# The Making of Hedonic Index Numbers

*Auno, Ville, Statistics Finland*
*Luomaranta-Helmivuo, Henri, Statistics Finland*
*Markkanen, Hannele, Statistics Finland*
*Montonen, Satu, Statistics Finland*
*Nieminen, Kristiina, Statistics Finland*
*Suoperä, Antti, Statistics Finland*

## Abstract

This study combines heterogeneously behaving cross-sectional regressions and hedonic quality adjusting in traditional index number framework. The approach provides a transparent mathematical representation of quality correction and quality adjustment of price changes in elementary aggregates. We propose an alternative to the standard Griliches-type time-dummy hedonic approach, which in the sense of index number theory is more interpretable and mathematically transparent between actual average price changes, quality correction and quality adjustment.

In the first stage, the problem of heterogeneously behaving cross-sectional models is handled using the principle of hierarchical, 'nested', price models. The price models are formulated by combining the proper partition of observations (categorization of observations) and the proper classification of observations into the most homogeneously behaving subgroups (heterogeneous between subgroups) using standard statistical inference. These are achieved using the FE-models (fixed effects) familiar to economists. In the second stage, the estimated price models are aggregated from observation level into the level of partition (i.e., into stratums), where the so-called Oaxaca decompositions are computed. This decomposition, although not unambiguous, consistently divides the actual price change into quality corrections and quality adjusted price change for each stratum. We show what is the ideal selection of decompositions based on the algebraic properties of the OLS method. In the third stage, the stratum level decompositions are aggregated into higher levels similarly as in a traditional index number calculation where 'a weighted-by-economic-importance'-variable takes a central role. We use several basic and excellent index number formulas. The study ends in empirical application of used cars in Finland.

**Keywords**
Partition, Unit Value, Logarithmic Representations, Index Number Formulas, Hedonic Method, FE-Model, OLS Method, Unbiased, Price Aggregation, Oaxaca Decomposition, Logarithmic mean, Conditional and Unconditional mean.

# 1 Introduction

In traditional index number theory direct price-links are based on comparisons $0 \rightarrow t$, $t = 1, 2, \ldots$, for commodities comparable in quality. Practically this means measurement of price changes from commodity prices having a unique code e.g. GTIN-identifier. This traditional method fits nicely for e.g. daily products but not generally. In most cases, like clothes, shoes, mobile phones, TV, home electronics etc., bilateral price-linking is not possible because of quality change. This property makes bilateral strategies less useful leading to indices being contingently biased caused by quality changes of quality characteristics. This happens for example for prices of houses and used cars. For that Bailey, Muth and Nourse (1963) developed a repeat-sales model (see, Case and Shiller,1989; Quigley, 1995) using a model based (or the stochastic) approach to measure changes of prices. These repeat-sales models are problematic, because they can capture a tiny fraction of the data because each transacted 'commodity', for example apartment or used car, appears rarely more than once in the data in a short time span. Another well-known model-based approach is the Griliches (1971) time-dummy hedonic method or the WTPD-model (Diewert and Fox, 2018, pp.15), which cover the entire data and resolve the comparability issue using hedonic quality adjusting. These methods suffer from several problems, but most importantly they are not connected any way with traditional index number theory (see Koev, 2003; Suoperä, Luomaranta, Nieminen and Markkanen. 2021; Kaila, Luomaranta & Suoperä, 2022). Therefore, these hedonic methods are abandoned in this study.

The focus of the study is to show 'How hedonic quality adjusting, and traditional index number theory may be combined using familiar regression analysis and its algebraic properties transparently?'. The work builds on two earlier papers (Koev, 2003; Suoperä, 2006; see also Vartia, Suoperä & Vuorio, 2021; Suoperä & Auno, 2021; Suoperä, Luomaranta, Nieminen and Markkanen. 2021; Kaila, Luomaranta & Suoperä, 2022) which address most of issues based on hedonic approach to index numbers. The main idea is that because effective matched pairs method or bilateral price-linking is not possible, the price-linking should be done for some coarse but the most homogeneous grouping of observations. We do this using econometric approach where price models include two-dimensional heterogeneity: 'intercept' or 'categorical heterogeneity' that arise from a detailed partition and 'slope coefficient heterogeneity' from different OLS regressions in several heterogeneously behaving subgroups (Suoperä and Vartia, 2011). In statistical textbooks this modelling is a well-known Fixed Effects (FE) model (Hsiao, 1986, s.29-32).

