Statistical Theory of Hedonic Price Indices

Hans Wolfgang Brachinger

Seminar of Statistics, University of Fribourg, Av. de Beauregard 13,
CH-1700 Fribourg, hanswolfgang.brachinger@unifr.ch

Summary. In the economic literature, essentially, hedonic techniques either are applied straightforwardly or the economic foundations of the hedonic hypothesis are discussed. In this paper, the statistical foundations of hedonic price indices are developed. After a short overview on well-known functional forms of hedonic equations, first, precise hedonic notions of a good and its price are specified. These specifications allow a clear-cut definition of true hedonic price indices. Then, the problem of estimating hedonic price indices is treated. It is shown, first, that the usual hedonic price index formulae result from estimating certain true indices in a special way and, second, that the techniques used in practice for estimating hedonic indices are just first approaches.

Key words: Hedonic regression, functional forms of hedonic equations, hedonic price of a good, true hedonic price index, true hedonic Laspeyres price index, true hedonic Paasche price index, true adjacent periods price index, estimated hedonic price indices

1 Introduction

One of the classical goals of price statistics is the quantification of the "true price change" of a good given a certain quality. The problem is that qualities change in time and the goods of today are no more the same as yesterday. So the goods actually available on the market are no more directly comparable with those which were available before. For price comparisons, prices have to be quality adjusted.

Quality adjustment commonly is regarded as one of the most complicated problems in price statistics. At the latest after the publication of the "Boskin-Report" [6] in 1996, the well-known classical techniques like "linking" or "overlap pricing" have come under attack because they are, under certain conditions, prone to considerable "biases". Hedonic methods have been recommended as a reasonable alternative.

Hedonic methods already have been proposed not later than more than 30 years ago in the well-known classical paper written by Griliches 1961 [10]. In the last years, official price statistics have begun to use hedonic methods on a regular basis (see, e.g., [14]).

In this paper, the statistical foundations of hedonic price indices are developed. In an introductory section, the hedonic hypothesis is briefly presented and a short overview of well-known functional forms of hedonic equations is given. In the first main section, to start with, precise hedonic notions of a good and its price are introduced. Then, these specifications are used to derive clear-cut definitions of true hedonic price indices. In the second main section, the problem of estimating hedonic price indices is treated. It is shown, first, that the usual hedonic price index formulae result form estimating certain true indices in a special way and, second, that the techniques used in practice for estimating hedonic indices are just first approaches. The paper closes with some final remarks.

2 Hedonic Regression

2.1 Hedonic Hypothesis

The starting point of every hedonic price index is the *hedonic hypothesis*. The core of this hypothesis is that each good is characterized by the set of all its characteristics. For any given good, let this set be ordered and denoted by $\mathbf{x} = (x_1, \ldots, x_K)'$. It is assumed that the preferences of the economic actors with respect to any good are solely determined by its corresponding characteristics vector.

Furthermore, it is assumed that, for any good, there is a functional relationship f between its price p and its characteristics vector $\mathbf{x}$, i.e.

$$p = f(\mathbf{x}) \ . \tag{1}$$

This function specifies the *hedonic relationship* or *hedonic regression* typical for the good.

This idea of the hedonic hypothesis has been advanced long ago by Lancaster [16] and Rosen [17] in famous articles. For a recent economic foundation of the hedonic hypothesis which is more easily accessible, see the paper by Diewert [9].

Based on the functional relationship (1) the important concept of *implicit* or *hedonic prices* can be introduced. These prices are defined to be the partial derivatives of the hedonic function (1), i.e., they are defined through

$$\frac{\partial p}{\partial x_k}(x) = \frac{\partial f}{\partial x_k}(x) \qquad (k = 1, \ldots, K) \ . \tag{2}$$

The hedonic price $\partial f / \partial x_k(x)$ indicates, how much the price p of a good (approximately) changes if this good is, ceteris paribus, endowed with an additional unity of the characteristic x_k.

For practical applications of the hedonic relationship (1) in price statistics, of course, the main problems are to determine the characteristics vector typical of a good and to specify the hedonic function (1). An overview of different functional forms of hedonic regressions is given in the next section. For concrete applications in price statistics see, e.g., the papers by Berndt [2] and Hoffmann [13].

2.2 Functional Forms of Hedonic Regressions

In hedonic approaches to price index problems four different functional forms have been employed in the past. In this section, an overview of these functional forms is given. Thereby, the vector $\mathbf{x}$ stands for a particular variant or model of a good considered.

The most simple approach is the ordinary *linear approach* given by

$$p = \beta_0 + \sum_{k=1}^{K} \beta_k \, x_k \tag{3}$$

with hedonic prices

$$\frac{\partial p}{\partial x_k} = \beta_k \; .$$

The regression coefficient β_k $(k = 1, \ldots, K)$ indicates the *marginal change* of the price with respect to a change of the k-th characteristic x_k of the good.

Another approach is the *exponential approach* characterized by

$$p = \beta_0 \prod_{k=1}^{K} \exp\bigl(\beta_k \, x_k\bigr) \tag{4}$$

or

$$\ln p = \ln \beta_0 + \sum_{k=1}^{K} \beta_k \, x_k \tag{5}$$

with hedonic prices

$$\frac{\partial p}{\partial x_k} = \beta_k p \; .$$

Obviously, in this approach, the regression coefficients can be interpreted as *growth rates.* The coefficient β_k $(k = 1, \ldots, K)$ indicates the rate at which the price increases at a certain level, given the characteristics $\mathbf{x}$.

