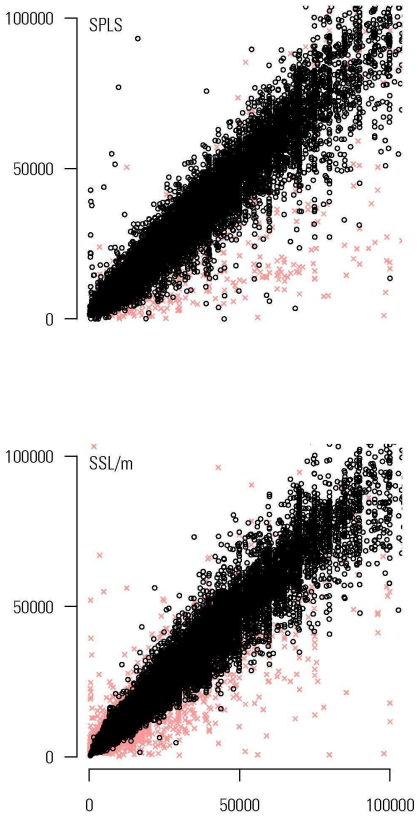


Figure 8.1: **Observed versus fitted price values for different hedonic function estimates in February 2005**



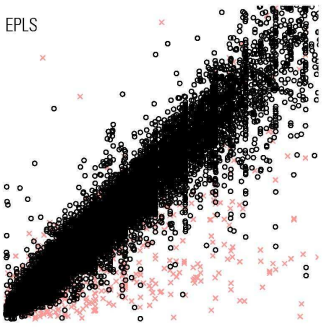


Figure 8.1: **(cont.)**

were calculated based on the formulae (5.9) and (5.10) using $R = 199$ bootstrap replications $\hat{h}_{\star r}$ generated by the case-based resampling algorithm 6.1. In each row, the lowest and thus most preferable value is printed in bold and the second lowest is printed using a regular font weight.

It is important to note that, in order to generate these replications, the model specifications have been imposed to be the same as in the corresponding hedonic functions. In other words, when the replications of the semi-log hedonic function estimate for February 2005 were created, the model formula for all these replications was chosen to be the one that resulted from the variable selection procedure of the original estimate for February 2005. In the per-model simple semi-log model, all the model-specific regressions were performed using the same formulae as in the original hedonic function estimate. Moreover, the number of latent variables in the enhanced PLS model, for example, was held constant for all replications within one time period.

It can be seen from Table 8.5 that the per-model simple semi-log modelling approach seems to perform best for both cost functions. Second best is the enhanced PLS model if the goodness of fit is measured by the absolute cost function. Conversely, the flexible double-log model seems to perform second best for the squared-log cost function. This confirms the observations of the plots in Fig. 8.1 where we had noticed that the PLS models seem to perform better than, e.g., the simple semi-log models in absolute terms over the whole price range.

In order to test which of these differences in estimated aggregate prediction error are statistically significant at a certain level, we pick up the idea presented in Section 5.4 of performing a Wilcoxon signed rank test on the differences of the estimated aggregate prediction errors between the models for the different time periods.

Table 8.4 displays all the alternatives that were supported by exact Wilcoxon signed rank tests through rejecting the corresponding (i.e. opposite) null hypotheses at the 5% level. Results are given for both cost functions individually. The results show that, for the squared-log cost function, the ordering of the models can be summarised by

$$\mathrm{SSL/m} \succ \mathrm{FDL} \succ \left\{ \begin{array}{c} \mathrm{SSL} \\ \mathrm{ESL} \end{array} \right\} \succ \mathrm{SDL} \succ \mathrm{EPLS} \succ \mathrm{SPLS}$$

whereas, for the absolute cost function, this results in

$$\mathrm{SSL/m} \succ \mathrm{EPLS} \succ \mathrm{SPLS} \succ \mathrm{ESL} \succ \mathrm{FDL} \succ \mathrm{SSL} \succ \mathrm{SDL}\,.$$

Table 8.4: **Testing the aggregate prediction error of the different models**

Alternatives supported at the 5% level for the squared-log cost function:

SSL/m $\succ$ FDL	SSL/m $\succ$ SSL	SSL/m $\succ$ ESL	SSL/m $\succ$ SDL	SSL/m $\succ$ EPLS	SSL/m $\succ$ SPLS
–	FDL $\succ$ SSL	FDL $\succ$ ESL	FDL $\succ$ SDL	FDL $\succ$ EPLS	FDL $\succ$ SPLS
	–		SSL $\succ$ SDL	SSL $\succ$ EPLS	SSL $\succ$ SPLS
		–	ESL $\succ$ SDL	ESL $\succ$ EPLS	ESL $\succ$ SPLS
			–	SDL $\succ$ EPLS	SDL $\succ$ SPLS
				–	EPLS $\succ$ SPLS
					–

Alternatives supported at the 5% level for the absolute cost function:

SSL/m $\succ$ EPLS	SSL/m $\succ$ SPLS	SSL/m $\succ$ ESL	SSL/m $\succ$ FDL	SSL/m $\succ$ SSL	SSL/m $\succ$ SDL
–	EPLS $\succ$ SPLS	EPLS $\succ$ ESL	EPLS $\succ$ FDL	EPLS $\succ$ SSL	EPLS $\succ$ SDL
	–	SPLS $\succ$ ESL	SPLS $\succ$ FDL	SPLS $\succ$ SSL	SPLS $\succ$ SDL
		–	ESL $\succ$ FDL	ESL $\succ$ SSL	ESL $\succ$ SDL
			–	FDL $\succ$ SSL	FDL $\succ$ SDL
				–	SSL $\succ$ SDL
					–

The choice of the cost function has thus a large impact on the evaluation of the models. The per-model simple semi-log hedonic function seems to perform best with regard to both cost functions. Apart from that, the only relations that hold for both cost functions are EPLS $\succ$ SPLS and FDL $\succ$ SSL $\succ$ SDL giving some support for preferring the flexible or enhanced models to their simple alternatives. No overall ordering, however, can be established between the simple or enhanced semi-log or double-log and the PLS models. This is particularly interesting if one takes into account that the PLS models use much more of the information contained in the data by including the `model` variable instead of `make`. The dominance of the per-model simple semi-log model shows that an appropriate inclusion of the `model` variable *does* yield better predictions.

An interesting question for further research would certainly be to estimate these aggregate prediction errors by other means such as adjusted cross-validation or the 0.632 bootstrap estimate (see, e.g., Davison and Hinkley, 1997, Sect. 6.4.1, Efron and Tibshirani, 1993, Chap. 17, or Mevik and Cederkvist, 2004) and to compare the results to those obtained here. Moreover, if outlier detection could be accomplished by a method that is independent from the regression model of the hedonic function, one could first 'clean' the database and then estimate all the models and bootstrap replications using the same data. This would significantly increase the relevance of such a comparison of estimated aggregate prediction errors. An important weakness of the results presented here is the fact that for each of the six model types, other data were finally left over for estimating the models and prediction errors. Differences between the prediction errors may thus to a certain extent be due to the fact that not all the models predict prices for the same set of observations. Therefore, the above preference orderings must, in any case, be interpreted with care.

One conclusion that seems to be supported by these results is that partitioning the data and building separate models for different sub-populations allows to increase the predictive power of the resulting hedonic function. One could now construct per-model hedonic functions not only for the simple semi-log but also for other regression approaches. Moreover, it could be studied whether partitioning the data with regard to other variables than `model` or `pricenew` could yield even better results. The extreme case of such a partitioning approach is probably the fitting of regression trees to such data. It is an open question, however, whether such an extreme solution is really the best one.

Table 8.5: **Goodness of fit estimations for different regression models**

Month	Criterion	SSL	ESL	SDL	FDL	SPLS	EPLS	SSL/m
Oct 2004	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.8081E−02	2.9349E−02	3.4221E−02	**2.7064E−02**	5.3260E−02	5.1217E−02	**2.5515E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	2.8530E+03	2.6783E+03	3.1239E+03	2.7555E+03	2.7448E+03	**2.6578E+03**	**2.4854E+03**
Nov 2004	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.7894E−02	2.9084E−02	3.3808E−02	**2.6785E−02**	5.2677E−02	4.8992E−02	**2.4082E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	2.9155E+03	2.7151E+03	3.1613E+03	2.8117E+03	2.6672E+03	**2.5914E+03**	**2.4073E+03**
Dec 2004	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.7987E−02	2.8632E−02	3.3810E−02	**2.7024E−02**	4.6870E−02	4.4499E−02	**2.2901E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0053E+03	2.8779E+03	3.2660E+03	2.8996E+03	2.8084E+03	**2.7036E+03**	**2.3978E+03**
Jan 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.8492E−02	3.0808E−02	3.5402E−02	**2.7519E−02**	5.3066E−02	4.9569E−02	**2.3951E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	2.9333E+03	2.7397E+03	3.2028E+03	2.8170E+03	2.7215E+03	**2.6507E+03**	**2.3684E+03**
Feb 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.8264E−02	2.8879E−02	3.4914E−02	**2.7328E−02**	5.1622E−02	4.7536E−02	**2.3182E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0445E+03	2.8187E+03	3.3236E+03	2.9333E+03	2.7729E+03	**2.6975E+03**	**2.4196E+03**
Mar 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.9010E−02	2.9989E−02	3.5016E−02	**2.7889E−02**	5.4081E−02	4.8530E−02	**2.3831E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0366E+03	2.8655E+03	3.2713E+03	2.9088E+03	2.7489E+03	**2.6950E+03**	**2.3669E+03**
Apr 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	2.9903E−02	3.0805E−02	3.6987E−02	**2.8797E−02**	5.6043E−02	5.1356E−02	**2.5101E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0487E+03	2.8589E+03	3.3232E+03	2.9184E+03	2.7978E+03	**2.7408E+03**	**2.4174E+03**
May 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	3.3074E−02	3.4221E−02	4.0109E−02	**3.1865E−02**	6.2354E−02	5.6992E−02	**2.7236E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.1321E+03	2.8783E+03	3.4033E+03	2.9955E+03	2.8539E+03	**2.7595E+03**	**2.4007E+03**
Jun 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	3.1732E−02	3.2496E−02	3.8015E−02	**3.0449E−02**	5.7900E−02	5.3107E−02	**2.6157E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0922E+03	2.8816E+03	3.3526E+03	2.9720E+03	2.7488E+03	**2.7084E+03**	**2.3877E+03**
Jul 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	3.2265E−02	3.2797E−02	3.8443E−02	**3.0969E−02**	6.0148E−02	5.6955E−02	**2.6691E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.1505E+03	3.0014E+03	3.4010E+03	3.0171E+03	2.9542E+03	**2.8524E+03**	**2.4374E+03**
Aug 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	3.1727E−02	3.2166E−02	3.7219E−02	**3.0454E−02**	5.6172E−02	5.3636E−02	**2.6246E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0075E+03	2.8645E+03	3.1997E+03	2.8786E+03	2.8023E+03	**2.7487E+03**	**2.3376E+03**
Sep 2005	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{ql}})$	3.1342E−02	3.2425E−02	3.8024E−02	**3.0257E−02**	5.5557E−02	5.1844E−02	**2.6372E−02**
	$\hat{\Delta}_{\mathrm{B}}(c_{\mathrm{a}})$	3.0473E+03	2.8862E+03	3.3139E+03	2.9337E+03	2.8304E+03	**2.8136E+03**	**2.4133E+03**

Oct 2005	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{ql}})$	3.1303E−02	3.2009E−02	3.7507E−02	**3.0168E−02**	5.7078E−02	5.3007E−02	**2.6074E−02**
	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{a}})$	2.9347E+03	2.7862E+03	3.2200E+03	2.8586E+03	2.7672E+03	**2.7468E+03**	**2.3294E+03**
Nov 2005	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{ql}})$	3.2568E−02	3.3678E−02	3.9504E−02	**3.1687E−02**	5.8283E−02	5.4599E−02	**2.7056E−02**
	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{a}})$	3.0758E+03	2.8723E+03	3.3412E+03	2.9671E+03	2.7822E+03	**2.7691E+03**	**2.4423E+03**
Dec 2005	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{ql}})$	3.2123E−02	3.1986E−02	3.9901E−02	**3.1181E−02**	5.7226E−02	5.4055E−02	**2.7024E−02**
	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{a}})$	2.9432E+03	2.7711E+03	3.2572E+03	2.8571E+03	2.6999E+03	**2.6774E+03**	**2.3125E+03**
Jan 2006	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{ql}})$	3.1527E−02	3.2834E−02	3.9339E−02	**3.0476E−02**	5.8688E−02	5.3886E−02	**2.6386E−02**
	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{a}})$	2.8989E+03	2.7432E+03	3.2094E+03	2.8159E+03	2.6708E+03	**2.6647E+03**	**2.2937E+03**
Feb 2006	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{ql}})$	3.1995E−02	3.2831E−02	4.0976E−02	**3.0843E−02**	5.8099E−02	5.3517E−02	**2.6292E−02**
	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{a}})$	2.9438E+03	2.7701E+03	3.2927E+03	2.8452E+03	2.7359E+03	**2.6722E+03**	**2.2964E+03**
Mar 2006	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{ql}})$	3.4133E−02	3.5175E−02	4.2632E−02	**3.3157E−02**	6.3423E−02	5.9193E−02	**2.8743E−02**
	$\hat{\Delta}_{\mathrm{B}}\,(c_{\mathrm{a}})$	2.9939E+03	2.7990E+03	3.3418E+03	2.9002E+03	2.7383E+03	**2.7248E+03**	**2.3581E+03**

8.5 Concluding remarks

If one tries to summarise the experience presented in this chapter about estimating hedonic functions, the following aspects have to be mentioned:

Variable selection has two different dimensions. In a first stage, it is important to carefully select variables that are potentially price-relevant for a certain elementary aggregate in order to survey them for further quantitative analysis. This needs to be done based on expert judgements. In a second stage, variable selection plays its role within the regression procedure and is there a purely statistical issue. For this second stage, standard techniques such as the one applied here minimising the AIC are applicable.

Outlier detection seems to be one of the most difficult issues inherent in the practical estimation of a hedonic function. Ideally, it should be accomplished independently from the estimation of a regression model. Only then, the goodness of fit estimations of different regression models can be compared with a clear conscience. Using robust techniques for estimating hedonic functions is certainly favourable. In this case, however, the estimates of aggregate prediction error should probably be generated using a cost function that does not give too much weight to large differences stemming from real outliers in the data.

The *functional form* of a hedonic function needs to be chosen in such a way that it fits the data and not necessarily any theoretical reasoning. The family of candidate models should be broad enough to potentially provide good approximations of the truth. Modern adaptive regression approaches are certainly interesting techniques for modelling hedonic functions, because they may approximate a very broad universe of functional forms. There is, however, no *a priori* evidence that they would substantially outperform conventional models, such as the semi-log linear regression, in terms of predictive power.

Automated processes need to be established for the estimation of hedonic functions in a productive environment. Second-stage variable selection, outlier detection, and model estimation should not require the intervention of experts. The methods need to be designed in such a manner that they can cope with different potential data sets for separate time periods. Nevertheless, an ongoing quality assessment of the estimations will in any case be necessary.

CHAPTER 9

Estimating Hedonic Elementary Price Indices for Used Cars

In this last chapter, we are going to apply the tools developed so far for estimating hedonic elementary price indices for used cars. These estimates relate on the theoretical concepts introduced in Chapter 6 and on the estimated hedonic functions presented in Chapter 8. The base period of all indices presented in this chapter will be set to October 2004.

Given the conceptual limitations of this data set (see Section 7.4), it is important not to overrate the numerical results that are going to be presented in this chapter. Therefore, we deliberately refrain from comparing these index estimates with the official results published by the Swiss Federal Statistical Office. Our interest lies in analysing the outcomes of different estimation methods, as far as this is possible.

9.1 Hedonic elementary price index estimates

It has been shown in the previous chapters that any bilateral hedonic elementary price index estimator is determined by essentially three mutually independent dimensions. The first dimension consists of the potential index estimators where the five formulae given in Table 6.1 are individual representatives. The second dimension reflects the type or functional form of the hedonic functions. Seven alternative approaches have been investigated in Chapter 8.