The process consists of three steps. In the first step, we define several hierarchical 'nested' FE price models and use statistical inference, that is the estimation of heterogeneously behaving price models and testing equality between them. Statistical inference helps us to identify the data generating process of prices and leads to selection of the best price models, that is the combination of the classification of price models and their partitions. Estimators of the price models are the best linear unbiased estimates (BLUE). In second step, we aggregate price models from observations into stratums of the partition. This will be done while satisfying the basic algebraic properties of the OLS method. Then the quality adjusting is performed using decomposition introduced by Oaxaca (1973). Even the decomposition is not unambiguous, it splits the true average price change consistently into quality changes and quality adjusted price changes for any stratum in question. In third step, we apply traditional index number theory for stratum level aggregates of the decomposition. We analyze two stratum aggregates and their decompositions – unweighted arithmetic and geometric averages. We perform our analysis of index numbers using several basic (Laspeyres (L), log-Laspeyres (l), Log-Paasche (p), Paasche (P)) and excellent index number formulas (Törnqvist (T), Montgomery-Vartia (MV), Sato-Vartia (SV), Fisher (F)).

The structure of the study is as follows. In chapter 2 we present the data, basic concepts and notations. In chapter 3 we present several nested partitions and combine them with heterogeneously behaving cross-sectional regressions. Theoretical methods are presented by their empirical counterparts. In chapter 4 we derive stratum aggregates and their Oaxaca decompositions. In chapter 5 we apply index number methods to our stratum aggregates and show some graphical figures comparing different basic and excellent index numbers. Chapter 6 concludes.

# 2 Data, Basic Concepts and Notation

## 2.1 Data

Data is received on a daily basis from one major selling portal for second-hand cars in Finland. The received data contains the sales announcements updated on the previous day. When daily announcements are compiled as monthly data, only the latest sales announcement of the month is considered. The sales announcement data is then supplemented with additional characteristics information from the vehicle register data from Finnish Transport and Communications Agency. If the weight or the power of the car are unavailable from abovementioned sources, they are imputed. The monthly data contains approximately 75 000 individual sales announcements of second-hand cars.

For index calculation purposes, only second-hand cars with ”sold”-status purchased from car dealers are taken into account. Second-hand cars aged between one and twenty years are taken into index calculation. Cars with price less than 2000 euros are excluded since they are not considered representative. Vans and recreational vehicles are deleted from index calculation data. Cars with outliers or clearly incorrect information in the categorical variables (such as mileage over one million kilometers, weight under 750 kilograms or over 3000 kilograms and power under 20 kilowatts or over 600 kilowatts) are also removed. Also, cars with mileage under one kilometer are deleted since they are not considered as second-hand cars.

## 2.2 Basic Concepts

Price is defined as car specific *unit value* measuring *price of a car*. In this study, the unit prices are in logarithmic scale, log-euros. All other variables are measured by their typical units of measurement, e.g. age of the car in years, selling time of the cars in months, and mileage in kilometers. Non-linearity is taken into account by calculating square roots of those explanatory variables that are not dummy variables. In short, our price model is specified as semilogarithmic.