A third approach is the *power function* or *double log approach* described by

$$p = \beta_0 \prod_{k=1}^{K} x_k^{\beta_k} \tag{6}$$

or

$$\ln p = \ln \beta_0 + \sum_{k=1}^{K} \beta_k \, \ln x_k \tag{7}$$

with hedonic prices

$$\frac{\partial p}{\partial x_k} = \frac{\beta_k}{x_k} p \, .$$

In this approach, the regression coefficients can be interpreted as *partial elasticities.* The coefficient β_k $(k = 1, \ldots, K)$ indicates how many percent the price p increases at a certain level if the k-th characteristic x_k changes by one percent.

A fourth approach is the *logarithmic approach* given by

$$p = \beta_0 + \sum_{k=1}^{K} \beta_k \, \ln x_k \tag{8}$$

with hedonic prices

$$\frac{\partial p}{\partial x_k} = \frac{\beta_k}{x_k} \, .$$

In the following section, now, it will be shown how the general hedonic hypothesis (1) can be used to introduce precise hedonic notions of a good and its price. Then, on the basis of these notions clear-cut definitions of true hedonic price indices are derived.

3 Hedonic Index Concepts

3.1 Hedonic Concept of a Good

In the economics literature the notion of a good is a primitive used as a basic concept for the development of most theoretical results. However, nowhere a general empirical statistical definition of the notion of a good can be found. In theory, this is not a problem but for the daily business of a price statistician it definitely is. She, always, is confronted with lots of different commodities and has to decide which of these commodities may still count as a variant of a certain good and which one no more.

The notion of a homogeneous good is an “Idealtyp” in the sense of Max Weber. Goods as such empirically do not exist. What empirically exists are concrete examples or variants of the idea of a certain good all of which are more or less different. All of those variants are empirical operationalizations of this good. There is always a certain difference between the idea of a good and an empirical variant of it. It is an empirical statistical question how far the differences between variants may go to count still as variants of a certain good. Every variant comprises characteristics which go further than the basic idea of that good.

The hedonic hypothesis (1) allows the following precise definition of the notion of a good: A *good* is characterized by the set of all those models or variants j which fit under one and the same hedonic equation, i.e., a *good* is characterized by the set of all variants j whose prices p_j can be explained by the same set of characteristics variables $\mathbf{x} = (x_1, \ldots, x_K)'$ and the same structure of a certain parametric family of hedonic regression functions, i.e., the same parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$ typical of that family.

Empirically, it is the prices of the different variants of a good which can be observed. But for each well-specified variant itself, the observable price varies. So the hedonic hypothesis (1) implies the following general *statistical model for the observable prices*: Any *observable price* p_j is a random variable

$$p_j = f(\mathbf{x}_j, \boldsymbol{\beta}) + u \tag{9}$$

where $\mathbf{x}_j = (x_{1j}, \ldots, x_{Kj})'$ is the K-vector of characteristics values identifying the variant for which p_j is measured and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$ the unknown parameter vector characterizing the good to which variant j belongs. The variable u denotes a stochastic error term.

Thereby, usually it is assumed that, given a certain variant characterized by $\mathbf{x}_j$, for the conditional expectation $\mathrm{E}(u \mid \mathbf{x}_j)$ and the variance $\mathrm{Var}(u \mid \mathbf{x}_j)$ of the error term u the classical assumptions hold, i.e., $\mathrm{E}(u \mid \mathbf{x}_j) = 0$ and $\mathrm{Var}(u \mid \mathbf{x}_j) = \sigma^2$. These assumptions imply $\mathrm{E}(p_j \mid \mathbf{x}_j) = f(\mathbf{x}_j, \boldsymbol{\beta})$ and $\mathrm{Var}(p_j \mid \mathbf{x}_j) = \sigma^2$.

3.2 Hedonic Price of a Good

In Sect. 3.1 is has been argued that the notion of a homogeneous good is an *Idealtyp* which as such empirically does not exist. As a consequence, *the price* of a good does not exist either. The basic idea of any hedonic price index concept is that the price $\mathbf{P}$ of a good is a latent, i.e. non-observable variable whose values are generated by a hedonic regression on the average variant belonging to that good. The values of this variable vary over time.

Statistically speaking, the price $\mathbf{P}$ of a good is assumed to be a non-stationary discrete stochastic process $\mathbf{P} = (P^t)_{t \in T}$ defined by the hedonic regression

$$P^t = f(\mathrm{E}(\mathbf{x}^t), \boldsymbol{\beta}^t) + v^t \tag{10}$$

where $\mathbf{x}^t = (x_1^t, \ldots, x_K^t)'$ is, at time t, the characteristics random vector varying over the population of all models or variants belonging to the good considered and $\mathrm{E}(\mathbf{x}^t)$ its expectation. The parameter vector $\boldsymbol{\beta}^t = (\beta_0^t, \beta_1^t, \ldots, \beta_K^t)'$ is, at time t, the parameter vector characterizing the good considered and the conditional expectation of the error term v^t is assumed to be 0, i.e. $\mathrm{E}(v^t \mid \mathrm{E}(\mathbf{x}^t)) = 0$. Note that in this definition both, the expectation $\mathrm{E}(\mathbf{x}^t)$ of the characteristics vector and the parameter vector $\boldsymbol{\beta}^t$ may vary over time.

Equation (10), now, allows a precise definition of the price of a good at time t: For any good, the *hedonic price* Π^t *at time* t is the expectation of the stochastic process (10) at time t, i.e.

$$\Pi^t := \mathrm{E}P^t = f(\mathrm{E}(\mathbf{x}^t), \boldsymbol{\beta}^t) \; , \tag{11}$$

where the expectation is taken over all models or variants belonging to the good considered, at time t. In general, a *hedonic price* Π *of a good* is a function

$$\Pi(\mathrm{E}(\mathbf{x}), \boldsymbol{\beta}) := f(\mathrm{E}(\mathbf{x}), \boldsymbol{\beta}) \; , \tag{12}$$

where the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$ is the parameter vector characterizing the good considered, at any time, and the expectation is taken over any population of models or variants belonging to the good considered.