Table 9.1: **Reference sample size for individual (base or current month) samples or symmetric samples combined with the base month October 2004.**

Month	Individual	Combined
Oct 2004	81209	–
Nov 2004	104161	118745
Dec 2004	100893	139327
Jan 2005	97319	147320
Feb 2005	99188	155781
Mar 2005	101354	162637
Apr 2005	105520	170754
May 2005	104816	173268
Jun 2005	107811	178406
Jul 2005	106159	178556
Aug 2005	104809	178494
Sep 2005	106935	181545
Oct 2005	109497	184990
Nov 2005	109992	186205
Dec 2005	109202	185998
Jan 2006	110301	187514
Feb 2006	109685	187443
Mar 2006	115618	193709

The third dimension assembles the potential reference characteristics samples to be used for the index estimation. Here, we are going to consider either the sample of cars available in the base period, the current period or in both periods symmetrically combined. Since the reference sample operationalises the reference quality spectrum of the index, choosing either of these three samples has an impact on the nature of the index. If the base period sample is used for the index estimation, this mimics a Laspeyres-type approach since, in the aspect already described in Section 3.3, the base period population is held constant. Using the current period sample, on the other hand, rather follows the philosophy of Paasche-type indices where the constant population represents the current period. Taking the symmetric sample finally leads to an implementation of an index that is similar to the true hedonic adjacent periods price index (Brachinger, 2002, pp. 6–8).

The samples used for the estimations are all extracted from the database using the procedure specified in Section 7.2.2. In contrast to the learning data sets fetched for fitting the models in Chapter 8, where only advertisements which had been changed during a specific month were included,

Table 9.2: **Bilateral hedonic elementary price index estimates for February 2005. Comparison of different hedonic functions, index formulae, and reference sample types.**

Formula	Reference	Functional form						
		SSL	ESL	SDL	FDL	SPLS	EPLS	SSL/m
$\hat{I}_{\mathrm{J}}^{0:1}$	base	1.0063	1.0041	1.0054	1.0068	1.0025	1.0056	1.0080
	symmetric	1.0072	1.0047	1.0054	1.0070	1.0027	1.0054	1.0067
	current	1.0079	1.0055	1.0058	1.0076	1.0033	1.0057	1.0060
$\hat{I}_{\mathrm{D}}^{0:1}$	base	1.0013	0.9991	1.0085	1.0047	1.0054	1.0035	1.0028
	symmetric	1.0026	1.0000	1.0092	1.0049	1.0063	1.0037	1.0032
	current	1.0032	1.0006	1.0094	1.0053	1.0075	1.0041	1.0031
$\hat{I}_{\mathrm{HD}}^{0:1}$	base	1.0106	0.9958	1.0019	1.0102	0.9823	1.0004	1.0127
	symmetric	1.0104	0.9801	1.0018	1.0092	0.9504	0.9902	1.0085
	current	1.0108	0.9760	1.0022	1.0094	0.9400	0.9872	1.0062
$\hat{I}_{\mathrm{C}}^{0:1}$	base	1.0068	1.0051	1.0060	1.0074	1.0061	1.0101	1.0143
	symmetric	1.0077	1.0057	1.0061	1.0076	1.0063	1.0100	1.0131
	current	1.0084	1.0064	1.0065	1.0082	1.0069	1.0106	1.0123
$\hat{I}_{\mathrm{HC}}^{0:1}$	base	1.0057	1.0027	1.0047	1.0062	0.9971	1.0011	1.0019
	symmetric	1.0066	1.0031	1.0048	1.0064	0.9957	1.0005	1.0005
	current	1.0073	1.0039	1.0051	1.0069	0.9960	1.0005	0.9998

all available advertisements are now retained within these reference samples. While it was important to have up-to-date data points for fitting the models, it is now more important to have a broad universe of items included in the reference quality spectrum. As a consequence, the number of characteristics vectors contained in reference samples of a single month is higher than the number of data points used for the estimation of the hedonic functions. (The latter is displayed in the column labelled 'Original data' of Table 8.1.) When combining a base and a current period in order to build a 'symmetric' reference sample, individual advertisements occurring in both months were only included once—using the version with the highest revision number available during the given time periods. Again, all advertisements with prices outside the range of 52 to 999 999 CHF were discarded (cf. Section 7.3). Table 9.1 shows the reference sample sizes both for individual one-month samples and for combined two-month samples.

Based on these three dimensions, we will have to choose from a $5 \times 7 \times 3$ matrix of potential estimates for each individual hedonic elementary price

index value to be estimated. This is a discrete grid with 105 edges in a three-dimensional space where each dimension is on its own potentially high-dimensional and continuous. (Table 9.2 displays these point estimates for February 2005 as an example.) It is difficult to overlook all these estimates as a whole, especially if this needs to be done over several time periods.

We therefore first identify a standard case and compare deviations from this case in only one or two dimensions at a time. Our standard index formula will be the one by Jevons since it emerges from both (6.3) and (6.4) and seems to have the largest support in the literature. As a standard functional form of the hedonic function, we are going to take the per-model simple semi-log approach since it performed best in terms of aggregate prediction error among all of the alternatives presented in Chapter 8. The standard reference sample will be the symmetric inclusion of base and current period observations.

Fig. 9.1 shows three times the bilateral hedonic elementary price index estimates obtained using the reference model. In each of the three frames, this model is confronted with alternatives obtained by moving along one of the three dimensions described above. In the top frame, results from the six alternative index formulae are presented, showing that the index estimates obtained by the Jevons formula lie somewhere in the center of the different alternatives. This is, at least for the three alternative implementations of (6.4), due to the theoretical inequality relations between the arithmetic, geometric, and harmonic means (see e.g. ILO et al., 2004, p. 361). In the middle frame, secondly, the seven different functional forms are varied, while the third frame displays the results for the three alternative specifications of the random sample that were analysed.

Fig. 9.2 shows the effect of changing the reference sample type individually for each index formula while holding the modelling approach of the hedonic function constant at its reference level 'SSL/m'. Fig. 9.3 holds constant the reference sample type ('symmetric') and shows the effect of changing the functional form of the hedonic regressions. The interaction between the functional forms and the reference sample type is finally depicted in Fig. 9.4 for estimates obtained using the Jevons formula.

A number of most prevalent conclusions that can be drawn from these results is now going to be summarised:

The price index estimates seem to be to a large extent *insensitive to the choice of the reference sample* among the base period, the current period, and the inclusion of both periods symmetrically. This finding emerges from the fact that the curves in the lower frame of Fig. 9.1 as well as those in Fig. 9.2 almost coincide. The average price change between the base and any

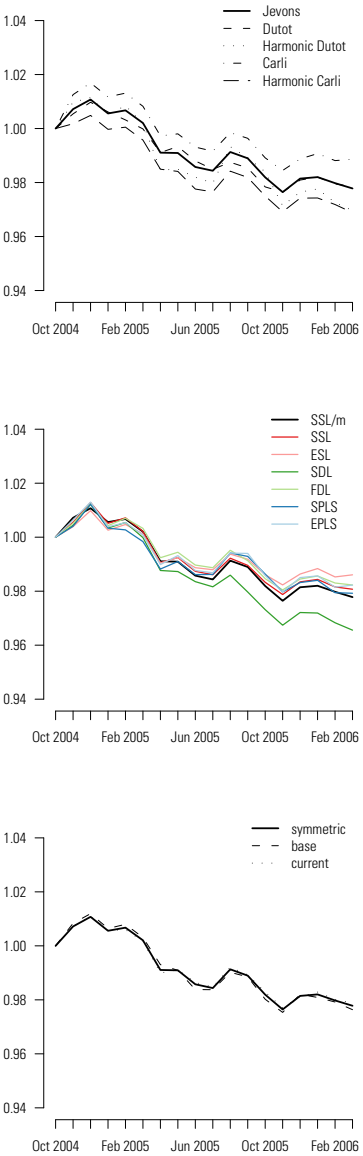


Figure 9.1: **Bilateral hedonic elementary price index estimates using the reference model. Comparison with alternatives in each of the three main dimensions.**

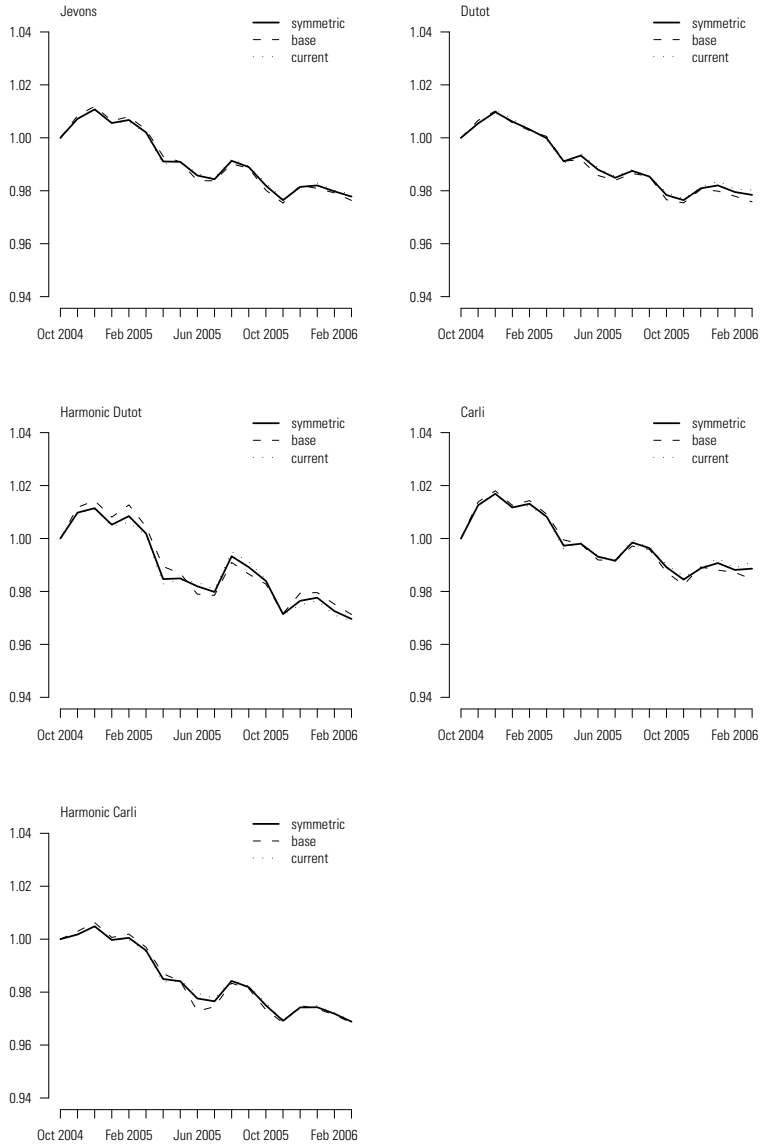


Figure 9.2: **Bilateral hedonic elementary price index estimates using the per-model simple semi-log modelling approach for the hedonic functions. Comparison of alternative index formulae and reference sample types**

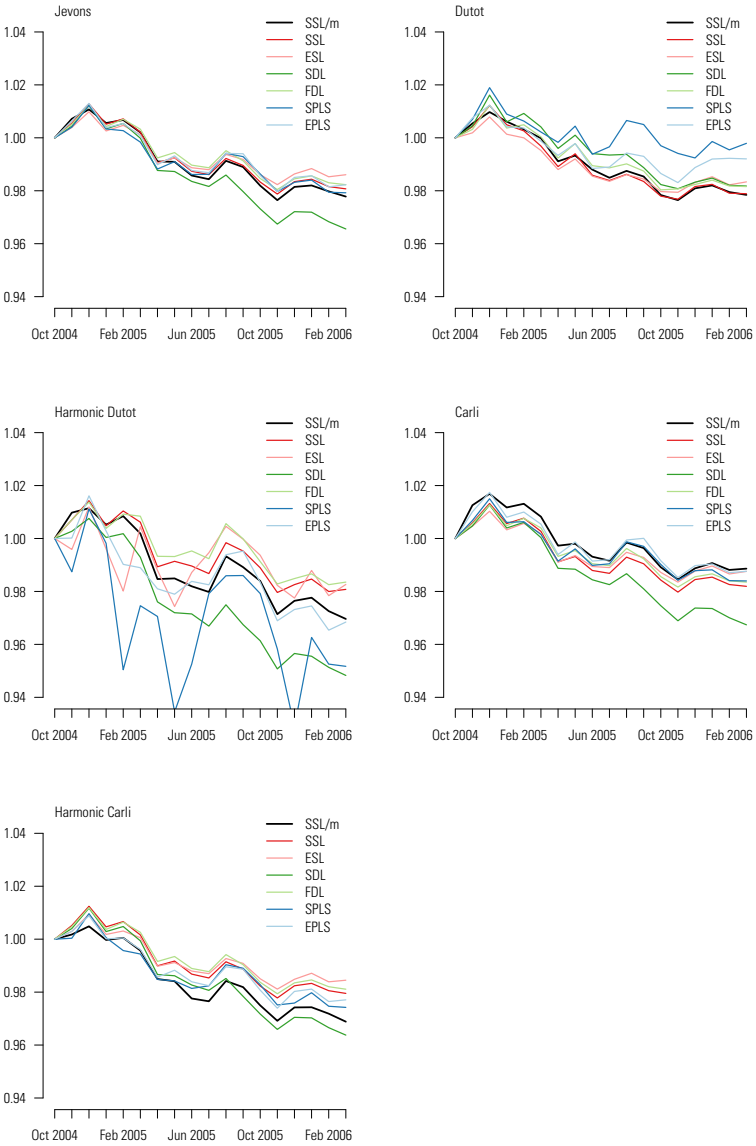


Figure 9.3: **Bilateral hedonic elementary price index estimates using the symmetric reference sample. Comparison of alternative index formulae and functional forms of the hedonic functions**

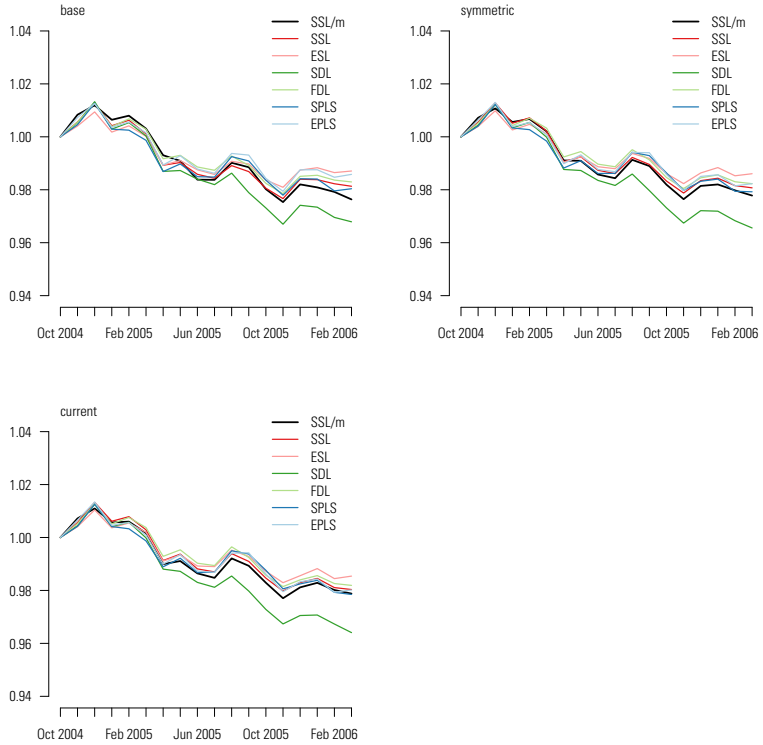


Figure 9.4: **Bilateral hedonic elementary price index estimates using the Jevons formula. Comparison of alternative reference sample types and functional forms of the hedonic functions.**

current period seems to be almost identical for any of the three populations used as reference quality. This might be due to the fact that the quality spectrum does not change dramatically between the different periods.

In general, indices estimated using the *simple double-log model* tend to have lower values than those estimated using the other functional forms, at least if they are calculated using the Jevons formula. The discrepancy, however, should not be overrated since the SDL model did not perform very well in terms of goodness of fit (cf. Section 8.4).

According to Fig. 9.4, the *variance of the index estimates over the different functional forms* of the hedonic functions tends to increase with time. Moreover, according to Fig. 9.3 as well as to its two counterparts Fig. A.1 and Fig. A.2 in the appendix which show the results based on the base period and the current period reference sample respectively, it seems to be lower for the Jevons, Carli, and Harmonic Carli than for the Dutot and Harmonic Dutot indices. It thus seems that estimates of the index type (6.4) are less sensitive to the specification of the hedonic functional form than estimates of the index type (6.3). This is probably due to the fact that the latter is essentially a ratio of means whereas the former is a mean of ratios.