## 2.3 Notation

The notations in this study are two-fold. First, in observation level we use typical econometric notation because we use model-based price analysis. Aggregation of variables (i.e., dependent, independent) from observations into strata (i.e., into index commodities or stratum aggregates) connect notations also into traditional notations of index number theory. The most important concepts are:

Observation level:
Commodities: $a_1, a_2, \dots, a_{n_t}$ are transacted used cars in period *t*.
Time periods: $t = 0, 1, 2, \dots$ are the compared months.
Quantity: $q_i^t = q_{it} = 1$ for $a_i$ in period *t*.
Unit value or unit price: $p_i^t = v_i^t / q_i^t$ or $p_{it} = v_{it}/q_{it}$ is the unit price of a used car $a_i$ in period *t*
Value: $v_i^t = v_{it} = q_{it} p_{it}$ is the value of a used car $a_i$ in period *t*.
Total value: $V^t = \sum_i v_i^t = \sum_i v_{it}$ is the total value of all used cars in period *t*.
Total quantity: $Q^t = \sum_i q_i^t = \sum_i q_{it}$ is the total quantity of all used cars in period *t*.
Explanatory variables in regressions: $\boldsymbol{x}_{it} = (x_{it1} \dots x_{itk})'$ is a *k*-vector of observed characteristics in period *t*.

Stratum level (i.e., elementary aggregates, for example conditional averages):
Price relatives: $\bar{p}_k^{t/0} = \bar{p}_{kt}/\bar{p}_{k0}$ is the price relative of averaged unit prices for stratum *k* from period 0 to *t*.
Quantity relatives: $q_k^{t/0} = q_{kt}/q_{k0}$ is the quantity relative for stratum *k* from period 0 to *t*.
Value relatives: $v_k^{t/0} = v_{kt}/v_{k0}$ is the value relative for stratum *k* from period 0 to *t*.
Value shares: $w_{kt} = v_{kt}/\sum_k v_{kt}$ is the value share for stratum *k* in period *t*.

Explanatory variables in regressions: $\bar{\boldsymbol{x}}_{kt} = (\bar{x}_{t1} \ldots \bar{x}_{tk})'$ is a *k*-vector of averaged characteristics for stratum *k* in period *t.*

## 3 The Regression Analysis

We underline the importance of the analysis of heterogeneous micro behaviors that includes two main sources of heterogeneity – intercept or categorical heterogeneity (problem of partition) and slope heterogeneity from different OLS regressions. Inadequate partition or inadequate classification of price models, or both, lead to biased estimates of the OLS regressions caused by omitted relevant variables. We analyze this problem using several hierarchical partitions of observations and several classifications of the nested OLS regressions.

Partition means for most statisticians the classification of statistical units into most 'homogenous' disjoint stratums. 'Homogeneous groupings' are not easy to come by. In this study, we use statistical inference to solve problem of partition. The same principle is used also in the decision-making of the classification of price models. Together they make possible to control quality differences of the characteristic's variables, that is $\boldsymbol{x}_{it} = (x_{it1} \ldots x_{itk})'$, inside stratum *k* and time periods $t \neq t'$.

We proceed similarly as in Suoperä and Vartia (2011) – we make partition of transacted used cars and then apply regression analysis for some subgroup of stratums included in partition. We combine them into fixed-effects dummy-variable approach (Hsiao, 1986, s.29-32). We show that regression analysis combined with the partition is operational especially in construction of hedonic index numbers (Koev, 2003; Suoperä, 2006; see also Vartia, Suoperä & Vuorio, 2021; Suoperä & Auno, 2021; Suoperä, Luomaranta, Nieminen and Markkanen. 2021; Kaila, Luomaranta & Suoperä, 2022).

We give simple examples how to make hierarchical 'competing price models' that combine intercept/categorical and slope heterogeneity into the FE models. We also show how to select the best price model for our hedonic quality adjusting using simple statistical inference for these 'nested models'. Following Table shows two sources of heterogeneity for used cars.