Note that equation (10) also allows a precise definition of the quality of a good at time t: For any good, its *hedonic quality at time* t is given by the expectation $\mathrm{E}(\mathbf{x}^t)$. In general, a *hedonic quality of a good* is given by the expectation $\mathrm{E}(\mathbf{x})$ over any population of models or variants belonging to the good considered.

3.3 True Hedonic Price Indices

The general definition (12) of a hedonic price, now, serves as the starting point for hedonic price index concepts. The basic idea of any such index is to compare the hedonic price of a particular good at two different times while holding the quality of the good constant. Depending on for which period the quality of the good is held constant different index concepts result.

The Classical True Hedonic Price Indices

A first hedonic price index concept results when according to Laspeyres' classical idea the quality of the base period is held constant. The *true hedonic Laspeyres price index* for a certain good, at time t relative to a base period 0, is defined by

$$HPI^{\mathrm{L}}_{0,t} := \frac{\Pi^t_{\mathrm{qcorr}}}{\Pi^0} := \frac{\Pi(\mathrm{E}(\mathbf{x}^0), \boldsymbol{\beta}^t)}{\Pi(\mathrm{E}(\mathbf{x}^0), \boldsymbol{\beta}^0)} \; . \tag{13}$$

A second hedonic price index concept results when according to Paasche's idea the quality of the comparison period is taken as reference quality. The *true hedonic Paasche price index* for a certain good, at time t relative to a base period 0, is defined by

$$HPI^{\mathrm{P}}_{0,t} := \frac{\Pi^t}{\Pi^0_{\mathrm{qcorr}}} := \frac{\Pi(\mathrm{E}(\mathbf{x}^t), \boldsymbol{\beta}^t)}{\Pi(\mathrm{E}(\mathbf{x}^t), \boldsymbol{\beta}^0)} \; . \tag{14}$$

The True Adjacent Periods Price Index

A further hedonic price index concept results when the good considered is identified with the population of all variants existing at least in one of the two periods, base and the comparison period. Then, according to the general definition of the hedonic quality of a good in Sect. 3.2, the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ serves as the reference quality to be held constant. Thereby $\mathbf{x}^{0\cup t}$ denotes the characteristics random vector of the good considered, when all variants existing at least in one of the two periods, base and the comparison period, are admitted.

The idea to identify a good with the population of all variants existing at least in one of the two periods, base and the comparison period, implies that the parameters β_k corresponding to the characteristics x_k $(k = 1, \ldots, K)$ cannot change between the two periods: these parameters are typical for the good considered. Therefore, it is assumed that these parameters remain constant between the two periods.

Only the intercept parameter β_0 may change and the extent of this change, then, measures the influence of the price change between the two periods on the hedonic price of the good considered. In other words, the two parameter vectors $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^t$ characterizing the hedonic prices (11) of the good for times 0 and t, respectively, are assumed to be identical except for their intercept values. Let $\boldsymbol{\beta}^0 = (\beta_0^0, \beta_1, \ldots, \beta_K)' = (\beta_0^0, \boldsymbol{\beta}'_{(-0)})'$ and $\boldsymbol{\beta}^t = (\beta_0^t, \beta_1, \ldots, \beta_K)' = (\beta_0^t, \boldsymbol{\beta}'_{(-0)})'$ denote these parameter vectors.

Starting from this conception and, therefore, taking the quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ as reference quality and holding the regression parameters belonging to the characteristics constant, a third hedonic price index concept results which may be called the *true adjacent periods price index*. For a given good, this index at time t relative to a base period 0, is defined by

$$HPI_{0,t}^{\mathrm{ap}} := \frac{\Pi_{\mathrm{ap}}^t}{\Pi_{\mathrm{ap}}^0} := \frac{\Pi\big(\mathrm{E}(\mathbf{x}^{0\cup t}), \boldsymbol{\beta}^t\big)}{\Pi\big(\mathrm{E}(\mathbf{x}^{0\cup t}), \boldsymbol{\beta}^0\big)} = \frac{\Pi\big(\mathrm{E}(\mathbf{x}^{0\cup t}), (\beta_0^t, \boldsymbol{\beta}'_{(-0)})'\big)}{\Pi\big(\mathrm{E}(\mathbf{x}^{0\cup t}), (\beta_0^0, \boldsymbol{\beta}'_{(-0)})'\big)} \ . \tag{15}$$

In Sect. 3.1 it has been pointed out that according to the hedonic hypothesis (1) a good is characterized by the set of all those models or variants j which fit under one and the same hedonic equation, i.e., by the set of all variants j whose hedonic price (11) can be explained by the same set of characteristics and the same structure of a certain parametric family of regression functions.

In the case of the adjacent periods price index, to get the base period and the comparison period variants of the good considered under one and the same hedonic equation, the set of the characteristics typical of that good is supplemented by the additional "characteristic" time, i.e., the dummy variable

$$D^\tau = \begin{cases} 0 & \text{for } \tau = 0 \\ 1 & \text{for } \tau = t \ . \end{cases}$$

This variable serves as an additional exogenous variable in the good's hedonic equation. The parameter vector characterizing the good is given with $\boldsymbol{\beta} = (\beta_0, \delta, \beta_1, \ldots, \beta_K)'$, where δ is the coefficient belonging to the time dummy variable, and $\beta_0^t = \beta_0^0 + \delta$.

Note that it is typical of this index concept that neither the base period nor the comparison period quality serves as reference quality. Identifying a good with the population of all variants existing at least in one of the two periods, base or comparison period, and using the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ as the reference quality to be held constant, in a certain sense, a kind of "compromise quality" is used for that purpose. In that, the hedonic adjacent periods price index resembles compromise index formulae like the well-known Fisher index or the recently very often forwarded Thørnquist index. In both indexes, the weighting scheme used is a kind of compromise between the consumption weights in base and comparison period.