Finally, the observed *variance of the index estimates over the five different index formulae* is not surprising because these formulae are all estimators of different theoretical indices. The choice of an index estimator depends first of all on the specification of the index we want to estimate. Comparing results delivered by the five formulae among each other is thus not informative with regard to the robustness of these estimates.

9.2 Estimates of bootstrap confidence intervals

In order to get some further idea about the quality of the point estimates presented so far, we are now going to estimate confidence intervals for some of the bilateral hedonic elementary price indices. In Section 6.2, three approaches for estimating such intervals using bootstrap replications have been presented. The aim of the current section is also to compare these three techniques for those hedonic functions where all of them are applicable.

The bootstrap replications were all built on the same data as the original hedonic function estimates. In other words, the outlier detection procedure was not repeated for each replication. Moreover, as in Section 8.4, the model specification was always fixed to be the same for each replication as far as this was possible in order to catch the variability in the data rather than any variability in the models. For example, the replications of the simple semi-

log hedonic functions were all built using the model formula that resulted from the model selection algorithm of the original estimate. This means that exactly the same subset of exogeneous variables was used for all replications, and there was no individual model selection involved any more. The same holds for both of the double-log models. In the enhanced semi-log model, the specification was held constant individually for both the part of the model including the `pricenew` variable and the other part excluding it. In the per-model simple semi-log model, the model formulae applied to the different car models were fixed individually as well. This regime was somewhat weaker for the PLS models, where only the number of latent variables was fixed for all replications in a month—always based on the number applied for the original estimates.

Choosing the number of replications R represents a trade-off between the precision of the estimates and the computing time needed. The higher R, the better the empirical distribution of the estimates approximates the underlying reality. The higher R, however, the more time is necessary to produce the estimates. For the current study, a number of $R = 199$ replications was chosen. This choice ensures that, when determining the empirical 2.5%, 5%, 95%, and 97.5% quantiles of $\zeta_\star^{0:1}$ as defined in (6.6), an integer number of at least five observations is left over at each tail of the distribution. Some analyses using a higher number of replications showed no relevant change in the resulting empirical quantiles. Table 9.3 shows the average time that was needed to produce one replication of $\zeta_\star^{0:1}$ along with the total time needed to produce $R = 199$ replications either using one single CPU or using four or eight CPUs in parallel.

The durations were measured when running the simulations on the Beowulf cluster nodes described in Section 7.2.3. They depend to a certain extent on the implementation of the algorithms. Linear reductions of all these time intervals could certainly be attained by purging the source code appropriately. In the current implementation, for example, replications using all five index formulae were simultaneously generated, and intermediate results were stored in order to estimate the aggregate prediction error of the hedonic functions afterwards. If all such ballast is removed, the computations would be somewhat quicker.

If we now look at the individual resampling algorithms, it is evident that the case-based version is the most generally applicable since it does not rely on the availability of variance-adjusted residuals. It is therefore similarly applicable to all seven hedonic modelling approaches proposed in the previous chapter. Its implementation was straightforward. $R = 199$ new estimates $(\hat{h}^0_{\star r}, \hat{h}^t_{\star r})$ $(r = 1, \ldots, R)$ were obtained on the basis of resampled price-

Table 9.3: **Average time needed for creating bootstrap replications.**

Model	**One replication**	$R = 199$ **replications**		
		1 CPU	4 CPUs	8 CPUs
SSL	3.2 min	10 h 33 min	2 h 38 min	1 h 19 min
ESL	5.4 min	17 h 48 min	4 h 27 min	2 h 13 min
SDL	3.2 min	10 h 37 min	2 h 39 min	1 h 20 min
FDL	3.7 min	12 h 7 min	3 h 2 min	1 h 31 min
SPLS	4.0 min	13 h 21 min	3 h 20 min	1 h 40 min
EPLS	4.8 min	15 h 48 min	3 h 57 min	1 h 58 min
SSL/m	28.7 min	3 d 23 h 10 min	23 h 47 min	11 h 54 min

characteristics combinations from the original data. One technical problem encountered, however, was that the three gigabytes of random access memory available were insufficient for routinely creating replications of the SPLS and EPLS models, due to the inclusion of the model variable. For this reason, the number of observations used for building the replications of the two PLS models was fixed to be half of the number of observations used for the original estimate. In other words, wherever the number N^t occurs in Algorithm 6.1, it was replaced by $N^t/2$ for the PLS models. The variance of the index estimators based on PLS hedonic regression models might therefore be somewhat overestimated.

The implementation of the model-based and the wild bootstrap algorithms was less straightforward. The main problem here was the proper specification of variance-adjusted residuals. With the choice of modelling the natural logarithm of the dependent price variable, one implicitly assumes that the variance of the error term ϵ^t is proportional to $(\mathrm{E}[P^t])^2$ (see e.g. Montgomery and Peck, 1992, p. 98), as has already been mentioned in Section 5.3.1. The model-based bootstrap algorithm, however, depends on the homoscedasticity of the error term. We therefore defined the modified residuals used in the algorithm as

$$r_n^t = \frac{\ln p_n^t - \ln \hat{h}^t(\boldsymbol{m}_n^t)}{(1-h_n)^{1/2}}\,. \tag{9.1}$$

instead of (6.7), where h_n is the nth diagonal element of the hat matrix of the corresponding linear regression model. This reflects that the assumption of homoscedasticity is translated to the error term η^t of the transformed model (5.4) or (5.5). As a consequence, the step 1b of Algorithm 6.2 needed to be changed into $p_{\star rn}^t = \exp(\ln \hat{h}^t(\boldsymbol{m}_n^t) + \epsilon_{\star n}^t)$ accordingly. In the wild bootstrap

Table 9.4: **Lengths of 90% and 95% bootstrap confidence intervals (multiplied by 100) over the period from Nov 2004 to Mar 2006 for Jevons-type index estimates. Comparison of different modelling and bootstrap approaches.**

Model	Bootstrap algorithm	Interval lengths		Mean length difference	
		90%	95%	90%	95%
SSL	case-based	0.41–0.52	0.49–0.63	0.003	0.000
	model-based	0.37–0.49	0.45–0.59	0.002	0.001
	wild	0.41–0.52	0.51–0.60	0.000	0.001
ESL	case-based	0.40–0.54	0.47–0.65	0.003	0.002
	model-based	0.39–0.50	0.45–0.63	−0.000	−0.000
	wild	0.43–0.51	0.51–0.60	0.002	−0.003
SDL	case-based	0.45–0.56	0.52–0.67	−0.000	0.003
	model-based	0.44–0.53	0.51–0.63	0.005	0.001
	wild	0.45–0.56	0.52–0.71	0.002	−0.002
FDL	case-based	0.42–0.53	0.50–0.64	−0.000	0.003
	model-based	0.38–0.52	0.47–0.59	0.001	0.003
	wild	0.40–0.55	0.49–0.62	0.001	0.000
SPLS	case-based	0.85–1.07	1.04–1.35	0.008	0.007
EPLS	case-based	0.86–1.07	1.06–1.34	0.008	−0.005
SSL/m	case-based	0.46–0.58	0.55–0.71	0.014	0.008
	model-based	0.37–0.51	0.43–0.61	0.013	0.010
	wild	0.38–0.50	0.44–0.63	0.012	0.020

algorithm 6.3, the step concerned is also 1b which needed to be adapted into $p^t_{\star rn} = \exp(\ln \hat{h}^t(\boldsymbol{m}^t_n) + r^t_n\, \epsilon^t_{\star rn})$.

Model-based and wild bootstrap confidence intervals were estimated for the SSL, ESL, SDL, FDL, and SSL/m models. In the per-model simple semi-log modelling approach, the residuals of the regressions for each car model were modified individually according to (9.1) using their specific hat matrix coefficients, but then pooled together for generating the bootstrap replications. In order to modify the residuals of observations where no model-specific regression model had been estimated, the hat matrix of the overall model was used. In the enhanced semi-log model, similarly, residuals of observations where `pricenew` was available were modified using the hat matrix of the regression model including `pricenew`, while all the other residuals were modified using the respective hat matrix element of the overall model.

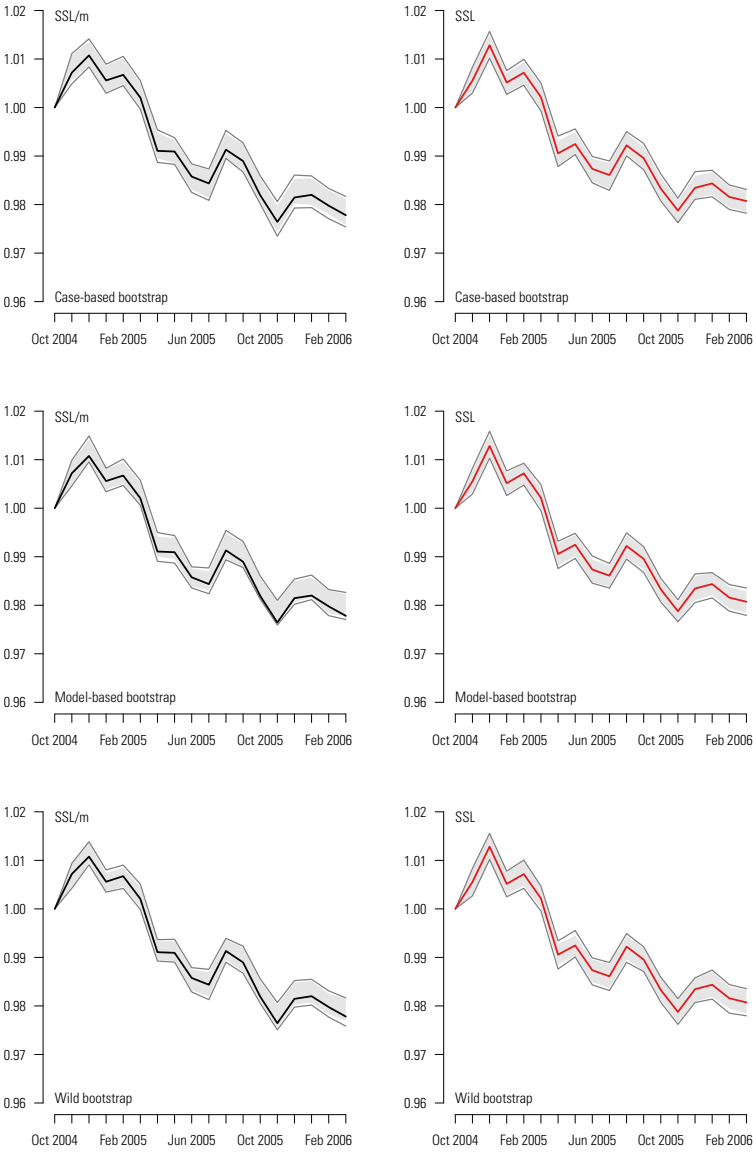


Figure 9.5: **Jevons-type bilateral hedonic elementary price index estimates for the symmetric reference samples. Comparison of different functional forms of the hedonic function and of different bootstrap approaches.**

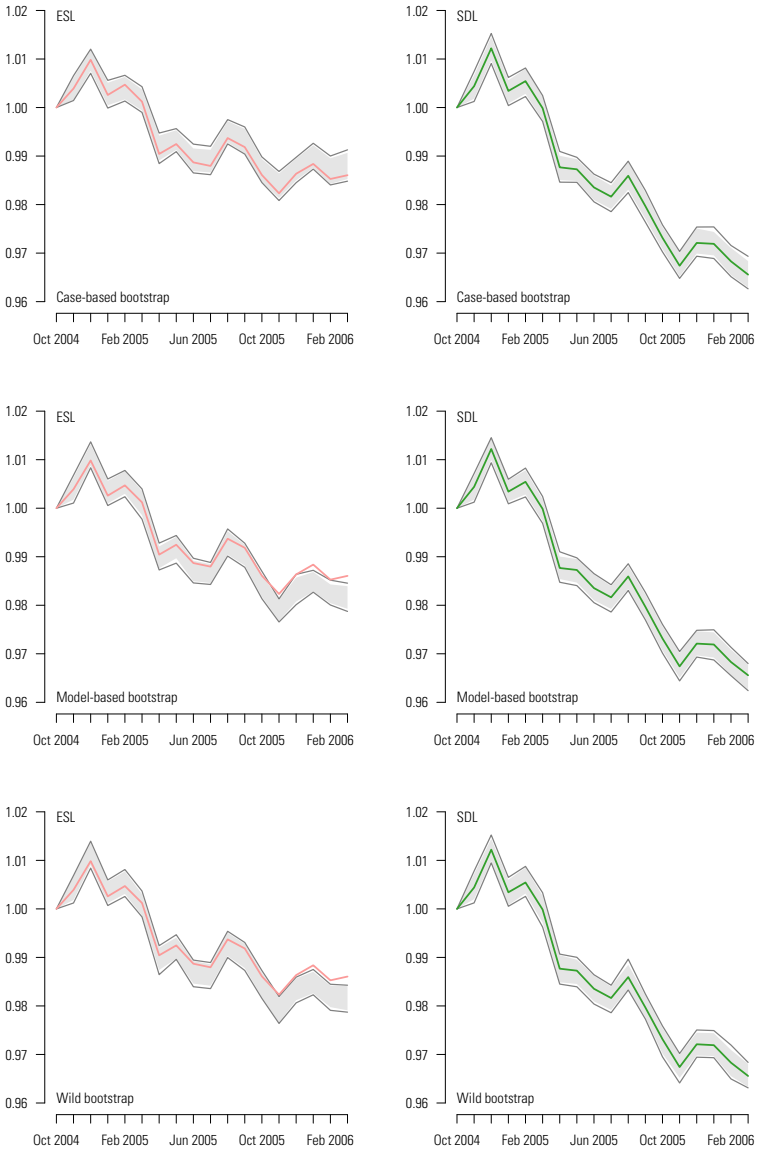


Figure 9.5: **(cont.)**

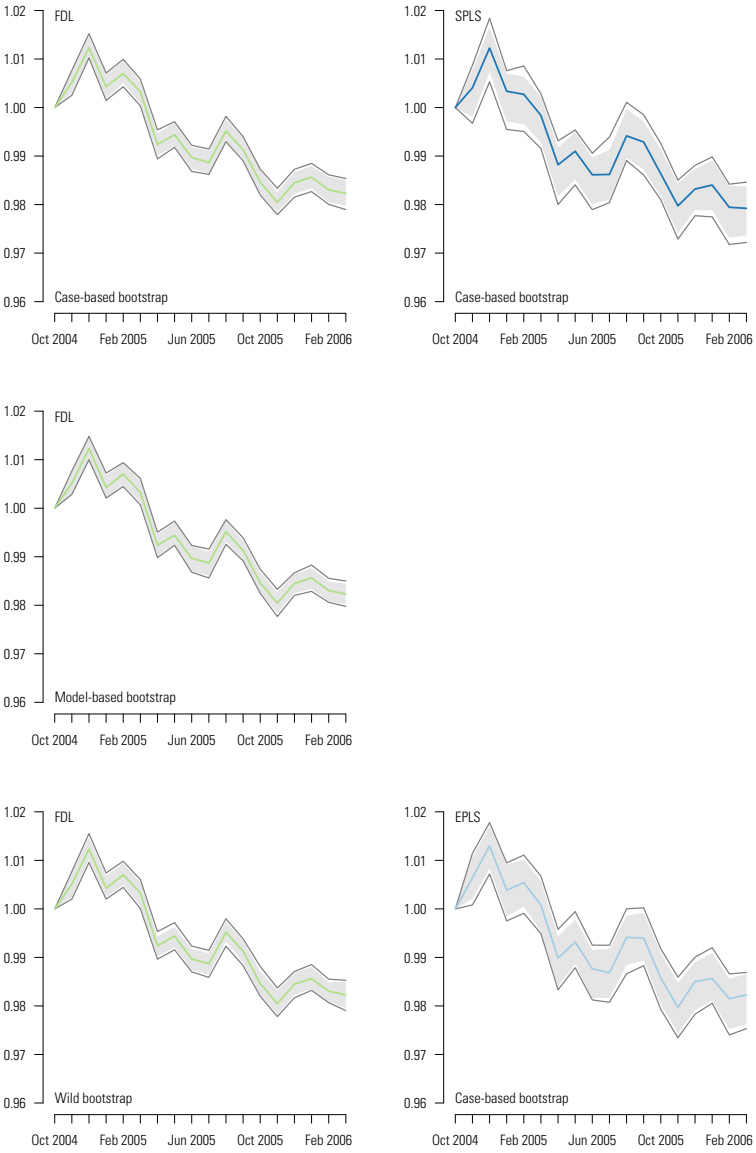


Figure 9.5: **(cont.)**

Using (6.9), 90% and 95% confidence intervals for bilateral elementary price indices in all the months under consideration were estimated. They are displayed for the Jevons formula along with the point estimates of the indices in Fig. 9.5. The gray areas represent the 90% confidence intervals while the gray lines show the borders of the 95% intervals. The respective ranges of interval lengths (upper minus lower bound) over time are listed in Table 9.4, each value being multiplied by the factor of 100 for reasons of better legibility.