Table 3.1: Two heterogeneity effects on price levels and price differences.

| **Intercept/categorical heterogeneity** | | | | | |
|---|---|---|---|---|---|
| **Partition 1** | **Partition 2** | **Partition 3** | **Partition 4** | **Partition 5** | **Partition 6** |
| No partition | Size of a car | Size of a car × Make | Size of a car × Make × Model | Size of a car × Make × Model × Driving Power | Size of a car × Make × Model × Driving Power × Type of a car |
| **Slope heterogeneity categories** | | | | | |
| **Naive** | | **Typical** | | **Good or 'Best'** | |
| No heterogeneity | | Size of a car | | Size of a car × Make | |

Size of a car-indicator is formed with internationally used segment-variable which classify cars into standard, SUV[1]- and MPV[2]- cars according to seven size categories from M to F. We group them into following four size categories: {M, A, B} ('Small), {C} (Normal), {D} (Big) and {E, F} (Maximum), which each includes their SUV- and MPV-models. SUV- and MPV-models are included into categorization by separate indicators, that are formed using 'Make' and 'Model' information. 'Make'-indicator classifies cars into, e.g. 'Audi', 'BMW', 'Ford' and its 'Model' into e.g. 'A4', 'Series-5', 'Focus'. Indicator 'Driving Power' classify cars into five categories: Diesel, Electric, Hybrid, Gasoline and Others. 'Type of a car'-indicator into estate and other

[1] SUV=sport utility vehicle
[2] MPV=multi-purpose vehicle

type cars. All indicators and their cartesian product, i.e. '×' in Table 3.1, form partition of disjoint sets with union of all observations.

In Table 3.1, we define six competing partitions and three different specification of slope heterogeneity. We proceed using following three steps: In first step, we combine 'naïve' model with all five partitions, estimate them separately and test the equality between them hierarchically (i.e., Partition 1 vs. Partition 2, Partition 2 vs. Partition 3, …). This step concludes the best partition in statistical sense. In second step, a naïve model is replaced by four equations based on 'Size of a car' categories, which are combined with the best partition selected in the first step. Price model from steps one and two are 'nested models' (certain linear restrictions on model two leads into model one) and their equality may be tested using standard F-statistics. This test is a measure of the loss of fit those results from imposing a linear restriction on price models of step two (see Greene, 1997, p. 343-344, 657). In third step, we estimate about 70 equations based on size and make of a car that are combined with the best partition selected in step one and two. The price models selected in each step (i.e., step one, two and three) are nested hierarchical models and their equality may be tested using the same F-test as before (see example: Suoperä and Vartia, 2011, p.21).

## 3.1 The Price Model for Heterogeneously Behaving Cross-sections

We start the analysis using the standard linearly additive price model in its most general representation:

(1) $$y_{ijt} = \sum_{k=1}^{K_j} i_{ikt}\alpha_{kt} + \boldsymbol{x}'_{ijt}\boldsymbol{\beta}_{jt} + \varepsilon_{ijt},$$

where the dependent variable $y_{ijt} = \log(p_{ijt})$ is a log-price for statistical unit *i* belonging into equation *j* in time period *t*. $\boldsymbol{x}_{ijt}$ is a *E*-dimensional vector of explanatory variables for equation *j* in time period *t*. $\boldsymbol{\beta}_{jt}$ is a *E*-dimensional vector of parameters presenting of mean changes in the log-prices y from a unit changes of $\boldsymbol{x}$. The explanatory variables are measured in their original units of measurements meaning that equation (1) is specified as semilogarithmic. Each equations includes $K_j$ categorical indicator or dummy variables (i.e., size of a car, make, model, driving power, type of a car) $i_{ikt}$ that gets value 1 if belongs into certain category otherwise 0. The categorical variables form the partition of observations for any equation *j*.

The price model is defined in its most general form because the sources of heterogeneity may be easily presented. Using simple algebra, the equation (1) may be represented as a sum of representative and deviation behaviors (heterogeneity effects):