Note that all of these three index concepts require the *preceding specification of a hedonic function f*. Note further that all these price indices are well-defined economic parameters which are unobservable and, therefore, have to be estimated appropriately. Any procedure to estimate one of these indices necessarily is a two-step procedure because, first, an expectation $\mathrm{E}(\mathbf{x})$ and than parameter vectors $\boldsymbol{\beta}^{\tau}$ $(\tau = 0, t)$ have to be estimated. Usually only the problem of estimating $\boldsymbol{\beta}^{\tau}$ is considered consciously. The problem of estimating $\mathrm{E}(\mathbf{x})$, regularly, is left unnoticed.

4 Estimation of Hedonic Price Indices

To compute a hedonic price index, one must first estimate a hedonic function. For that empirical data is needed. In the following section, a statistical model for such data is developed.

4.1 Statistical Model for the Data

For any index concept making use of a hedonic regression, the structure of the good considered, i.e., the parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$ of its hedonic regression has to be estimated. For that, a random sample of size M has to be drawn from the population of all variants of that good, i.e. from all commodities regarded as fitting under the hedonic function of that good in the sense of Sect. 3.1.

As each variant j of a good, in general, is characterized not only by a different price p_j but also by a different vector $\mathbf{x}_j$ of characteristics, in a suitable model for the data, price and characteristics vector, as well, have to be regarded as random variables. This means that a suitable model for the data used to estimate a hedonic regression has to be a regression model with stochastic exogenous variables.

Based on the hedonic regression (1) specific for the good considered, the model for the data is given with

$$\mathbf{p} = f(\mathbf{X}, \boldsymbol{\beta}) + \mathbf{u} \tag{16}$$

where $\mathbf{p} = (p_1, \ldots, p_M)'$ is a M-vector of the price observations of the variants drawn randomly and $\mathbf{X} = (\mathbf{x}_j')$ is the stochastic $(M \times K)$-matrix of the corresponding characteristics vectors $\mathbf{x}_j' = (x_{j1}, \ldots, x_{jK})$, $j = 1, \ldots, M$. The vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$ is the parameter vector which is typical for the good and has to be estimated appropriately.

4.2 Estimated Hedonic Price Indices

The distinguishing feature of the Laspeyres and the Paasche index concept is that the population of all variants or models of a good available during at least one of the two periods, base and comparison period, is separated in two sub-populations, the sub-population of all base period variants and the sub-population of all comparison period variants. According to that idea, for the estimation of these indexes, two data sets have to be sampled, one for the base period and one for the comparison period.

A first step in estimating a hedonic Laspeyres or Paasche price index is to draw a random sample from the population of all the base period variants. Let $(p_j^0, x_{j1}^0, \ldots, x_{jK}^0)$, $j = 1, \ldots, M^0$, denote the data sampled from the base period 0 population of all variants of the good considered. On the basis of these cross sectional data the parameter vector $\boldsymbol{\beta}^0$ of the statistical model (16) for time 0 is estimated.

In the next step a random sample from the population of all the comparison period variants has to be drawn. Let $(p_j^t, x_{j1}^t, \ldots, x_{jK}^t)$, $j = 1, \ldots, M^t$, denote the data sampled from the comparison period t population of all variants of the good considered. On the basis of these cross sectional data the parameter vector $\boldsymbol{\beta}^t$ of the statistical model (16) for time t is estimated. Both parameters, $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^t$, usually are estimated using the OLS-procedure leading to the OLS-estimates $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^t$.

Estimated True Hedonic Laspeyres Index

In a third step, to estimate the hedonic Laspeyres index, the time-0-quality of the good considered, i.e. the expectation $\mathrm{E}(\mathbf{x}^0)$ has to be estimated. As an estimator, usually, the vector

$$\overline{\mathbf{x}}^0 = (\overline{x}_1^0, \ldots, \overline{x}_K^0)' \tag{17}$$

of the arithmetic means

$$\overline{x}_k^0 = \frac{1}{M^0} \sum_{j=1}^{M^0} x_{jk}^0 \qquad (k = 1, \ldots, K) \tag{18}$$

over all the characteristics values sampled for the base period is employed. This estimation leads to the estimated quality $\hat{\mathrm{E}}(\mathbf{x}^0)$.

This estimatior together with the OLS-estimators $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^t$ leads to the estimation

$$\widehat{HPI}^{\mathrm{L}}_{0,t} := \frac{\hat{\Pi}^t_{\mathrm{qcorr}}}{\hat{\Pi}^0} = \frac{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^0), \hat{\boldsymbol{\beta}}^t\big)}{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^0), \hat{\boldsymbol{\beta}}^0\big)} = \frac{f\big(\overline{\mathbf{x}}^0, \hat{\boldsymbol{\beta}}^t\big)}{f\big(\overline{\mathbf{x}}^0, \hat{\boldsymbol{\beta}}^0\big)} \tag{19}$$

of the hedonic Laspeyres price index. This estimation of (13) is what usually is meant when, in applications, is spoken about the hedonic Laspeyres price index.

Estimated True Hedonic Paasche Index

To estimate the hedonic Paasche index, the time-t-quality of the good considered, i.e. the expectation $\mathrm{E}(\mathbf{x}^t)$ has to be estimated. As an estimator, usually, the vector

$$\overline{\mathbf{x}}^t = (\overline{x}^t_1, \ldots, \overline{x}^t_K)' \tag{20}$$

of the arithmetic means

$$\overline{x}^t_k = \frac{1}{M^t} \sum_{j=1}^{M^t} x^t_{jk} \qquad (k = 1, \ldots, K) \tag{21}$$

over all the characteristics values sampled for the comparison period is employed. This estimation leads to the estimated quality $\hat{\mathrm{E}}(\mathbf{x}^t)$.