It is interesting to compare the interval lengths calculated above to those obtained when the reference characteristics $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$ are not resampled in (6.8), i.e. when the original values $\boldsymbol{m}_n$ are used instead of $\boldsymbol{m}_{\star n}$. This second case imitates the idea that the reference characteristics distribution $\mathbb{P}_{\boldsymbol{M}}$ puts equal probability weight on the values $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$, whereas in the above case $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N$ is seen as a sample from an unknown $\mathbb{P}_{\boldsymbol{M}}$. If we subtract the length of the confidence intervals where no resampling of the characteristics is done from those in the original model, we see hardly any difference. The respective average differences—again multiplied by 100—are contained in the two columns labelled 'Mean length difference' of Table 9.4. We conclude that the variance of the index estimators with regard to the randomness of the reference sample seems to be negligible—presumably due to the large size of the sample.

If we compare the various results obtained, several phenomena can be observed. These will now be shortly summarised and commented:

The confidence intervals presented here are all *conditioned on the choice of the supposed functional form of the hedonic functions.* Given the hedonic regression model, they contain the true but unknown index values with a certain probability (90 or 95 percent). The variance observed is the variance due to sampling from all the five price and characteristics distributions the index depends on (cf. Section 4.1). It does, however, not cover the variance due to choosing a specific regression model for approximating the true hedonic functions. We see from Fig. 9.5 that there are interval estimates for different regression approaches that do not even intersect (e.g. the confidence intervals for the ESL and SDL models in March 2006). Unconditioned confidence interval for hedonic elementary price indices would therefore be considerably larger.

The *lengths of the confidence intervals* obtained using the essentially linear hedonic regression models (SSL, ESL, SDL, FDL, and SSL/m) are all comparable and mostly independent of the bootstrap algorithm used. In contrast, the lengths of the intervals obtained using the two PLS hedonic regression models are generally about twice as large as the former. To a

certain extent, this might be due to the fact that, for the PLS models, the replications of the hedonic function estimates were based on only half of the original observations as mentioned above. Increasing the sample size, however, does not usually decrease linearly the variance of any estimator, especially for samples of this size. It seems highly probable that most of the additional variance observed here is really due to the choice of the PLS approach.

The most prominent *differences between the three bootstrapping approaches* can be identified for the estimates based on the ESL hedonic functional model. Whereas the two residual-based resampling approaches are comparable between each other, they contrast somewhat with the results obtained using the case-based resampling algorithm. There is no apparent explanation for this phenomenon. In general, however, the resulting confidence intervals are comparable for all the three bootstrap approaches—in spite of their individual advantages and disadvantages. It appears that the case-based resampling procedure is sufficient and, due to its ease of use, preferable to the two other approaches.

Fig. A.3 and Table A.5 in the appendix show the same results for the index estimated using the *Dutot formula*, and all the observations mentioned above similarly hold. As a general rule, the confidence intervals are longer for the Dutot than for the Jevons indices. This means that the index (6.3) with $\varphi(x) = \ln x$ can be estimated more precisely than the alternative implementation of (6.3) with $\varphi(x) = x$—which is certainly an empirical argument for using the former of the two in practice.

Due to the nonparametric nature of these confidence intervals, they are *not generally centered* around the respective point estimates. It may even happen that the point estimate lies outside any of these intervals. This is the case when the point estimate lies in the upper or lower tails of the corresponding distribution (see also Beer, 2007 for an example of such a situation). When implementing this kind of hedonic elementary price index estimators in a production environment, it should therefore be discussed whether it is wiser to use the original point estimate or rather an appropriate summary statistic, e.g. the median, of the estimated probability distribution as an estimate for the corresponding index.

CHAPTER 10

Conclusions, Recommendations, and Open Questions

Why such an effort? In his critical review on price hedonics, HULTEN (2003) insists on the demand for 'perceived credibility' formulated by SCHULTZE and MACKIE (2002, p. 7) with regard to the application of hedonic methods for building and estimating consumer price indices. 'Policy', he says, 'ultimately relies on the consent of the public, not the vision of convinced experts.' However, the more the academic world is convinced of hedonic methods as being the most promising to account for changes in product quality, the higher the acceptance by the public and the policymakers will be.

The present piece of work is a contribution to the academic discussion of hedonic methods in price statistics. In the first part, we constructed an axiomatic framework for hedonic elementary price indices and touched the fundamental question of what we want to measure with such indices. We then embedded well-known elementary index estimators into the proposed framework. Finally, we studied the problems arising when a hedonic elementary price index needs to be implemented in practice, and we tried to estimate how sensitive such an index may be with regard to sampling and specification errors.

WHELAN (2005, p. 66) recently argued that the consumer or retail price index in general as a 'method of measuring inflation adopted in developed economies is profligately over-engineered for the phenomenon it measures'. In this sense, it is certainly appropriate to discuss whether hedonic elemen-

tary price indices, which need a respectable effort to be properly defined and estimated, are really necessary in practice. WHELAN is probably right in saying that 'Merely because the index is important does not mean that it requires huge effort'. Users of an inflation measure, however, 'have a right to demand a quality product from the supplier and to define quality in their own terms' (HULTEN, 2003, p. 13). Statistical offices estimating and publishing a CPI are thus urged to work with 'best practice' methods.

Conclusions and recommendations

In the first part of the thesis, we identified two different approaches of defining a hedonic elementary price index for an elementary aggregate. Both of them still exhibit a large degree of flexibility, as they build upon a transformation function that is not intrinsically specified. The list of axioms developed in Chapter 4 provides guidelines that help to decide which of the two approaches and what kind of transformation function are preferable. However, the proposed system of axioms is not restrictive enough to allow for only one ideal definition of a hedonic elementary price index.

This has both advantages and disadvantages. From a theoretical point of view, it would certainly be preferable not having to choose among different definitions of a quality-adjusted inflation measure for an elementary aggregate of consumption. If one ideal index could be defined, there would be no ambiguity concerning the interpretation of such a measure. On the other hand, the flexibility offered by the theory also allows us to rely on more practical criteria in the choice of the parameter we wish to estimate. One such criterion might be the expected precision of the index estimates. If among the theoretically admissible index definitions, there is one where estimates are generally more robust against misspecifications of the hedonic functions or against sampling errors, it is just pragmatic to prefer this definition to the others.

In this sense, we come to the empirical conclusion that a hedonic elementary price index of an elementary aggregate $\mathcal{G}$ and for any given reference quality distribution $\mathbb{P}_{\boldsymbol{M}}$ should be defined by

$$HPI^{0:1}(\mathcal{G}) = \exp\left(\mathrm{E}\left[\ln\left(\frac{h^1_{\mathcal{G}}(\boldsymbol{M})}{h^0_{\mathcal{G}}(\boldsymbol{M})}\right)\right]\right) . \tag{10.1}$$

An estimator of this index is given by the Jevons formula

$$\hat{I}^{0:1}_{\mathrm{J}} = \sqrt[N]{\prod_{n=1}^{N} \frac{\hat{h}^1(\boldsymbol{M}_n)}{\hat{h}^0(\boldsymbol{M}_n)}} \tag{10.2}$$

based on estimates $\hat{h}^0$ and $\hat{h}^t$ of the hedonic functions and on an i.i.d. sample $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N$ of reference characteristics vectors.

The reasons for this choice are various. First of all, the index (10.1) is the only one that incorporates both of the two definition approaches mentioned above. Due to the properties of the exponential function and of the natural logarithm, (10.1)—which shows to be the backtransformed mean of transformed price ratios—can also be interpreted as the ratio of backtransformed means of transformed prices:

$$\exp\left(\mathrm{E}\left[\ln\left(\frac{h^1_{\mathcal{G}}(\boldsymbol{M})}{h^0_{\mathcal{G}}(\boldsymbol{M})}\right)\right]\right) = \frac{\exp(\mathrm{E}\ln(h^1_{\mathcal{G}}(\boldsymbol{M})))}{\exp(\mathrm{E}\ln(h^0_{\mathcal{G}}(\boldsymbol{M})))}\,.$$

Secondly, an important argument for using (10.1) as the preferred index definition is that it satisfies all of the axioms developed in Chapter 4. All the theoretical requirements on the index as a quality-adjusted inflation measure are therefore guaranteed.

Thirdly, with the Jevons formula (10.2), the index (10.1) has a natural estimator that enjoys a large support in the literature in its non-hedonic original form. It is currently the preferred base aggregator for unweighed price observations. From an empirical point of view, we could finally observe that index estimates generated by (10.2) were less sensitive to the choice of a specific functional form of the hedonic function than alternatives such as the Dutot and Harmonic Dutot estimators of other candidate index definitions. Moreover, compared to the alternatives, the Jevons index estimator showed a lower variance due to the sampling of prices and characteristics for both the estimations of the hedonic functions and of the reference quality distribution.

The choice of the reference sample among the items available in the base period, the current period, and both periods symmetrically taken together showed to have almost no influence on the estimates. We therefore recommend to use the symmetric reference sample, as it englobes the largest possible spectrum of items. In particular, it englobes both items that appeared and items that disappeared between the base and the current period. With higher-level price indices, practical reasons speak for using a Laspeyres-type approach, since expenditure shares of the current period are often not available early enough. This argument, however, does not hold in our context, as a sample of observations for the current period is in any case needed for estimating the hedonic function. Building a symmetric sample of items should therefore be straightforward.

Our investigations suggest that the estimation of the hedonic functions at different time periods is a question on its own that can and should be treated completely independently. A regression model needs to be identified,

which, depending on the nature of the elementary aggregate under review, best fits the assumed relationship between the price and the price-relevant characteristics. If no satisfactory fit can be found, this probably means that either the wrong characteristics are taken into account or the hedonic hypothesis does just not hold for this specific elementary aggregate.

Whether a relatively simple regression model is sufficient for modelling a hedonic function depends on the underlying elementary aggregate and the available observations. With our per-model simple semi-log approach building essentially on linear regression models, we showed that it was possible to approximate the hedonic function of used cars comparably well in terms of aggregate prediction error. In contrast, modern adaptive regression approaches might have the advantage that they fit an unknown functional dependence with no *a priori* theoretical background more easily and flexibly than an estimator that is manually built upon linear models. Moreover, they sometimes offer straightforward measures on the overall importance of individual exogenous variables. This information could be essential for deciding which characteristics should be surveyed in a productive estimation framework of a hedonic elementary price index.

If a statistical office seeks to introduce hedonic methods, in our opinion, the following steps are necessary:

1. Determine elementary aggregates where it is difficult to find items of equal quality in different time periods or to undertake traditional *ad hoc* quality adjustments. Hedonic elementary price indices as one possible approach to the problem of measuring 'pure' inflation are certainly most interesting for elementary aggregates where the quality spectrum in individual time periods as well as the change between periods is important.

2. Check the weight these elementary aggregates have in the overall CPI. The higher the individual weights are, the higher the effect of measurement errors in the corresponding elementary indices for higher-level indices is. Since hedonic methods generally demand more effort than traditional quality-adjustment methods, it is important to focus on elementary aggregates where the potential impact on the reduction of the overall measurement error is greater.

3. Start with a prospective study on any of the candidate elementary aggregates just identified. Use market surveys and expert judgements, for instance, to acquire insight into the range of quality differences observable on the market, the list of potentially price-relevant character-

istics of this elementary aggregate, the potential channels for surveying the necessary range of items, and so forth.

4. Build up a starting data set for exploratory purposes. This data should reflect the quality spectrum available on the market at a certain point of time. Moreover, it should include a full range of characteristics variables. Based on these observations, one would search for a regression model that best fits the characteristics-price relationship within the data. A fair technical knowledge of the elementary aggregate under review is certainly helpful for building a hedonic regression model. Once the model is specified, it is possible to analyse the importance of the individual exogenous variables. Based on the results of such an analysis, a reduced list of the most important variables that are sufficient for estimating the hedonic function can be compiled.

5. Survey regularly the identified characteristics and the price for a representative sample of the respective elementary aggregate. These samples are then used both as samples of reference quality as well as for estimating the hedonic function at fixed time intervals. Indices are finally estimated using (10.2), for instance.

While the above steps are sufficient for a regular estimation of hedonic elementary price indices, it seems judicious that an ongoing quality control accompany the publication of such estimates. It is recommended to monitor the goodness of fit of the estimated hedonic functions as well as the variance of the index estimates. Indicators for both of these aspects may be generated using, e.g., bootstrap methods, as we outlined in the empirical part of our research project.

In the Swiss CPI, where bilateral elementary indices are only calculated within a year and then chained over longer periods, hedonic methods could technically be introduced for any elementary aggregate within one year's time. Since the base period for the bilateral elementary indices within one year is always the month of December of the previous year (Bundesamt für Statistik, 2006), these indices could be replaced by hedonic counterparts each time the base period changes. Moreover, the Federal Statistical Office could reassess the actually employed hedonic regression model during the year in order to update it for the following year's estimations if necessary. Such an update, however, would not mean changing the index estimators but potentially adapting the learners of the hedonic functions, in order to obtain a better fit of the data.

It is important to note that hedonic methods, in general, do not provide a magic solution to any quality adjustment problem occurring within the estimation of an elementary price index. If, however, the relationship between quality and price can be modelled well enough using any regression approach and thus the hedonic hypothesis seems to hold, hedonic methods certainly provide a sophisticated tool for resolving the problem of quality change.

Open questions for further research

Today, the research on hedonic elementary price indices is far from being complete. Even during the elaboration of the present thesis, we had to leave many important and interesting questions aside, in order not to overload this specific research project. We will now address a few issues that are open for further investigations.

In the axiomatic framework of hedonic elementary price indices, it would be interesting to find and formulate other axioms that concentrate more on the characteristics than on the price variables of individual items of an elementary aggregate. How can the continuity of the index with regard to the characteristics random vectors be defined? Is it possible to narrow the potential field of index definitions in specifying further requirements? Is there a sensible axiom that would not be satisfied by our preferred index definition (10.1)? Or are there other axioms that further support the use of (10.1)?

There are as many theoretical models for hedonic functions as there are multivariate regression models. The open research questions in this context are thus the same as those related to any regression problem, and none of them is particularly different for hedonic regressions. The main aim here is to predict the price of an item from its characteristics. Any method that provides an answer to this problem is a candidate estimator of the hedonic function.

Research questions related to estimators of hedonic indices are, again, more specific. Is it possible, for instance, to find theoretical results on the distribution of such index estimators? Under what conditions does the estimator (10.2) of (10.1) have a lower variance than another estimator of any other admissible index definition? What other particular properties does (10.2) have? What about double or single imputation formulae—which approach is theoretically more satisfactory?

More general investigations could tend towards the application of hedonic methods for interregional instead of intertemporal price comparisons. Are price differences of certain goods between different countries due to quality

differences of the product or to other economic factors? Are used cars of the same quality more expensive in Zurich than in Fribourg?

The list of open research questions is long and could still be extended. We hope, therefore, that the results presented in this thesis are going to stimulate further research, especially in the area of the axiomatic foundations of hedonic elementary price indices. Moreover, we hope that the use of hedonic methods for estimating quality-adjusted elementary price indices will become increasingly standard in statistical offices for elementary aggregates where these methods are appropriate.