(2) $$y_{ijt} = \bar{\alpha}_t + \boldsymbol{x}'_{ijt}\overline{\boldsymbol{\beta}}_t + \sum_{k=1}^{K} i_{ikt}(\alpha_{kt} - \bar{\alpha}_t) + \boldsymbol{x}'_{ijt}(\boldsymbol{\beta}_{jt} - \overline{\boldsymbol{\beta}}_t) + \varepsilon_{ijt},$$

where the representative behavior is $\bar{\alpha}_t + \boldsymbol{x}'_{ijt}\overline{\boldsymbol{\beta}}_t$ and two sources of heterogeneity behaviors, that are categorial $\sum_{k=1}^{K} i_{ikt}(\alpha_{kt} - \bar{\alpha}_t)$, $k = 1,\ldots, K$ (number of categories/stratums) and behavioral heterogeneity $\boldsymbol{x}'_{ijt}(\boldsymbol{\beta}_{jt} - \overline{\boldsymbol{\beta}}_t)$. Interpretation of these two terms is presented in Vartia, (1979, 2008a); Suoperä and Vartia (2011, pp.6) and may be noted simply as

Categorial: $i_{ikt}(\alpha_{kt} - \bar{\alpha}_t) = c_{ikt}$, for $k = 1,\ldots, K$ and
Behavioral: $\boldsymbol{x}'_{ijt}(\boldsymbol{\beta}_{jt} - \overline{\boldsymbol{\beta}}_t) = \boldsymbol{b}_{ijt}$ for $j = 1,\ldots, J$.

Before empirical solution of (2) we put all things together using deterministic mathematics and matrix notations for equation (2), that is

(3a) $$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta}^*_t + \boldsymbol{H}_t\mathbf{1}_t + \boldsymbol{\varepsilon}_t, \text{ where } \boldsymbol{H}_t = [\,\boldsymbol{C}_t\ \boldsymbol{B}_t]$$

or more compactly as

(3b) $$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\phi}_t + \boldsymbol{\varepsilon}_t, \text{ where } \boldsymbol{Z}_t = [\boldsymbol{X}_t\ \boldsymbol{H}_t\,] \text{ and } \boldsymbol{\phi}_t = (\boldsymbol{\beta}^{*\prime}_t\ \mathbf{1}'_t\,)', \text{ where } \boldsymbol{\beta}^{*\prime}_t = (\alpha_t\ \boldsymbol{\beta}_t)'$$

$\boldsymbol{y}_t$ is $N_t$-vector of log-prices, $\boldsymbol{X}_t$ is $(N_t * (E+1))$-matrix having unity vector in the first column (constant) and rest columns are the $E$ explanatory variables. $\boldsymbol{H}_t$ matrix includes two heterogeneity matrices - $\boldsymbol{C}_t$ is $(N_t * K)$-matrix including categorial heterogeneity covariates and $\boldsymbol{B}_t$ is $(N_t * E)$-matrix including behavioral slope heterogeneity covariates, that is

$$\begin{bmatrix} y_{1t} \\ \vdots \\ y_{N_t t} \end{bmatrix}, \boldsymbol{X}_t = \begin{bmatrix} 1 & x_{11t} & \cdots & x_{1Et} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{N_t 1t} & \cdots & x_{N_t Et} \end{bmatrix}, \boldsymbol{C}_t = \begin{bmatrix} \boldsymbol{c}_{1t} & \mathbf{0} & \ldots & & \mathbf{0} \\ 0 & \boldsymbol{c}_{2t} & \mathbf{0} & \ldots & \vdots \\ \vdots & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & & \mathbf{0} & \boldsymbol{c}_{Kt} \end{bmatrix}, \boldsymbol{B}_t = \begin{bmatrix} \boldsymbol{b}_{11t} & \cdots & \boldsymbol{b}_{1Et} \\ \vdots & \ldots & \vdots \\ \boldsymbol{b}_{J1t} & \cdots & \boldsymbol{b}_{JEt} \end{bmatrix}$$

It is true that the estimation of equation (2) and (3) is impossible or at least difficult. Next, we show how it can be done using the OLS method. Looking carefully, the analysis from (1) to (3), one may understand our idea - the method reproduces separately specified price equations exactly in the observation level, but now in the mean-deviation re-parameterized form (3). The first part of it consists of the common behavior described by the mean parameter part of the equation and the second part the heterogeneity effects described by the covariates.