This estimator together with the OLS-estimators $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^t$ leads to the estimation

$$\widehat{HPI}^{\mathrm{P}}_{0,t} := \frac{\hat{\Pi}^t}{\hat{\Pi}^0_{\mathrm{qcorr}}} = \frac{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^t), \hat{\boldsymbol{\beta}}^t\big)}{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^t), \hat{\boldsymbol{\beta}}^0\big)} = \frac{f\big(\overline{\mathbf{x}}^t, \hat{\boldsymbol{\beta}}^t\big)}{f\big(\overline{\mathbf{x}}^t, \hat{\boldsymbol{\beta}}^0\big)} \tag{22}$$

of the hedonic Paasche price index for any good. This estimation of (14) is what usually is meant when, in applications, is spoken about the hedonic Paasche price index.

Estimated True Hedonic Adjacent Periods Index

The distinguishing feature of the adjacent periods price index concept is that neither the base nor the comparison period quality of the good considered serve as reference quality. It is the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ resulting when the good considered is identified with the population of all variants existing at least in one of the two periods, base or comparison period, which is taken as reference quality.

According to that idea the population of all variants or models of a good available during at least one of the two periods, base or comparison period, is no more separated in two sub-populations as it is the case for the Laspeyres and Paasche indices. For the estimation of this index only one data set has to be sampled.

To estimate the adjacent periods price index, therefore, a random sample has to be drawn from the population of all the variants of the good considered available during at least one of the two periods, base or comparison period. Let $(p_j, D_j^\tau, x_{j1}, \dots, x_{jK})$, $j = 1, \dots, M$, denote the data sampled. Thereby, D^τ denotes the time dummy variable introduced in Sect. 3.3. Note that this time dummy allows to identify if an observation stems from a variant of the base or the comparison period.

On the basis of these data, first, the parameter vector

$$\boldsymbol{\beta} = (\beta_0, \delta, \beta_1, \dots, \beta_K)'$$

of the statistical model (16) has to be estimated. The most simple estimation procedure would, of course, be the OLS-method. Let any estimate of $\boldsymbol{\beta}$ be denoted by $\hat{\boldsymbol{\beta}}$.

Then, in a second step to estimate the hedonic adjacent periods index, the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ of the good considered has to be estimated. The "natural" estimator for $\mathrm{E}(\mathbf{x}^{0\cup t})$ is the vector

$$\overline{\mathbf{x}}^{0\cup t} = (\overline{x}_1^{0\cup t}, \dots, \overline{x}_K^{0\cup t})' \tag{23}$$

of the arithmetic means

$$\overline{x}_k^{0\cup t} = \frac{1}{M} \sum_{j=1}^{M} x_{jk} \qquad (k = 1, \dots, K) \tag{24}$$

over all the characteristics values sampled. This estimation leads to the estimated quality $\hat{\mathrm{E}}(\mathbf{x}^{0\cup t})$.

This estimator together with any estimator $\hat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$ leads to the estimation

$$\begin{aligned} \widehat{HPI}_{0,t}^{\mathrm{ap}\cup} &:= \frac{\hat{\Pi}_{\mathrm{ap}}^t}{\hat{\Pi}_{\mathrm{ap}}^0} := \frac{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^{0\cup t}), \hat{\boldsymbol{\beta}}^t\big)}{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^{0\cup t}), \hat{\boldsymbol{\beta}}^0\big)} = \frac{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^{0\cup t}), (\hat{\beta}_0^t, \hat{\boldsymbol{\beta}}'_{(-0)})'\big)}{\Pi\big(\hat{\mathrm{E}}(\mathbf{x}^{0\cup t}), (\hat{\beta}_0^0, \hat{\boldsymbol{\beta}}'_{(-0)})'\big)} \\ &= \frac{f\big(\overline{\mathbf{x}}^{0\cup t}, (\hat{\beta}_0^t, \hat{\boldsymbol{\beta}}'_{(-0)})'\big)}{f\big(\overline{\mathbf{x}}^{0\cup t}, (\hat{\beta}_0^0, \hat{\boldsymbol{\beta}}'_{(-0)})'\big)} \; . \end{aligned} \tag{25}$$

of the hedonic adjacent periods price index.

Alternatively Estimated Hedonic Adjacent Periods Index

In practice, usually, to get a sample where the observations from the base and the comparison period are "balanced", a different estimation procedure is employed (see, e.g., [7], [5]). First a sample of M variants of the good considered is drawn from the base period population and then all the variants drawn are reconsidered in the comparison period. This sampling procedure leads to two samples. Let the first sample concerning the base period 0 be denoted by $(p_j^0, x_{j1}^0, \dots, x_{jK}^0)$ and the second sample concerning the comparison period t by $(p_j^t, x_{j1}^t, \dots, x_{jK}^t)$, $j = 1, \dots, M$.

These data are pooled together to estimate the parameter vector $\boldsymbol{\beta} = (\beta_0, \delta, \beta_1, \ldots, \beta_K)'$ of the statistical model (16). As estimation procedure the OLS-method is used. Then, to estimate the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ of the good considered, as above, an estimator of arithmetic means is employed. But, this time, average is only taken over all the base period observations, i.e. for estimating $\mathrm{E}(\mathbf{x}^{0\cup t})$ the estimator

$$\overline{\mathbf{x}}^0 = (\overline{x}_1^0, \ldots, \overline{x}_K^0)' \tag{26}$$

with

$$\overline{x}_k^0 := \frac{1}{M} \sum_{j=1}^{M} x_{jk}^0 \qquad (k = 1, \ldots, K) \; . \tag{27}$$

is used.