Appendix

APPENDIX A

Tables and Figures

Table A.1: **Terms included in the simple semi-log models**

	Oct 2004	Nov 2004	Dec 2004	Jan 2005	Feb 2005	Mar 2005	Apr 2005	May 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005	Oct 2005	Nov 2005	Dec 2005	Jan 2006	Feb 2006	Mar 2006
make	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
body	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
bodycol	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
bodycol_met		•	•		•	•		•	•	•	•	•	•	•	•	•	•	•
fuel	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
gear	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
doors	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
seats	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
cylinders	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
ccm	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
hp	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
age	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
mileage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
warranty	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
accident	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
X4x4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
X8fach.bereift		•		•			•	•	•	•							•	•
ABS				•	•		•				•		•	•	•	•	•	•
Airbag		•		•	•		•	•	•	•	•	•	•	•	•	•	•	•
Airbag.Beifahrer	•	•	•	•	•	•	•	•	•	•					•	•	•	•
Alarmanlage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Alufelgen	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Anh.ngerkupplung							•	•								•		
Elektr..Sitze	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Elektr.Fensterheber	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Elektr.Parkhilfe	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Elektr.Schiebedach	•		•	•	•	•	•	•		•	•	•		•	•			
Faltdach	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•
Hardtop	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Klima	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Klima.Automatik	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Ledersitze	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Navigationssystem	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Nebelscheinwerfer	•	•	•	•	•	•	•	•	•	•	•	•	•	•				
Partikelfilter							•	•	•	•	•	•	•	•	•	•	•	•
Radio.CD	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Radio.Tonband	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Schiebedach		•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Servo	•	•			•	•	•	•		•		•	•	•	•	•	•	•
Sidebags	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Sitzheizung			•	•		•	•	•	•	•	•	•	•		•	•	•	•
Sonnendach	•		•	•	•	•	•	•			•			•				•
Sperrdiff.	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Sportsitze		•	•		•		•	•	•	•	•	•						
Standheizung	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•
Tempomat	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Wegfahrsperre	•	•	•	•	•	•	•	•	•	•	•		•		•	•	•	•
Xenon.Scheinwerfer	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Zentralverriegelung	•			•	•	•		•	•	•	•	•	•	•	•	•	•	•
age:mileage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Table A.2: **Terms included in the enhanced semi-log models**

	Oct 2004	Nov 2004	Dec 2004	Jan 2005	Feb 2005	Mar 2005	Apr 2005	May 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005	Oct 2005	Nov 2005	Dec 2005	Jan 2006	Feb 2006	Mar 2006
make	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
body	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
bodycol	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
bodycol_met	| ●	●| ●	●| ●	| ●	●|	●|	| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
fuel	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
gear	●|	●| ●	●|	●| ●	●| ●	●|	●|	●| ●	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
doors	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
seats	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
cylinders	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
ccm	●| ●	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
hp	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
age	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
mileage	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
warranty	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
accident	●| ●	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●|	●| ●	●|	●| ●	●| ●
X4x4	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
X8fach.bereift	|	|	|	●|	|	| ●	●| ●	●|	●|	●|	|	|	| ●	|	|	●|	●|
ABS	| ●	| ●	| ●	●| ●	●| ●	| ●	●| ●	| ●	| ●	| ●	|	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
Airbag	| ●	●| ●	| ●	●| ●	●| ●	| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
Airbag.Beifahrer	●| ●	●| ●	●|	●| ●	●| ●	●| ●	●|	●|	●| ●	●| ●	| ●	|	|	●| ●	●| ●	●| ●	●| ●
Alarmanlage	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●|	●| ●	●|	●|	●| ●	●| ●	●|	●| ●	●| ●	●| ●	●| ●
Alufelgen	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●	●| ●
Anh.ngerkupplung	|	|	|	|	| ●	| ●	●|	●| ●	|	|	|	|	| ●	|	●| ●	| ●	| ●
Elektr..Sitze	●|	●| ●	●|	●|	●| ●	●|	●|	●|	●| ●	●| ●	●| ●	●| ●	●| ●	●|	●|	●|	●|	●|

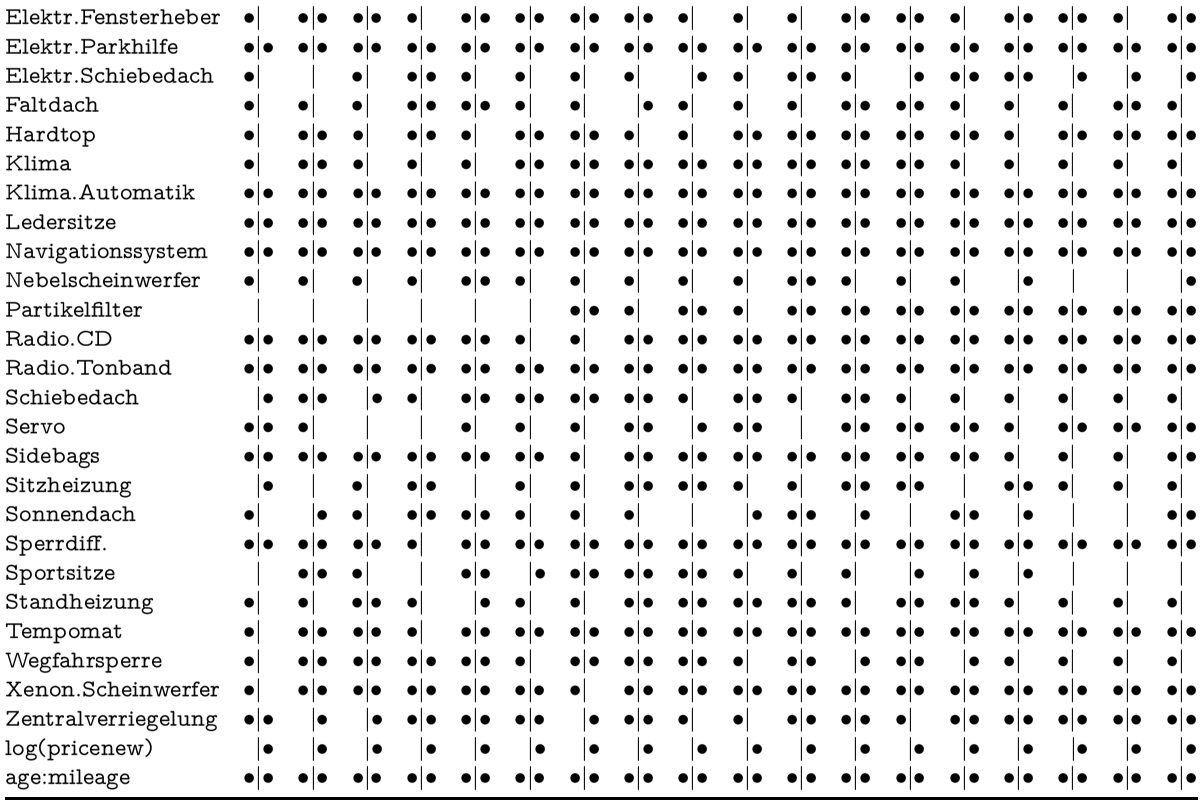

Elektr.Fensterheber
Elektr.Parkhilfe
Elektr.Schiebedach
Faltdach
Hardtop
Klima
Klima.Automatik
Ledersitze
Navigationssystem
Nebelscheinwerfer
Partikelfilter
Radio.CD
Radio.Tonband
Schiebedach
Servo
Sidebags
Sitzheizung
Sonnendach
Sperrdiff.
Sportsitze
Standheizung
Tempomat
Wegfahrsperre
Xenon.Scheinwerfer
Zentralverriegelung
log(pricenew)
age:mileage

Table A.3: **Terms included in the simple double-log models**

	Oct 2004	Nov 2004	Dec 2004	Jan 2005	Feb 2005	Mar 2005	Apr 2005	May 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005	Oct 2005	Nov 2005	Dec 2005	Jan 2006	Feb 2006	Mar 2006
make	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
body	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
bodycol	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
bodycol_met		•			•	•		•	•	•	•	•	•	•	•	•	•	•
fuel	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
gear		•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(doors)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(seats)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(cylinders)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(ccm)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(hp)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(age)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(mileage)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
warranty	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
accident	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
X4x4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
X8fach.bereift	•		•		•	•	•	•	•	•	•	•	•	•	•	•	•	•
ABS	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Airbag	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Airbag.Beifahrer	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Alarmanlage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Alufelgen	•	•	•			•	•						•		•			•
Anh.ngerkupplung		•	•	•			•	•	•				•	•		•		
Elektr..Sitze	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Elektr.Fensterheber	•	•	•	•		•		•			•	•	•	•	•	•	•	•	
Elektr.Parkhilfe	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Elektr.Schiebedach			•	•	•	•	•	•		•	•	•	•	•	•				
Faltdach	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	
Hardtop	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Klima	•	•	•				•	•	•	•	•	•	•	•	•	•	•	•	
Klima.Automatik	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Ledersitze	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Navigationssystem	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Nebelscheinwerfer	•	•	•	•	•	•		•		•	•	•	•	•	•	•	•	•	
Partikelfilter							•			•	•	•	•	•		•	•	•	
Radio.CD	•	•	•	•	•	•	•	•	•	•	•	•	•	•				•	
Radio.Tonband	•	•	•	•	•	•			•		•	•	•	•	•	•	•	•	
Schiebedach	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Servo	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Sidebags	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Sitzheizung		•	•	•			•	•		•	•				•	•	•	•	
Sonnendach				•	•	•	•	•		•	•	•	•	•	•	•	•	•	
Sperrdiff.	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Sportsitze			•				•	•	•	•	•	•							
Standheizung	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Tempomat	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Wegfahrsperre	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Xenon.Scheinwerfer	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Zentralverriegelung	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
log(age):log(mileage)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	

Table A.4: **Terms included in the flexible double-log models**

	Oct 2004	Nov 2004	Dec 2004	Jan 2005	Feb 2005	Mar 2005	Apr 2005	May 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005	Oct 2005	Nov 2005	Dec 2005	Jan 2006	Feb 2006	Mar 2006
make	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
body	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
bodycol	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
bodycol_met		•	•			•		•	•	•	•	•	•	•	•	•	•	•
fuel	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
gear	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(doors)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
doors	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(seats)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
seats	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(cylinders)		•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	
cylinders	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(ccm)	•	•	•	•	•	•	•	•	•	•					•		•	
ccm	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(hp)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
hp	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(age)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
log(mileage)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
age	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
mileage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
warranty	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
accident	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
X4x4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
X8fach.bereift				•				•	•	•	•	•	•					•

ABS	•	•	•	•		•	•	•	•	•	•	•	•	•					
Airbag		•				•		•			•	•	•	•	•	•	•	•	
Airbag.Beifahrer		•	•	•	•	•	•	•		•					•	•	•	•	
Alarmanlage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Alufelgen	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Anh.ngerkupplung			•				•	•	•				•						
Elektr..Sitze	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Elektr.Fensterheber	•	•	•	•		•	•	•	•	•	•				•	•	•	•	
Elektr.Parkhilfe	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Elektr.Schiebedach				•	•	•	•	•		•	•	•	•	•	•	•	•		
Faltdach					•	•	•	•		•		•		•			•		
Hardtop	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Klima	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•		
Klima.Automatik	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Ledersitze	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Navigationssystem	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Nebelscheinwerfer			•	•	•	•	•	•	•						•	•		•	
Partikelfilter							•	•	•	•	•	•	•	•	•	•	•	•	
Radio.CD	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Radio.Tonband	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Schiebedach		•		•	•	•	•	•	•	•	•	•	•	•	•	•	•		
Servo	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Sidebags	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Sitzheizung	•		•	•			•	•			•				•	•	•	•	
Sonnendach			•	•	•	•	•	•			•	•	•	•	•	•	•		
Sperrdiff.	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Sportsitze		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Standheizung	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	
Tempomat	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	

Table A.4: **(cont.)**

	Oct 2004	Nov 2004	Dec 2004	Jan 2005	Feb 2005	Mar 2005	Apr 2005	May 2005	Jun 2005	Jul 2005	Aug 2005	Sep 2005	Oct 2005	Nov 2005	Dec 2005	Jan 2006	Feb 2006	Mar 2006
Wegfahrsperre	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Xenon.Scheinwerfer	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Zentralverriegelung	•			•	•	•		•	•	•	•	•	•	•	•	•	•	•
log(age):log(mileage)		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
age:mileage	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

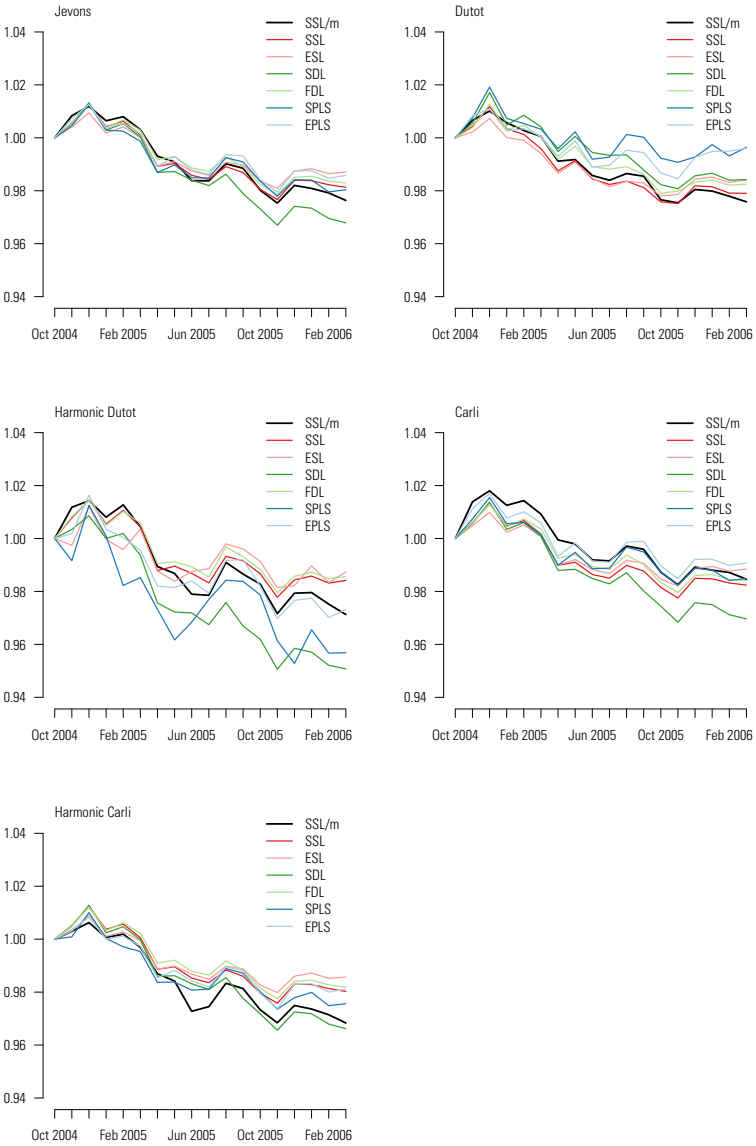


Figure A.1: **Bilateral hedonic elementary price index estimates using the base period reference sample. Comparison of alternative index formulae and functional forms of the hedonic functions**