## 3.2 The OLS solution for Heterogeneously Behaving Cross-sections

The price models (1) are familiar Fixed Effects models (FE) (Hsiao, 1986, s.29-32) that we specify as semilogarithmic. The price equations for log-prices are specified as non-linear with respect to age of a car (years), mileage (ten thousand), power/weight ratio of a car and selling time (months). All explanatory variables of eq. (1) are listed in Table 3.2.

Table 3.2: The exogenous variables used in the price models for used cars in Finland.

| Variable | Description |
|---|---|
| Categorical variables | Size of a car × Make × Model × Driving Power × Type of a car or some special cases of these categorial variables (see Table 3.1). The size of a car is determined using international segment-variable:<br>Small cars: Segment = {'A', 'A_SUV', 'B', 'B_MPV', 'B_SUV', 'M'}<br>Normal cars: Segment = {'C', 'C_SUV', 'C_MPV'}<br>Big cars: Segment = {'D', 'D_SUV', 'D_MPV'}<br>Maximum size cars: Segment = {'E', 'E_MPV', 'E_SUV', 'F'} |
| $x_1$ | Gearbox type: If automatic $x_1 = 1$, else $x_1 = 0$. |
| $x_2$ | Towing hook: If towing hook $x_2 = 1$, else $x_2 = 0$. |
| $x_3$ | Service history: If service history is available $x_3 = 1$, else $x_3 = 0$. |
| $x_4$ | Cruise control: If cruise control $x_4 = 1$, else $x_4 = 0$. |
| $x_5$ | Selling time of a car, months. |
| $x_6 = sqrt(x_5)$ | Square root of the selling time of a car. |
| $x_7$ | Age of a car, years. |
| $x_8 = sqrt(x_7)$ | Square root of the age of a car. |
| $x_9$ | Mileage (ten thousand). |
| $x_{10} = sqrt(x_9)$ | Square root of mileage. |
| $x_{11}$ | Power/Weight ratio of a car. |
| $x_{12} = sqrt(x_{11})$ | Square root of Power/Weight of a car. |

It is assumed, that $E(\varepsilon_{ijt}|\boldsymbol{x'}_{ijt}) = 0$ and $Var(\varepsilon_{ijt}|\boldsymbol{x'}_{ijt}) = \sigma_{jt}^2 < \infty$ and the error covariance matrices are diagonal for all $j$ =1,…, $J$ (number of equations) . Practically this means that the OLS estimation assumes homoscedastic, uncorrelated model errors with zero mean for all equations - normality of the model errors is not necessary for parameter estimation. According to the Frisch, Waugh and Lovell -theorem (Davidson & MacKinnon, 1993), the OLS –estimation of the slopes can always be carried out via categorially centralized variables. The constant term for category/stratum $k$ is estimated by forcing the regression plane through the point of averages, that is

$$\widehat{\boldsymbol{\beta}}_{jt} = \left[\sum_i \sum_k (\boldsymbol{x}_{ikjt} - \overline{\boldsymbol{x}}_{kjt})(\boldsymbol{x}_{ikjt} - \overline{\boldsymbol{x}}_{kjt})'\right]^{-1} \sum_i \sum_k (\boldsymbol{x}_{ikjt} - \overline{\boldsymbol{x}}_{kjt})(y_{ikjt} - \bar{y}_{kjt})$$
$$\hat{\alpha}_{kt} = \bar{y}_{kjt} - \overline{\boldsymbol{x}}'_{kjt}\widehat{\boldsymbol{\beta}}_{jt},\ k \in j$$