This estimation, in general, of course leads to an estimated quality $\hat{\mathrm{E}}(\mathbf{x}^{0\cup t})$ different to the estimation introduced above. As a consequence, a different estimation of the hedonic adjacent periods price index results. This estimation is given by

$$\widehat{HPI}_{0,t}^{\mathrm{ap}0} := \frac{f\big(\overline{\mathbf{x}}^0, (\hat{\beta}_0^t, \hat{\boldsymbol{\beta}}_{(-0)}')'\big)}{f\big(\overline{\mathbf{x}}^0, (\hat{\beta}_0^0, \hat{\boldsymbol{\beta}}_{(-0)}')'\big)} \; . \tag{28}$$

This estimation of (15) is what usually is meant when, in applications, is spoken of hedonic adjacent periods price index.

However, it should be noted that this practice is in contradiction to the basic idea of the true hedonic adjacent periods price index (15). It has been mentioned in Sect. 3.3 that it is typical of this index concept that neither the base period nor the comparison period quality serves as reference quality. With the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$, in a certain sense, deliberately, a kind of "compromise quality" is used for that purpose. Estimating the hedonic quality $\mathrm{E}(\mathbf{x}^{0\cup t})$ by (26) means to come back to Laspeyres' concept and to favor the base period.

Some well-known hedonic price index formulae result from (25) as special cases when special functional forms of hedonic functions are assumed. If, e.g., the power function approach (6) is employed then (25) leads to

$$\begin{aligned} \widehat{HPI}_{0,t}^{\mathrm{ap}\cup} &:= \frac{f\big(\overline{\mathbf{x}}^{0\cup t}, (\hat{\beta}_0^t, \hat{\boldsymbol{\beta}}_{(-0)}')'\big)}{f\big(\overline{\mathbf{x}}^{0\cup t}, (\hat{\beta}_0^0, \hat{\boldsymbol{\beta}}_{(-0)}')'\big)} \\ &= \frac{\exp(\ln \hat{\beta}_0^0 + \hat{\delta}) \prod_{k=1}^{K} (\overline{x}_k)^{\hat{\beta}_k}}{\exp(\ln \hat{\beta}_0^0) \prod_{k=1}^{K} (\overline{x}_k)^{\hat{\beta}_k}} = \exp(\hat{\delta}) \; . \end{aligned} \tag{29}$$

Note that this index is independent of how the hedonic quality of the good considered is estimated.

Another well-known hedonic price index formula results from (25) when the linear approach (3) is employed. In that case (25) leads to

$$\begin{aligned}\widehat{HPI}_{0,t}^{\mathrm{ap}\cup} &:= \frac{f\big(\overline{\mathbf{x}}^{0\cup t}, (\hat{\beta}_0^t, \hat{\boldsymbol{\beta}}'_{(-0)})'\big)}{f\big(\overline{\mathbf{x}}^{0\cup t}, (\hat{\beta}_0^0, \hat{\boldsymbol{\beta}}'_{(-0)})'\big)} \\ &= \frac{\hat{\beta}_0^0 + \hat{\delta} + \sum_{k=1}^K \hat{\beta}_k \overline{x}_k^{0\cup t}}{\hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k \overline{x}_k^{0\cup t}} = 1 + \frac{\hat{\delta}}{\hat{\beta}_0^0 + \sum_{k=1}^K \hat{\beta}_k \overline{x}_k^{0\cup t}} \, . \qquad (30)\end{aligned}$$

Contrary to the index (29) this index depends on how the hedonic quality of the good considered is estimated.

4.3 Statistical Problems and Techniques

There are *three main statistical problems* involved in the concept of a true hedonic price index.

Specification of the Good

As the hedonic price index concept relies on the hedonic notion of a good, first, the population of all variants defining a certain good in the sense of Sect. 3.1 has to be determined empirically. I.e., it has to be decided which commodities fit under one and the same hedonic equation in the sense that their prices p_j can be explained by the same set of characteristics variables $\mathbf{x} = (x_1, \ldots, x_K)'$ and the same parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$ typical of a certain parametric family of hedonic regression functions.

In practice, when starting to consider a certain good, the first decision to be taken is about the characteristics characterizing that good in the sense of the hedonic hypothesis. This problem is, for some goods, e.g., personal computers, solvable. But for others a sufficient specification of characteristics might be very hard to reach. So this is a first reason why, for conducting a hedonic study, it is so important to "know your product" [20, p. 64].

A second decision to be taken is about the functional form of the hedonic regression. There is no immediate statistical technique for helping the practitioner to make the "best" choice. In the most intensely studied case of a good, personal computer, the double log approach (7) has proved to be the most preferable one (see, e.g., [5]). In other cases other approaches have been used. When starting to introduce a hedonic index in practice it is recommendable to start with one, e.g., the double log approach, and then try out other approaches and see which one yields the results which seem most reasonable given the practical knowledge of the commodities considered. So this is a second reason why, for conducting a hedonic study, it is so important to "know your product".

Assume that, on the basis of empirical a-priori reasons (e.g., practical knowledge of the commodities), for a larger class of commodities a certain functional form of the hedonic regression has already been chosen. Then, the decision which of these commodities really are variants of a certain good, i.e. are characterized by the same parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$, should be taken on the basis of the results of suitable statistical test procedures.

For that purpose, again on the basis of practical knowledge, this a-priori set of commodities will, e.g., be separated in two disjoint subsets. Then, for each subset, the hedonic parameters have to be estimated on the basis of samples from these subsets. Finally, the null hypothesis that the slope coefficients of the different subsets are equal has to be tested. Suitable procedures for this test problem can be found in the statistics and econometrics literature (see, e.g., [15]).

If the null hypothesis is rejected, then the two populations of variants from which the samples had been drawn should be regarded as populations characterizing different goods. If this is not the case, however, there is no empirical evidence against treating these two populations as one population characterizing one certain good.

An illustrative example for a good hedonic study where first it had to be decided which commodities should be treated as variants of a certain good, is the empirical study of Berndt [2]. In their study, the well-known Chow test procedure is employed for testing for equality of regression parameters.