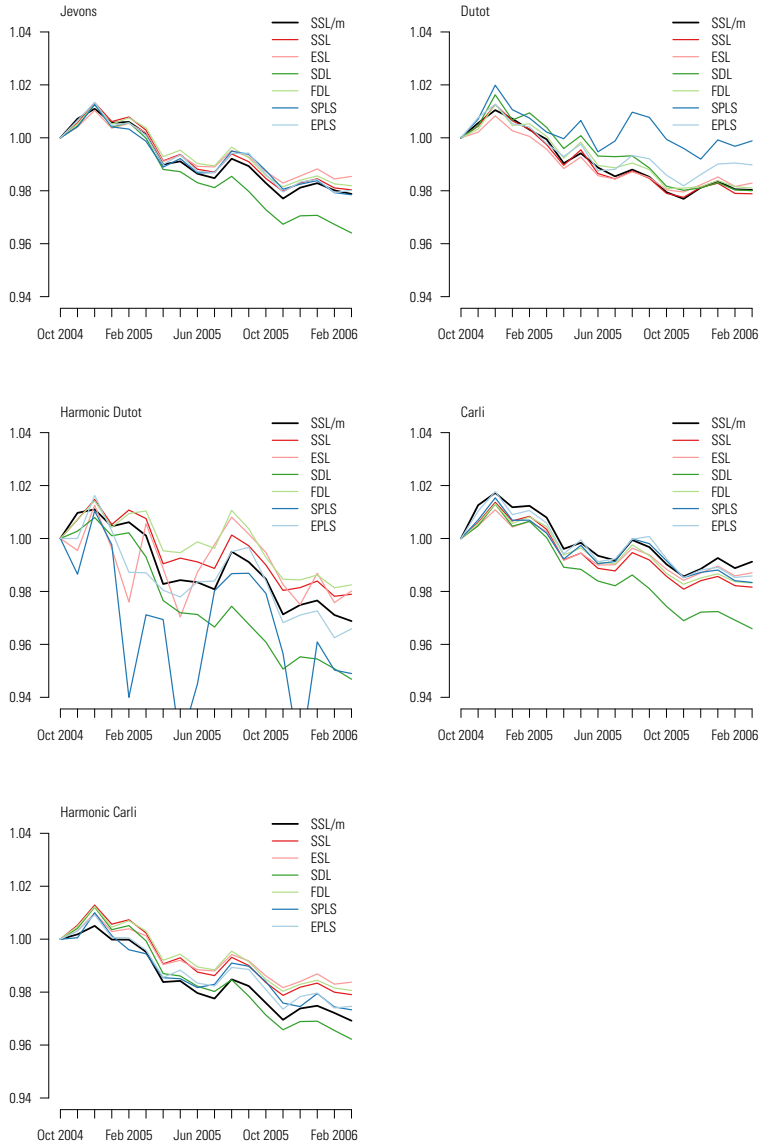


Figure A.2: **Bilateral hedonic elementary price index estimates using the current period reference sample. Comparison of alternative index formulae and functional forms of the hedonic functions**

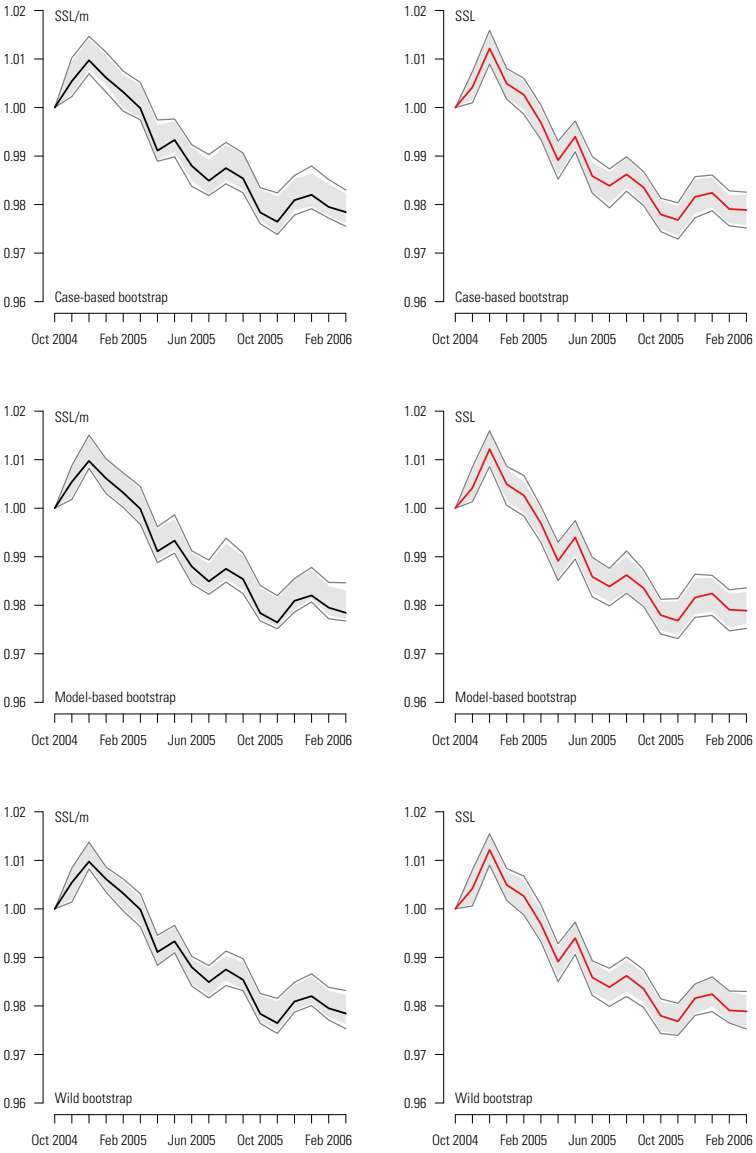


Figure A.3: **Dutot-type bilateral hedonic elementary price index estimates for the symmetric reference samples. Comparison of different functional forms of the hedonic function and of different bootstrap approaches.**

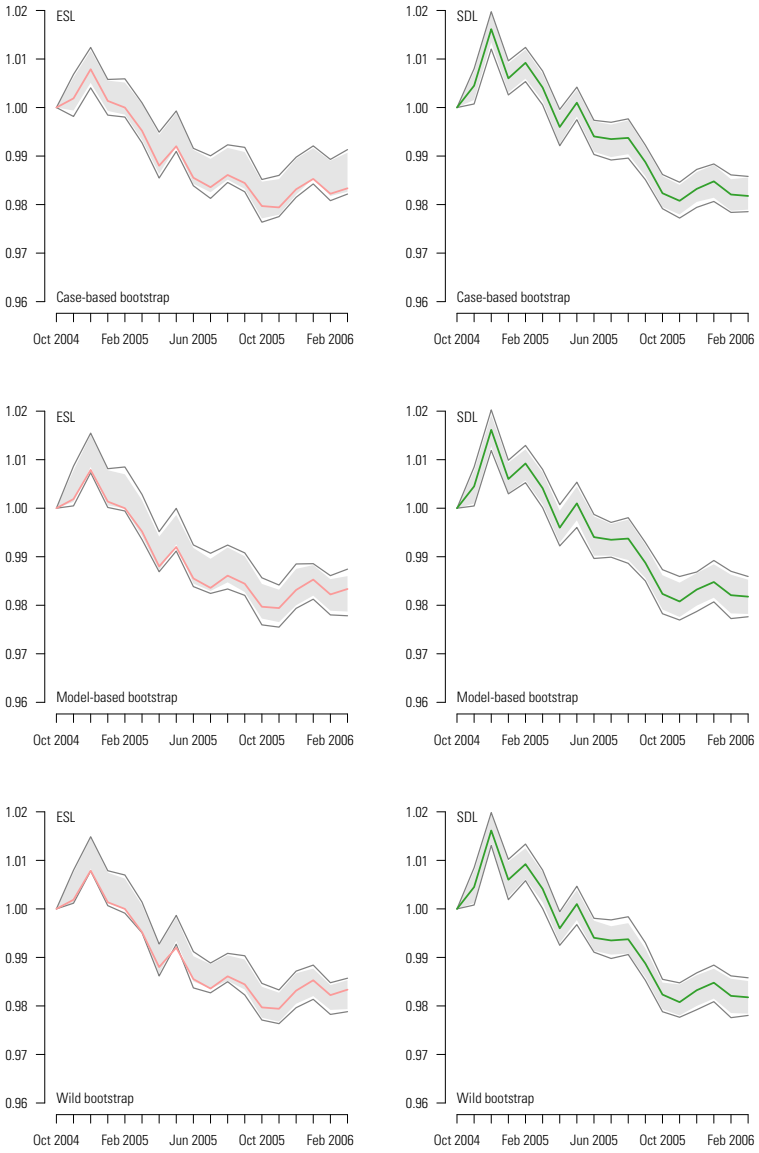


Figure A.3: **(cont.)**

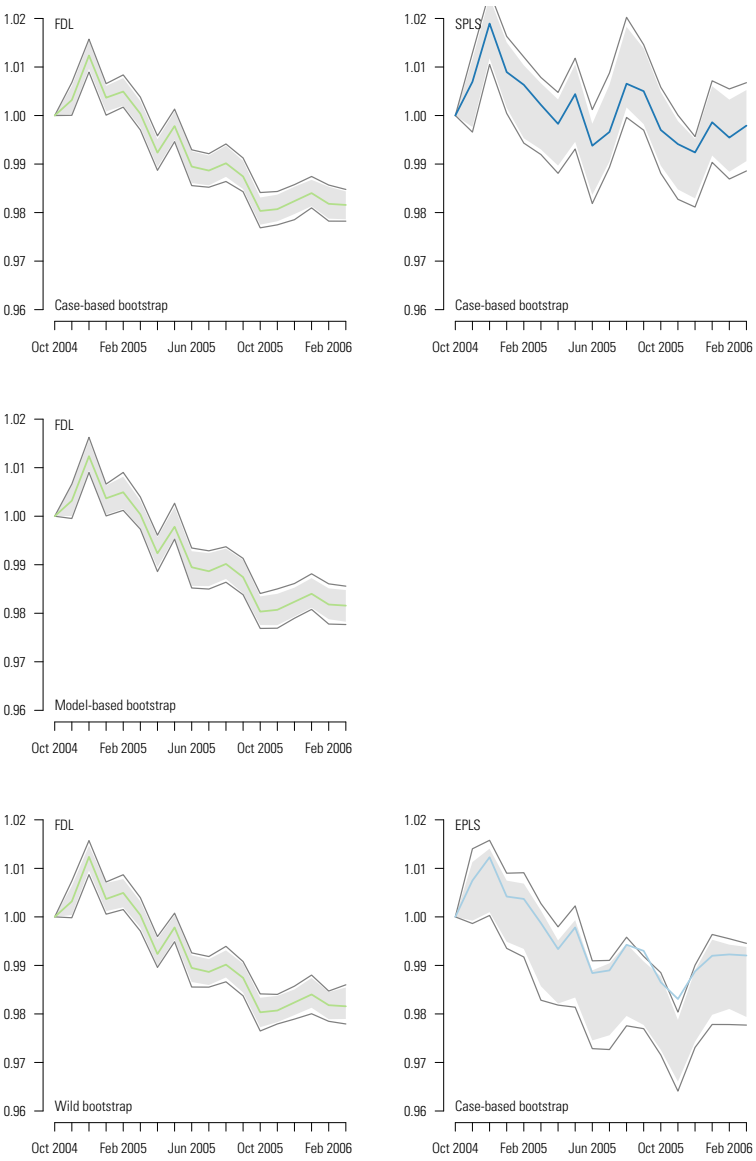


Figure A.3: **(cont.)**

Table A.5: **Lengths of 90% and 95% bootstrap confidence intervals (multiplied by 100) over the period from Nov 2004 to Mar 2006 for Dutot-type index estimates. Comparison of different modelling and bootstrap approaches.**

Model	Bootstrap algorithm	Interval lengths		Mean length difference	
		90%	95%	90%	95%
SSL	case-based	0.56–0.71	0.63–0.85	0.007	−0.001
	model-based	0.59–0.74	0.71–0.89	0.007	0.010
	wild	0.57–0.69	0.65–0.82	0.007	0.015
ESL	case-based	0.62–0.81	0.73–0.94	0.011	0.007
	model-based	0.60–0.75	0.73–0.97	0.018	0.039
	wild	0.48–0.64	0.58–0.81	0.030	0.020
SDL	case-based	0.55–0.70	0.67–0.81	0.010	0.006
	model-based	0.62–0.83	0.69–0.97	0.001	−0.000
	wild	0.55–0.73	0.67–0.86	0.005	0.010
FDL	case-based	0.53–0.66	0.65–0.77	0.003	0.006
	model-based	0.56–0.71	0.66–0.83	0.006	0.011
	wild	0.51–0.65	0.59–0.81	−0.001	0.004
SPLS	case-based	1.22–1.66	1.45–2.06	0.005	−0.006
EPLS	case-based	1.21–1.59	1.49–2.09	0.022	0.011
SSL/m	case-based	0.60–0.75	0.74–0.88	0.018	0.019
	model-based	0.57–0.72	0.68–0.91	0.018	0.014
	wild	0.43–0.61	0.52–0.79	0.023	0.031

APPENDIX B

The R package 'hepi'

The real workhorse in the empirical part of our research on hedonic elementary price indices were the algorithms and objects developed in R (R Development Core Team, 2006). Many parts of the code were closely related to the specific structure of our data and therefore not directly applicable or even useless in other contexts. However, the elements of the code related to the representation of hedonic functions and to the estimation of hedonic elementary price indices as well as some of the essential algorithms for estimating the hedonic functions could be assembled in a R package, which is independent of the data set. This package has been made publicly available under the GNU General Public License (Free Software Foundation, 1991). It may be downloaded from

`http://www.michael.beer.name/r-hepi`

or requested directly from the author.

On the following pages, the original R-style documentation of the package in its current state as well as of the functions and classes included is reproduced. For the sake of brevity, some non-informative or repeated sections such as the author name and the references have been omitted from the listing here below. They are, however, contained in the original documentation accompanying the package.

hepi-package *Functions for Estimating Hedonic Elementary Price Indices*

Description

This package provides a set of functions for estimating hedonic functions and hedonic elementary price indices.

Details

Package:	hepi
Date:	2006-09-14
Version:	1.0-0
License:	GPL (version 2 or later) See file LICENCE.
Depends:	R (>= 2.2.0)
Suggests:	MASS, spdep
LazyLoad:	yes
SaveImage:	yes
URL:	http://www.michael.beer.name/r-hepi

Index:

hedonic.function-class Class "hedonic.function"

hedonic.function Constructor for a "hedonic.function" Object

analyse.in.hf Evaluate Expressions Within the Environment of a Hedonic Function

is.applicable.hf Test Applicability of Hedonic Function to Data

build.hf.lm Hedonic Function Based on a Linear Model

build.hf.lm.split Hedonic Function Based on a List of Linear Models

hepi Bilateral Hedonic Elementary Price Indices

hedonic.function-class *Class "hedonic.function"*

Description

Hedonic functions for the use in price statistics.

A hedonic function predicts a price value for an item based on a list of characteristics expressions.

Slots

.Data: An object of class "function" having at least the named argument data representing the data frame for which new prices are to be estimated.

characteristics.names: A vector containing the names of all variables that must be available in data in order to provide valid price predictions.

call: An object of class "call", describing how .Data was created.

description: A textual description of the hedonic function .Data.

Extends

Class "function", from data part. Class "OptionalFunction", by class "function". Class "PossibleMethod", by class "function".

Methods

No methods defined with class "hedonic.function" in the signature.

See Also

hedonic.function

---

hedonic.function *Constructor for a "hedonic.function" Object*

Description

This function constructs objects of class "hedonic.function".

Usage

```
hedonic.function(hf, characteristics.names, env = NULL,
                 call = match.call(), description = "")
```

hf	The hedonic function.
characteristics.names	A vector of characteristics names.
env	An optional list of objects needed internally for the evaluation of hf.
call	An object of class "call", describing how hf was created.
description	A textual description of the hedonic function hf.

Details

A hedonic function predicts a price value for an item based on a list of characteristics expressions. hf is thus an object of class "function" having at least the named argument data representing the data frame for which new prices are to be estimated.

characteristics.names should contain the names of all variables that must be available in data in order to provide valid price predictions. It can be used to check whether the returned hedonic function is applicable for a given data set.

Value

An object of class "hedonic.function".

See Also

hedonic.function-class

Examples

```
## Build hedonic function from training data set
build.hf.loglin <- function(traindata) {
   model <- lm(log(price) ~ ., traindata)

   hf <- function(data) {
      predict(model, newdata = data)
   }

   hedonic.function(
      hf = hf,
      characteristics.names = all.vars(formula(model)),
      env = list(model = model),
```

```
        call = match.call(),
        description = "Semi-logarithmic hedonic function."
    )
}
```

---

analyse.in.hf — *Evaluate Expressions Within the Environment of a Hedonic Function*

Description

This function evaluates R expressions within the environment of an object of class "hedonic.function".

Usage

```
analyse.in.hf(expr, hf)
```

Arguments

expr	the expression to be evaluated
hf	an object of class "hedonic.function"

Details

Hedonic functions are often predictions from a regression model. This function allows to easily access the elements in the environment of the hedonic function for further analysis.

See Also

hedonic.function-class, hedonic.function

Examples

```
data(boston, package="spdep")

hf0 <- build.hf.lm(
   learndata = boston.c,
   full.formula = log(MEDV) ~ CRIM + ZN + INDUS + CHAS +
     I(NOX^2) + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +
     PTRATIO + B + log(LSTAT),
```

```
    backtrans = exp,
    rm.infl = FALSE,
    description = NULL,
    return.row.labels = FALSE,
    allow.variable.selection = FALSE)

analyse.in.hf(ls(), hf0)
analyse.in.hf(summary(learndata), hf0)
analyse.in.hf(summary(hf.model), hf0)
```

---

is.applicable.hf *Test Applicability of Hedonic Function to Data*

Description

This function tests whether a "hedonic.function" object is applicable to a given data set.

Usage

```
is.applicable.hf(hf, data)
```

Arguments

hf	an object of class "hedonic.function"
data	a named list or data frame

Details

Hedonic functions are usually estimated for a very specific data structure and cannot be applied to other situations. This function checks whether the given hedonic function hf is applicable to a certain data set data. This is done by checking whether data contains at least the variables mentioned in the slot characteristics.names of the object hf of class "hedonic.function".

Value

is.applicable.hf returns TRUE if data is suitable for hf and FALSE otherwise.

See Also

hedonic.function-class, hedonic.function

Description

This function estimates hedonic functions based on a linear regression model for a given dataset.

Usage

```
build.hf.lm(learndata, full.formula, min.formula,
        backtrans = I, rm.infl = TRUE,
        description = NULL, return.row.labels = FALSE,
        allow.variable.selection = TRUE)
```

Arguments

learndata	A data.frame containing the training data set.
full.formula	The formula of the full linear model. See Details.
min.formula	If variable selection is wanted, the formula of the minimal linear model. See Details.
backtrans	A backtransformation function applied to all predictions. See Details.
rm.infl	A logical value indicating whether influential observations should be removed.
description	A character string describing the hedonic function.
return.row.labels	A logical value indicating whether the row labels of the cleaned training data should be returned.
allow.variable.selection	A logical value indicating whether variable selection should be carried out.

Details

This function estimates a hedonic function based on a linear regression model. An appropriate model formula must be given in full.formula. (See lm for more details about specifying formulae.)

The function given in backtrans is used to backtransform any predicted value using the linear model and defauls to the identity function I. If, for example, log(price) stands on the left-hand side of the model formula, any predicted

value needs to be transformed with the exponential function to a valid price. This can be accomplished by indicating `backtrans = exp`.

If `rm.infl` is `TRUE`, influential observations having

$$\text{DFFITS}_i > 2\sqrt{\frac{K}{N}}$$

with K being the number of exogenous variables and N the dimension of the learning data set are removed before fitting the final model. There, the DFFITS_i values are calculated based on the residuals of a first fit of a linear model using the model formula `full.formula`.

If `allow.variable.selection` is `TRUE`, a stepwise model selection based on exact AIC is carried out (see `stepAIC` for more details). In this case, `full.formula` acts as upper and `min.formula` as lower limit of the search algorithm. If `allow.variable.selection` is `FALSE` the hedonic function is estimated using exactly the formula given in `full.formula`.

In `description`, a character string describing the hedonic function may be given which is saved within the returned `"hedonic.function"` object.

Value

If `return.row.labels == FALSE`, the function returns a `"hedonic.function"` object representing the fitted regression model.

If `return.row.labels == TRUE`, the function returns a list with following elements:

`hf`	The resulting `"hedonic.function"` object.
`row.labels`	A vector containing the row labels of the cleaned training data set.

See Also

`build.hf.lm.split`

Examples

```
data(boston, package = "spdep")

hf0 <- build.hf.lm(
   learndata = boston.c,
   full.formula = log(MEDV) ~ CRIM + ZN + INDUS + CHAS +
     I(NOX^2) + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +
     PTRATIO + B + log(LSTAT),
   backtrans = exp,
```

```
    rm.infl =  FALSE,
    description = NULL,
    return.row.labels = FALSE,
    allow.variable.selection = FALSE)

is.applicable.hf(hf0, boston.c)
summary(hf0(boston.c))

plot(boston.c$MEDV, hf0(boston.c), xlab = "Observed", ylab = "Predicted")
abline(0,1)

hf1 <- build.hf.lm(
    learndata = boston.c,
    full.formula = log(MEDV) ~ CRIM + ZN + INDUS + CHAS +
      I(NOX^2) + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +
      PTRATIO + B + log(LSTAT),
    min.formula = log(MEDV) ~ 1,
    backtrans = exp,
    rm.infl =  FALSE,
    description = NULL,
    return.row.labels = FALSE,
    allow.variable.selection = TRUE)
summary(hf1(boston.c))
```

---

build.hf.lm.split — *Hedonic Function Based on a List of Linear Models*

Description

This function estimates hedonic functions based on linear regression models for a given dataset. Individual sub-models are fit to subsets of the data split up according to a factor variable.

Usage

```
build.hf.lm.split(learndata, split.var, full.formula, min.formula,
    backtrans = I, rm.infl = FALSE, description = NULL,
    return.row.labels = FALSE, allow.variable.selection = TRUE,
    use.overall.hf = TRUE, split.threshold = 100)
```

Arguments

learndata	A data.frame containing the training data set.
split.var	The name of the factor variable used to split the data into subsets. See Details.
full.formula	The formula of the full linear model. See Details.
min.formula	If variable selection is wanted, the formula of the minimal linear model. See Details.
backtrans	A backtransformation function applied to all predictions. See Details.
rm.infl	A logical value indicating whether influential observations should be removed.
description	A character string describing the hedonic function.
return.row.labels	A logical value indicating whether the row labels of the cleaned training data should be returned.
allow.variable.selection	A logical value indicating whether variable selection should be carried out.
use.overall.hf	A logical value indicating whether an overall model should be fit to the whole data set.
split.threshold	The minimal number of observations required for fitting any sub-model.

Details

This function estimates a hedonic function based on linear regression models. In contrast to build.hf.lm, however, individual linear models are fit to several subsets of the data. These subsets are determined through a factor variable named split.var which needs to be contained in learndata. The minimal size a subset needs to have in order to fit a linear model is given by split.threshold.

If use.overall.hf is TRUE, an overall model for the whole data set is fit and stored additionally in order to predict prices for characteristics vectors beloning to categories of split.var where less than split.threshold observations are available in the learning data set.

See the documentation of build.hf.lm for an explanation of the other arguments of the function. Removal of influential observations and variable selection, if required, is carried out for each sub-model individually.

Value

If return.row.labels == FALSE, the function returns a "hedonic.function" object representing the fitted regression model.

If return.row.labels == TRUE, the function returns a list with following elements:

hf	The resulting "hedonic.function" object.
row.labels	A vector containing the row labels of the cleaned training data set.

See Also

build.hf.lm

Examples

```
data(boston, package = "spdep")

hf0 <- build.hf.lm.split(
   learndata = boston.c,
   split.threshold = 15,
   split.var = "TOWN",
   full.formula = log(MEDV) ~ CRIM + ZN + INDUS + CHAS +
     I(NOX^2) + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +
     PTRATIO + B + log(LSTAT),
   backtrans = exp,
   rm.infl = FALSE,
   description = NULL,
   return.row.labels = FALSE,
   allow.variable.selection = FALSE)

is.applicable.hf(hf0, boston.c)
summary(hf0(boston.c))

plot(boston.c$MEDV, hf0(boston.c), xlab = "Observed", ylab = "Predicted")
abline(0,1)

hf1 <- build.hf.lm.split(
   learndata = boston.c,
   split.var = "TOWN",
   split.threshold = 15,
   full.formula = log(MEDV) ~ CRIM + ZN + INDUS + CHAS +
     I(NOX^2) + I(RM^2) + AGE + log(DIS) + log(RAD) + TAX +
     PTRATIO + B + log(LSTAT),
   min.formula = log(MEDV) ~ 1,
   backtrans = exp,
   rm.infl = FALSE,
```

```
    description = NULL,
    return.row.labels = FALSE,
    allow.variable.selection = TRUE)
summary(hf1(boston.c))
```

---

hepi *Bilateral Hedonic Elementary Price Indices*

Description

This function estimates bilateral hedonic elementary price indices based on estimations of the hedonic function in the base and the current period as well as a reference sample of quality characteristics.

Usage

```
hepi(hf0, hf1, M,
    type = c("jevons", "dutot", "carli", "hdutot", "hcarli"),
    na.rm = TRUE, debug = FALSE)

hepi.jevons(hf0, hf1, M)
hepi.carli(hf0, hf1, M)
hepi.dutot(hf0, hf1, M)
hepi.hcarli(hf0, hf1, M)
hepi.hdutot(hf0, hf1, M)
```

Arguments

hf0	The base period hedonic function estimate. Must be of class "hedonic.function".
hf1	The current period hedonic function estimate. Must be of class "hedonic.function".
M	The reference sample.
type	The type of the index estimator(s) to be used. Can be a vector if estimates of several types are requested.
na.rm	A logical value indicating whether observations containing NA values should be stripped before the computation proceeds.
debug	A logical value indicating whether predicted prices should be returned for debugging purposes.

This function yields an estimate of a bilateral hedonic elementary price index. Inputs are the two estimated hedonic functions `hf0` and `hf1` of the base and current period respectively. Both of these must be of class `"hedonic.function"`. (See `hedonic.function` for a constructor of a `hedonic.function` object.)

The third input is the reference sample `M` to be used for the estimation of the index. This is usually a data frame containing N characteristics vectors to which both hedonic functions are applicable.

The `type` argument lets one choose the index formula to be used (and yet the index to be estimated). Currently, we implemented five alternative estimators, namely the

$$\sqrt[n]{\prod_{i=1}^{N} \frac{\hat{h}^1(m_n)}{\hat{h}^0(m_n)}} \qquad \text{(Jevons)},$$

$$\frac{\sum_{n=1}^{N} \hat{h}^1(m_n)}{\sum_{n=1}^{N} \hat{h}^0(m_n)} \qquad \text{(Dutot)},$$

$$\frac{1}{N} \sum_{n=1}^{N} \frac{\hat{h}^1(m_n)}{\hat{h}^0(m_n)} \qquad \text{(Carli)},$$

$$\frac{\left(\sum_{n=1}^{N} \left(\hat{h}^1(m_n)\right)^{-1}\right)^{-1}}{\left(\sum_{n=1}^{N} \left(\hat{h}^0(m_n)\right)^{-1}\right)^{-1}} \qquad \text{(Harmonic Dutot) and}$$

$$\left(\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{h}^1(m_n)}{\hat{h}^0(m_n)}\right)^{-1}\right)^{-1} \qquad \text{(Harmonic Carli)}$$

formulae. Details can be found in Chapter 6 of the reference mentioned below.

Value

If `debug == FALSE`, this function returns a vector with the same length as `type` containing the index estimates. They are returned in the same order as given by `type`.

If `debug == TRUE`, this function returns a list with the following entries

`index`	The vector of index estimates as above.
`p0hat`	The vector of predicted prices $\hat{\boldsymbol{p}}^0 = \hat{h}^0(\boldsymbol{M})$ in the base period.
`p1hat`	The vector of predicted prices $\hat{\boldsymbol{p}}^1 = \hat{h}^1(\boldsymbol{M})$ in the current period.
`ratios`	The vector of price ratios $\hat{p}_n^1/\hat{p}_n^0$ $(n = 1, \ldots, N)$.

APPENDIX C

Bibliography

Abdi, H. (2004) Partial Least Squares Regression. *In* M. S. Lewis-Beck, A. Bryman, and T. F. Liao (eds.), *The SAGE encyclopedia of social science research methods*, vol. 2, pp. 792–795. SAGE Publications, Thousand Oaks. [56]

Ahnert, H. and Kenny, G. (2004) Quality adjustment of European price statistics and the role for hedonics. Occasional Paper Series 15, European Central Bank. URL `http://ideas.repec.org/p/ecb/ecbops/20040015.html`, accessed: 24.01.2006. [78]

Andersson, D. E. (2000) Hypothesis testing in hedonic price estimation – On the selection of independent variables. *The Annals of Regional Science* **34**(2): 293–304. doi:10.1007/s001689900010. [59]

Anglin, P. M. and Gençay, R. (1996) Semiparametric Estimation of a Hedonic Price Function. *Journal of Applied Econometrics* **11**(6): 633–648. [55]

Arguea, N. M. and Hsiao, C. (1993) Econometric issues of estimating hedonic price functions. *Journal of Econometrics* **56**: 243–267. [59, 78]

Balk, B. M. (1995) Axiomatic Price Index Theory: A Survey. *International Statistical Review* **63**(1): 69–93. ISSN 0306-7734. [34]

Balk, B. M. (2005) Price Indexes for Elementary Aggregates: The Sampling Approach. *Journal of Official Statistics* **21**(4): 675–699. [26]

Beer, M. (2007) Bootstrapping a Hedonic Price Index: Experience from Used Cars Data. *AStA Advances in Statistical Analysis* doi:10.1007/s10182-006-0015-9. [109, 137]

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York. ISBN 0-471-05856-4. [101]

Benkard, C. L. and Bajari, P. (2003) Hedonic Price Indexes with Unobserved Product Characteristics, and Application to PC's. Working Paper 9980, National Bureau of Economic Research, Inc. URL `http://www.nber.org/papers/w9980.pdf`, accessed: 17.05.2004. [59]

Berndt, E. R. (1991) *The Practice of Econometrics: Classic and Contemporary*, chap. 4, pp. 102–149. Addison-Wesley Publishing Company, Reading. [54]

Billor, N., Hadi, A. S., and Velleman, P. F. (2000) BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis* **34**(3): 279–298. doi:10.1016/S0167-9473(99)00101-2. [109]

Bitros, G. C. and Panas, E. E. (1988) Measuring Product Prices Under Conditions of Quality Change: The Case of Passenger Cars in Greece. *The Journal of Industrial Economics* **37**(2): 167–186. [55, 78]

Boskin, M. J., Dulberger, E. R., Gordon, R. J., Griliches, Z., and Jorgenson, D. W. (1996) Toward A More Accurate Measure Of The Cost Of Living. Final report to the Senate Finance Committee, Advisory Commission To Study The Consumer Price Index. URL `http://www.ssa.gov/history/reports/boskinrpt.html`, accessed: 23.03.2005. [78]

Boskin, M. J., Dulberger, E. R., Gordon, R. J., Griliches, Z., and Jorgenson, D. W. (1998) Consumer Prices, the Consumer Price Index, and the Cost of Living. *Journal of Economic Perspectives* **12**(1): 3–26. [2]

Boulding, W. and Purohit, D. (1996) The Price of Safety. *Journal of Consumer Research* **23**: 12–25. [78]

Brachinger, H. W. (1990) Identifikation einflussreicher Daten: Ein Überblick über die Regression Diagnostics (Teil I). *Allgemeines Statistisches Archiv* **74**(2): 188–212. ISSN 0002-6018. [101, 103]

Brachinger, H. W. (2002) Statistical Theory of Hedonic Price Indices. DQE Working Paper 1, Department of Quantitative Economics, University of Freiburg/Fribourg Switzerland. URL http://ideas.repec.org/p/fri/dqewps/wp0001.html. [10, 13, 25, 28, 62, 68, 122]

Brachinger, H. W., Schips, B., and Stier, W. (1999) *Revision Landesindex 2000: Expertise zur Relevanz des «Boskin-Reports» für den schweizerischen Landesindex der Konsumentenpreise*. Statistik der Schweiz. Bundesamt für Statistik, Neuchâtel. ISBN 3-303-05385-5. [2, 3]

Bundesamt für Statistik (ed.) (2006) *Der neue Landesindex der Konsumentenpreise: Dezember 2005 = 100 – Methodenübersicht und Gewichtung 2006*. BFS aktuell. Bundesamt für Statistik, Neuchâtel. [3, 143]

Chambers, J. M. (1998) *Programming with data: a guide to the S language*. Springer-Verlag, New York. ISBN 0-387-98503-4. [101]

Chambers, J. M. and Hastie, T. J. (1993) *Statistical Models in S*. Chapman & Hall, New York. ISBN 0-412-05291-1. [101]

Court, A. T. (1939) Hedonic Price Indexes with Automotive Examples. *In The Dynamics of Automobile Demand*, pp. 99–117. General Motors Corporation, New York. [3, 78]

Curry, B., Morgan, P., and Silver, M. (2001) Hedonic Regressions: Mis-specification and Neural Networks. *Applied Economics* **33**(5): 659–671. doi:10.1080/00036840122335. URL http://ideas.repec.org/a/taf/applec/v33y2001i5p659-71.html. [55]