Estimation of Average Quality

It has already been mentioned in Sect. 3.3 that any procedure to estimate a true hedonic price index must comprise an estimator for a hedonic quality $\mathrm{E}(\mathbf{x})$ of the good considered. In the applied literature on hedonic price indices, regularly, this estimation problem is not explicitly treated. As an estimator, usually, a vector

$$\overline{\mathbf{x}} = (\overline{x}_1, \ldots, \overline{x}_K)' \tag{31}$$

of arithmetic means is employed.

The task of estimating the hedonic quality $\mathrm{E}(\mathbf{x})$ of a good consists of estimating the mathematical expectation of the random vector $\mathbf{x}$. It is well-known that the estimator $\overline{\mathbf{x}}$, statistically, is optimal in many senses: it is unbiased independently of the underlying distribution and it is efficient, i.e., minimum variance estimator when the underlying distribution is multivariate normal.

However, the routine application of the arithmetic mean, as it is common practice in applied hedonic studies, is a dangerous endeavor. Despite its qualities, it is also well-known that the arithmetic mean is very sensitive against outliers. In the case of heavy-tailed distributions, the arithmetic mean looses much of its efficiency. Then the estimator $\overline{\mathbf{x}}$ can be heavily biased and quite misleading.

Therefore, an important advice for the practice of hedonic price indices is, first, to be aware that there is that problem of estimating hedonic qualities at all. And, second, before estimating the hedonic quality $\mathrm{E}(\mathbf{x})$ of a good, to look at the multivariate distribution of the characteristics vector $\mathbf{x}$ and check for outliers. The detection of multivariate outliers is by no means a trivial task. For modern outlier detection methods see, e.g., [8].

Furthermore, there are well-known alternatives to the simple arithmetic mean, e.g. the multivariate median. For practical purposes, the classical device is, to estimate the hedonic quality $\mathrm{E}(\mathbf{x})$ by the multivariate mean and the multivariate median. When mean and median are quite similar then the $\mathbf{x}$-observations are "well-behaved" and one can proceed to estimate the parameter vector of the hedonic regression. If this is not the case, the data should be scrutinized further for outliers. Depending on their character they should be removed or not.

Estimation of Hedonic Parameters

The third statistical problem involved in the concept of a true hedonic price index is the problem of estimating the parameter vector $\boldsymbol{\beta}$ of a hedonic equation. In the literature on hedonic price indices, this problem usually is regarded as a standard econometric estimation problem. "From the statistical or econometric point of view, there is nothing very complicated about estimating hedonic functions" [20, p. 64].

In that vein, in the model (16) for the data it is simply, more or less implicitly, assumed that for the error term $\mathbf{u}$ the classical assumptions where the ordinary least squares (OLS) method is best hold. I.e., it is assumed that, given the matrix $\mathbf{X}$, for the conditional expectation of the error term holds $\mathrm{E}(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$ and for its conditional covariance matrix $\mathrm{Cov}(\mathbf{u} \mid \mathbf{X}) = \sigma^2\mathbf{I}$. Thus the conditional errors are assumed to be uncorrelated and homoscedastic, i.e. to have equal variances. The OLS-assumptions will, in general, be reasonable in cases where all the price observations are independent. Then, the well-known OLS-estimator is best.

However, in practice, the OLS-assumptions may not be fulfilled and, e.g., at least some of the prices surveyed may be correlated. One reason could be that some of the prices are collected by the same collector. This will, in particular, be a problem when a hedonic adjacent periods index is alternatively estimated in the sense of Sect. 4.2.

In that case, a sample of M variants of the good considered is drawn from the base period population and then all the variants drawn are reconsidered in the comparison period. Then the observations $(p_j^\tau, \mathbf{x}_j^\tau)$, $\tau = 0, t$, $j = 1, \ldots, M$, are stacked together to estimate the parameter vector $\boldsymbol{\beta}$.

I.e., in the model (16), the vector of the price observations is of the form

$$\mathbf{p} = (\mathbf{p}^{0\,\prime}, \mathbf{p}^{t\,\prime})' = (p_1^0, \ldots, p_M^0, p_1^t, \ldots, p_M^t)'$$

and the matrix $\mathbf{X}$ and the error term $\mathbf{u}$ have the forms

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^0 \\ \vdots \\ \mathbf{x}_M^0 \\ \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_M^t \end{pmatrix} = \begin{pmatrix} x_{11}^0 & \cdots & x_{1K}^0 \\ \vdots & & \vdots \\ x_{M1}^0 & \cdots & x_{MK}^0 \\ x_{11}^t & \cdots & x_{1K}^t \\ \vdots & & \vdots \\ x_{M1}^t & \cdots & x_{MK}^t \end{pmatrix} \tag{32}$$

and

$$\mathbf{u} = (\mathbf{u}^{0\,\prime}, \mathbf{u}^{t\,\prime}) = (u_1^0, \ldots, u_M^0, u_1^t, \ldots, u_M^t)' ,$$

respectively. Then the disturbance terms u_j^0 and u_j^t, $j = 1, \ldots, M$, corresponding to observations of the same commodity j, in general, will be correlated and the matrix $\mathrm{Cov}(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{\Sigma}$ may no more be diagonal but of the form

$$\sigma^2 \mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \cdots & 0 & \varrho_1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & \varrho_M \\ \varrho_1 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \varrho_M & 0 & \cdots & 1 \end{pmatrix} , \tag{33}$$

where $\varrho_j = \mathrm{E}(u_j u_{j+M})/\sigma^2$, $j = 1, \ldots, M$, denotes the correlation between two disturbance terms corresponding to the same commodity, sampled at different times 0 and t.