Davidson, R. and Flachaire, E. (2001) The Wild Bootstrap, Tamed at Last. IER Working Paper 1000, Queen's Institute for Economic Research, Ontario. URL http://qed.econ.queensu.ca/pub/papers/abstracts/download/2001/1000.pdf, accessed: 30.09.2005. [71]

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap methods and their application*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. ISBN 0-521-57471-4. [57, 69, 70, 71, 117]

Dayal, B. S. and MacGregor, J. F. (1997) Improved PLS algorithms. *Journal of Chemometrics* **11**(1): 73–85. doi:10.1002/(SICI)1099-128X(199701)11:1<73::AID-CEM435>3.0.CO;2-#. [57, 108]

de Haan, J. (2004) Direct and indirect time dummy approaches to hedonic price measurement. *Journal of Economic and Social Measurement* **29**(4): 427–443. [67]

Dickie, M., Delorme, C. D., Jr, and Humphreys, J. M. (1997) Hedonic prices, goods-specific effects and functional form: inferences from cross-section time series data. *Applied Economics* **29**(2): 239–249. [13, 55, 59]

Diewert, W. E. (1993) Fisher Ideal Output, Input and Productivity Indexes Revisited. *In* W. E. Diewert and A. O. Nakamura (eds.), *Essays in Index Number Theory*, vol. 1, pp. 317–353. North-Holland, Amsterdam. ISBN 0-444-87471-2. [34]

Diewert, W. E. (1995) Axiomatic and Economic Approaches to Elementary Price Indexes. NBER Working Paper 5104, National Bureau of Economic Research, Vancouver. URL http://www.nber.org/papers/w5104.pdf, accessed: 13.03.2006. [34]

Diewert, W. E. (1999) Axiomatic and Economic Approaches to International Comparisons. *In* A. Heston and R. E. Lipsey (eds.), *International and Interarea Comparisons of Income, Output, and Prices*, vol. 61 of *Studies in Income and Wealth*, pp. 13–87. National Bureau of Economic Research, The University of Chicago Press, Chicago and London. [34]

Diewert, W. E. (2003a) Hedonic Regressions: A Consumer Theory Approach. *In* Feenstra and Shapiro (2003), pp. 317–348. URL http://www.econ.ubc.ca/diewert/scan.pdf, accessed: 24.03.2005. [16, 52]

Diewert, W. E. (2003b) Hedonic Regressions: A Review of Some Unresolved Issues. URL http://www.ipeer.ca/papers/Diewert,Jan.2,2003,HedonicsandUnresolvedIssues.pdf, accessed: 08.06.2005. [104]

Dorfman, A. H., Leaver, S., and Lent, J. (1999) Some Observations on Price Index Estimators. Statistical Policy Working Paper 29, Bureau of Labor Statistics. URL http://www.bls.gov/ore/pdf/st990080.pdf, accessed: 11.09.2006. [vii, 26]

Eatwell, J., Milgate, M., and Newman, P. (eds.) (1987) *The New Palgrave: A Dictionary of Economics*. The Macmillan Press Limited, London. ISBN 0-935859-10-1. [185, 187]

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. No. 57 in Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton. ISBN 0-412-04231-2. [117]

Eichhorn, W. (1976) Fisher's Tests Revisited. *Econometrica* **44**(2): 247–256. [34]

Eichhorn, W. (1978) *Functional Equations in Economics*, vol. 11 of *Applied Mathematics and Computation*. Addison-Wesley, Reading. ISBN 0-201-01949-3. [34, 37]

Eichhorn, W. and Voeller, J. (1976) *Theory of the Price Index*. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, Berlin. ISBN 0-387-08059-7. [34]

Feenstra, R. C. and Shapiro, M. D. (eds.) (2003) *Scanner Data and Price Indexes*, vol. 64 of *Studies in Income and Wealth*. The University of Chicago Press, Chicago. [182, 187]

Fisher, I. (1922) *The Making of Index Numbers*. Houghton Mifflin, Boston. [34]

Free Software Foundation (1991) GNU General Public License. Version 2. URL `http://www.gnu.org/licenses/gpl.html`. [165]

Fristedt, B. and Gray, L. (1997) *A Modern Approach to Probability Theory*. Probability and its Applications. Birkhäuser, Boston. ISBN 0-8176-3807-5. [34, 39]

German Federal Statistical Office (2003) Hedonic Methods of Price Measurement for Used Cars. URL `http://www.destatis.de/download/e/preise/hed_used_cars.pdf`, accessed: 20.01.2006. [78]

Gordon, R. J. (1990) *The Measurement of Durable Goods Prices*. University of Chicago Press, Chicago. ISBN 0-226-30455-6. [78, 96]

Griliches, Z. (1971a) Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change. *In* Griliches (1971c), pp. 55–87. [52, 67, 78]

Griliches, Z. (1971b) Hedonic Price Indexes Revisited. *In* Griliches (1971c), pp. 3–15. [78]

Griliches, Z. (ed.) (1971c) *Price Indexes and Quality Change*. Harvard University Press, Cambridge. ISBN 0-67470420-7. [183]

Harrell, F. E., Jr (2001) *Regression Modeling Strategies*. Springer-Verlag, New York. ISBN 0-387-95232-2. [101]

Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York. ISBN 0-387-95284-5. [27, 50, 55, 58]

Hirshleifer, J. (1980) *Price theory and applications*. Prentice Hall, Englewood Cliffs, 2nd edn. ISBN 0-13-699744-9. [9]

Hulten, C. R. (2003) Price Hedonics: A Critical Review. *Economic Policy Review* **9**(3): 5–15. URL http://www.ny.frb.org/research/epr/03v09n3/0309hult.pdf, accessed: 17.05.2004. [4, 66, 139, 140]

ILO, IMF, OECD, UNECE, Eurostat, and The World Bank (eds.) (2004) *Consumer Price Index Manual: Theory and Practice*. International Labour Office, Geneva. ISBN 92-2-113699-X. URL http://www.ilo.org/public/english/bureau/stat/guides/cpi/. [1, 2, 3, 9, 10, 12, 21, 26, 34, 37, 44, 46, 64, 124]

James, D. A. and DebRoy, S. (2006) *RMySQL: R interface to the MySQL database*. URL http://stat.bell-labs.com/RS-DBI. [88]

Kondylis, A. and Hadi, A. S. (2005) Derived components regression using the BACON algorithm. *Computational Statistics & Data Analysis* doi:10.1016/j.csda.2005.11.004. [109]

Lancaster, K. J. (1971) *Consumer Demand: A New Approach*. No. 5 in Columbia Studies in Economics. Columbia University Press, New York. ISBN 0-231-03357-5. [10, 13]

Lehmann, E. L. (1986) *Testing Statistical Hypotheses*. Springer-Verlag, New York, 2nd edn. ISBN 0-387-94919-4. [39]

Liu, R. Y. (1988) Bootstrap procedures under some non-I.I.D. models. *The Annals of Statistics* **16**(4): 1696–1708. [71]

MacKinnon, J. G. (2002) Bootstrap inference in econometrics. *Canadian Journal of Economics* **35**(4): 615–645. [71]

Mammen, E. (1993) Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics* **21**(1): 255–285. [72]

Martens, H. and Næs, T. (1989) *Multivariate Calibration*. John Wiley & Sons, Chichester. ISBN 0-471-93047-4. [56, 107]

Mevik, B.-H. and Cederkvist, H. R. (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics* **18**(9): 422–429. doi:10.1002/cem.887. [108, 117]

Milgate, M. (1987) Goods and Commodities. *In* Eatwell et al. (1987), pp. 546–549. [9, 13]

Montgomery, D. C. and Peck, E. A. (1992) *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York, 2nd edn. ISBN 0-471-53387-4. [54, 131]

Müller, A. and Stoyan, D. (2002) *Comparison Methods for Stochastic Models and Risks*. Wiley Series in Probability and Statistics. Wiley, Chichester. ISBN 0-471-49446-1. [39]

Murray, J. and Sarantis, N. (1999) Price-Quality Relations and Hedonic Price Indexes for Cars in the United Kingdom. *International Journal of the Economics of Business* **6**(1): 5–27. doi:10.1080/13571519984287. URL http://ideas.repec.org/a/taf/ijecbs/v6y1999i1p5-27.html. [54, 78]

Nair, B. P. (2004) Use of hedonic regression methods for quality adjustments in Statistics NZ. Tech. Rep., Statistics New Zealand, Wellington. URL http://www.stats.govt.nz/NR/rdonlyres/73C85200-3FE5-4E7E-8C8D-B9B178D8565F/0/UseofHedonicRegressionMethodsforQualityAdjustmentsinSNZ.pdf, accessed: 24.05.2005. [78]

Neter, J., Wasserman, W., and Kutner, M. H. (1985) *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Richard D. Irwin, Homewood, 2nd edn. ISBN 0-256-02447-2. [111]

Pakes, A. (2003) A Reconsideration of Hedonic Price Indexes with an Application to PC's. *American Economic Review* **93**(5): 1578–1596. doi:10.1257/000282803322655455. [66, 67, 70]

Pakes, A. (2005) Hedonics and the Consumer Price Index. Working paper, Harvard University and the N.B.E.R. URL http://post.economics.harvard.edu/faculty/pakes/papers/Hedonics-CPI_5-11-05.pdf, accessed: 09.06.2005. [66]

Pashigian, B. P. (2001) The Used Car Price Index: A Checkup and Suggested Repairs. Working Paper 338, Bureau of Labor Statistics. URL http://www.bls.gov/ore/pdf/ec010060.pdf, accessed: 24.01.2006. [78]

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org. [88, 165]

R Special Interest Group on Databases (R-SIG-DB) (2006) *DBI: R Database Interface*. URL http://stat.bell-labs.com/RS-DBI. [88]

Reis, H. J. and Santos Silva, J. M. C. (2002) Hedonic Prices Indexes for New Passenger Cars in Portugal (1997–2001). Working Paper 10-02, Banco de Portugal Economic Research Department, Lisboa. URL http://ideas.repec.org/p/wpa/wuwpem/0303003.html, accessed: 21.05.2004. [71, 78]

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *The Journal of Political Economy* **82**(1): 34–55. [16, 52]

Ross, S. M. (1983) *Stochastic processes*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York. ISBN 0-471-09942-2. [39]

Schultze, C. L. and Mackie, C. (eds.) (2002) *At What Price? Conceptualizing and Measuring Cost-of-Living and Price Indexes*. National Academy Press, Washington, DC. [139]

Settele, C. (2005) Der Autohandel rollt ins Netz: Der Online-Markt für Motorfahrzeuge erlebt einen Boom. *Neue Zürcher Zeitung* **226**(224): 36. 26.09.2005. [78]

Silver, M. (2003) Use of hedonic regressions. URL http://www.ottawagroup.org/pdf/07/Session%20Hedonic%20Regression.pdf, accessed: 24.03.2006. Report of the session entitled 'Use of hedonic regressions' at the seventh meeting of the 'Ottawa' International Working Group on Price Indices. [67]

Silver, M. and Heravi, S. (2003) The Measurement of Quality-Adjusted Price Changes. *In* Feenstra and Shapiro (2003), pp. 277–316. [67, 68]

Silver, M. and Heravi, S. (2004a) The Difference Between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes for Desktop PCs. Working Paper 2004/02, University of New South Wales, Centre for Applied Economic Research, Sydney. URL http://www.caer.unsw.edu.au/DP/CAER0402.pdf, accessed: 24.05.2005. [67, 68]

Silver, M. and Heravi, S. (2004b) Hedonic Price Indexes and the Matched Models Approach. *The Manchester School* **72**(1): 24–49. ISSN 1463-6786. [68]

Silver, M. and Heravi, S. (2006) Why elementary price index number formulas differ: Evidence on price dispersion. *Journal of Econometrics* doi:10.1016/j.jeconom.2006.07.017. [26, 64]

Tenenhaus, M. (1998) *La régression PLS*. Technip, Paris. [56]

Tenenhaus, M., Pagès, J., Ambroisine, L., and Guinot, C. (2004) PLS methodology to study relationships between hedonic judgments and product characteristics. *Food Quality and Preference* **16**(4): 315–325. [55]

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2004) *snow: Simple Network of Workstations*. URL http://www.stat.uiowa.edu/~luke/R/cluster/cluster.html. [88]

Triplett, J. E. (1987) Hedonic Functions and Hedonic Indexes. *In* Eatwell et al. (1987), pp. 630–634. [13]

Triplett, J. E. (2004) Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products. Working Paper 2004/9, OECD Directorate for Science, Technology and Industry, Paris. doi:10.1787/643587187107. URL http://ideas.repec.org/p/oec/stiaaa/2004-9-en.html, accessed: 17.09.2005. [3, 4, 52, 53, 54, 55, 57, 59, 66, 67, 71, 77, 105]

United Nations (ed.) (1993) *System of National Accounts 1993*, chap. XVI. Price and Volume Measures. United Nations. ISBN 92-1-161352-3. URL http://unstats.un.org/unsd/sna1993/, accessed: 26.01.2004. [1, 13, 14, 15]

van Dalen, J. and Bode, B. (2004) Quality-corrected price indices: the case of the Dutch new passenger car market, 1990–1999. *Applied Economics* **36**(11): 1169–1197. doi:10.1080/0003684042000247361. URL http://journalsonline.tandf.co.uk/link.asp?id=prehqywlvakucn1v. [78]

Van Reenen, J. (2005) The Growth of Network Computing: Quality Adjusted Price Changes for Network Servers. Discussion Paper 702, Centre for Economic Performance, London School of Economics and Political Science. URL http://cep.lse.ac.uk/pubs/download/dp0702.pdf, accessed: 24.08.2005. [67, 70]

van Wezel, M., Kagie, M., and Potharst, R. (2005) Boosting the Accuracy of Hedonic Pricing Models. Econometric Institute Report EI 2005-50, Econometric Institute, Erasmus University, Rotterdam. URL https://ep.eur.nl/bitstream/1765/7145/1/ei2005-50.pdf, accessed: 4.1.2006. [55]

Velleman, P. F. and Welsch, R. E. (1981) Efficient Computing of Regression Diagnostics. *The American Statistician* **35**(4): 234–242. [101]

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Statistics and Computing. Springer-Verlag, New York, 4th edn. ISBN 0-387-95457-0. [101]

Vogt, A. and Barta, J. (1997) *The Making of Tests for Index Numbers*. Physica-Verlag, Heidelberg. ISBN 3-7908-1011-8. [34]

Wehrens, R. and Mevik, B.-H. (2006) *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*. URL http://mevik.net/work/software/pls.html. R package version 1.2-0. [57, 108]

Whelan, S. (2005) RPI: a Really Profligate Index? *Significance* **2**(2): 66–68. [139, 140]

White, A. G., Abel, J. R., Berndt, E. R., and Monroe, C. W. (2004) Hedonic Price Indexes for Personal Computer Operating Systems and Productivity Suites. Working Paper 10427, National Bureau of Economic Research. URL http://ideas.repec.org/p/nbr/nberwo/10427.html, accessed: 22.06.2004. [4, 78]

Yu, H. (2004) *Rmpi: Interface (Wrapper) to MPI (Message-Passing Interface)*. URL http://www.stats.uwo.ca/faculty/yu/Rmpi. [88]

Yu, K. (2003) An Elementary Price Index for Internet Service Providers in Canada: A Hedonic Study. Working paper, Department of Economics, Lakehead University, Thunder Bay, Ontario. URL http://flash.lakeheadu.ca/~kyu/Papers/ISP2001.pdf, accessed: 12.11.2004. [54, 55, 64]

The hedonic approach is currently seen as the most promising method for constructing quality-adjusted price indices. Building upon a novel axiomatic framework, the current piece of work tackles the fundamental question of what hedonic elementary price indices actually measure. They reflect the average price change over time for a set of products of constant quality. Furthermore, they are latent economic parameters that require a precise definition before being estimated from empirical data.

Once the set of suitable definitions is specified, it turns out that most of the well-known index formulae are natural estimators of particular hedonic elementary price indices. They are all based on estimates of the hedonic function relating the characteristics (and thus the quality) of a product to its price. Adapted bootstrap resampling methods may be used to explore the stochastic nature of such estimators of 'pure' inflation.

The concepts presented in this study are illustrated by an empirical analysis of the market of used cars in Switzerland.

ISBN 978-3-033-01099-4