In such cases the OLS-assumptions are no more fulfilled. If, nevertheless, the OLS-method is used for estimating the parameter vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)'$, it is well-known that the estimation is still unbiased but no more efficient. The efficient estimator, then, is the generalized least squares (GLS-) estimator.

5 Conclusion

In this paper, a statistical theory of hedonic price indices has been developed. This theory, first, elaborates the basic ideas underlying the current statistical practice of hedonic price indices. It specifies precise hedonic notions of a good and its price as well as clear-cut definitions of true hedonic price indices. These specifications provide a general framework within which a careful analysis of the current hedonic practice is possible.

Within this framework, e.g., it becomes clear that the widely applied adjacent periods price index can be regarded as an index which is, in a certain sense, a kind of compromise index between classical Laspeyres and Paasche type approaches. In that sense it corresponds to well-known compromise index formulae like the Fisher or the Thørnquist index.

The analysis carried out in this paper shows that the well-known hedonic price index formulae actually are estimations of well-defined theoretical concepts, the true hedonic price indices. It elucidates that there are, in fact, three main statistical problems involved in these concepts: The problem of specifying a good, and the problems of estimating a hedonic quality as well as the parameter vector of a hedonic equation.

Unless these problems are well-known in general statistical theory, this analysis shows further that each of these problems, can be a serious one in practical applications of hedonic concepts. Furthermore, it becomes clear that the techniques used in practice for estimating hedonic indices are just first approaches. More statistical and econometric efforts are needed.

Sometimes, in official statistics, it is deplored that there is no international "standard procedure" to calculate a hedonic price index. Some even argue that, for them, this is a necessary prerequisite before routinely applying hedonic price index concepts.

The theoretical statistical analysis of the concept of hedonic price indices carried out in this paper, finally, shows that this hope will, at least in a narrower sense, be in vain. A careful realization of a hedonic price index will forever require a careful exploration of the specific case under consideration. This exploration necessitates a sound practical knowledge of the good considered as well as more than average econometric and statistical competence. One would render the dissemination of the hedonic idea a disservice when the calculation of a hedonic index, generally, would be regarded as a simple standard estimation problem.

The calculation of hedonic price indices is and will remain a necessary but highly challenging statistical task.

References

1. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, New York (1994).
2. Berndt, E.R.: **. . . .** In: Brachinger, H.W., Diewert, E.: Hedonic Methods in Price Statistics: Theory and Practice. Springer, Heidelberg (2002)
3. Berndt, E.R., Griliches, Z.: Price Indexes for Microcomputers: An Exploratory Study. In: Foss, M.F., Manser, M.E., Young, A.H. (eds) Price Measurements and their Uses: Studies in Income and Wealth. University of Chicago Press, Chicago (1993)
4. Berndt, E.R., Griliches, Z., Rappaport, N.: Econometric Estimates of Price Indexes for Personal Computers in the 1990's. Journal of Econometrics, **68**, 243–268 (1995)
5. Berndt, E.R., Rappaport, N.: Price and Quality of Desktop and Mobile Personal Computers: A Quarter-Century Historical Overview. American Economic Review, **91**, 268–273 (2001)
6. Boskin, M.J., Dulberger, E.R., Gordon, R.J., Griliches, Z., Jorgensen, D.: Toward a More Accurate Measure of the Cost of Living. U.S. Government Printing Office, Washington (1996)

7. Cole, R., Chen, Y.C., Barquin-Stolleman, J.A., Dulberger, E., Helvacian, N., Hodge, J.H.: Quality-Adjusted Price Indexes for Computer Processors and Selected Peripheral Equipment. Survey of Current Business, **66**, 41–50 (1986)
8. Davies, P.L., Gather, U.: The Identification of Multiple Outliers. Journal of the American Statistical Association, **88**, 782–792 (1993)
9. Diewert, E.: **. . . .** In: Brachinger, H.W., Diewert, E.: Hedonic Methods in Price Statistics: Theory and Practice. Springer, Heidelberg (2002)
10. Griliches, Z.: Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change. In: Stigler, G. (chairman) The Price Statistics of the Federal Government. Columbia University Press, New York (1961)
11. Griliches, Z.: Hedonic Price Indexes Revisited. In: Griliches, Z. (ed) Price Indexes and Quality Change: Studies in New Methods of Measurement. Harvard University Press, Cambridge (1971)
12. Harhoff, D., Moch, D.: Price indexes for PC database software and the value of code compatibility. Research Policy, **26**, 509–520 (1997)
13. Hoffmann, J.: Hedonic Methods for Assessing the Accuracy of Price Statistics: Evidence from Microwave ovens in Germany. In: Brachinger, H.W., Diewert, E.: Hedonic Methods in Price Statistics: Theory and Practice. Springer, Heidelberg (2002)
14. Moulton, B.R.: The Expanding Role of Hedonic Methods in the Official Statistics of the United States. In: Brachinger, H.W., Diewert, E.: Hedonic Methods in Price Statistics: Theory and Practice. Springer, Heidelberg (2002)
15. Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H., Lee, T-C.: The Theory and Practice of Econometrics. Wiley, New York, 2nd edn. (1985)
16. Lancaster, K.J.: A New Approach to Consumer Theory. Journal of Political Economy, **74**, 132–157 (1966)
17. Rosen, S.: Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, **82**, 34–55 (1974)
18. Rousseeuw, P.J., van Zomeren, B.C.: Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, **85**, 633–639 (1990)
19. Triplett, J.E.: Hedonic Methods in Statistical Agency Environments: An Intellectual Biopsy. In: Berndt, E.R., Triplett, J.E. (eds) Fifty Years of Economic Measurement. University of Chicago Press, Chicago (1990)
20. Triplett J.E.: **. . . .** In: Brachinger, H.W., Diewert, E.: Hedonic Methods in Price Statistics: Theory and Practice. Springer, Heidelberg (2002)