This method is computationally extremely effective especially when partition includes hundreds/thousands of categories/strata (see Suoperä & Vartia, 2011). After estimation of (1) for all $j$ we may construct equations (2) and (3) and estimate them using the OLS method. These estimated models, based on the mean-deviation re-parameterization, are mathematically exactly equal in all arguments compared with the price equation (1) together taken – even the residuals are equal observation by observation. This is a known result mentioned shortly e.g., by Balestra and Nerlove in their introduction in Matyás and Sevestre (1996). They just simply state that the total sum of squares of one large seemingly unrelated regression model (SUR) reduces to the sum of squares summed over the equations. This means, that the separately estimated price equations by the OLS method are in fact equivalent to one large SUR estimation with diagonal covariance matrix. Therefore, minimizing the sum of squared residual first in the equation level is equivalent to the minimizing all of them at the same time in the mean-deviation re-parameterized form for all observations as a whole. So, the estimation of the price equation (3) reproduces exactly the average OLS-estimates and the unity coefficients (i.e., $\hat{1}_t = 1_t$) for the covariances. The re-parameterization has a more central goal – the model (3) can be used to estimate the variance-covariance matrix for the estimates of the model (2) or (3). We end our analysis and show the variance-covariance matrix for the estimator of the model (3) by the OLS method. We know that the slope coefficients or linear estimator $\boldsymbol{\phi}_t$ is a linear function of disturbances. When we have no stochastic $\boldsymbol{Z}_t$, that is $E(\boldsymbol{\varepsilon}_t|\boldsymbol{Z}_t) = \boldsymbol{0}$, regardless of the distribution of $\boldsymbol{\varepsilon}_t$, the OLS estimator $\widehat{\boldsymbol{\phi}}_t$ is a best linear, unbiased estimator of $\boldsymbol{\phi}_t$ and its variance-covariance estimator is

(4) $Var(\widehat{\boldsymbol{\phi}}_t) = \sigma_t^2(\boldsymbol{Z}_t{}'\boldsymbol{Z}_t)^{-1}$, where

$$= \sigma_t^2 \begin{bmatrix} (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}(\boldsymbol{I}_t + \boldsymbol{X}_t'\boldsymbol{H}_t\boldsymbol{R}_t\boldsymbol{H}_t'\boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}) & (\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{H}_t\boldsymbol{R}_t \\ -\boldsymbol{R}_t\boldsymbol{H}_t'\boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1} & \left(\boldsymbol{X}_t'\boldsymbol{X}_t - \boldsymbol{X}_t'\boldsymbol{H}_t(\boldsymbol{H}_t'\boldsymbol{H}_t)^{-1}\boldsymbol{H}_t'\boldsymbol{X}_t\right)^{-1} \end{bmatrix}$$

where $\boldsymbol{R}_t = \left(\boldsymbol{H}_t'\boldsymbol{H}_t - \boldsymbol{H}_t'\boldsymbol{X}_t(\boldsymbol{X}_t'\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t'\boldsymbol{H}_t\right)^{-1}$. This is a new result by which we may look at not only significance of parameters of representative behavior but also significance of any single heterogeneity variables, categorial and behavioral covariates that otherwise is impossible. We show some important properties of (4) when significant categorial or/and behavioral heterogeneity components are deleted. The whole mathematical and statistical story of this chapter is shown in Suoperä and Vartia (2011).

## 3.3 Statistical inference of Price Models

Now we turn into empirical analysis where we use statistical inference in selection of the best price model for hedonic quality adjusting. We proceed above mathematical/statistical analysis in spirit of the Table 3.1: First, we make statistical inference about partition/categorization of observations restricting behavioral slope heterogeneity $\boldsymbol{\beta}_{jt} = \overline{\boldsymbol{\beta}}_t$ for all $j$. We get four hierarchical tests about five different partitions and select the best one. Second, we relax the restriction $\boldsymbol{\beta}_{jt} = \overline{\boldsymbol{\beta}}_t$ and estimate price models according to three slope heterogeneity categories using the best partition/categorization selected in the first stage. We get two hierarchical tests about three different slope heterogeneity modelling and select the best one.

The statistical inference - estimation and hypothesis testing - is based on the OLS estimation and hypothesis test on the well-known loss of fit test. We already know that the OLS estimator $\widehat{\boldsymbol{\phi}}_t$ is a best linear, unbiased estimator of $\boldsymbol{\phi}_t$ that is chosen to minimize the sum of squared errors, SSE. Because the coefficient of determination $R^2$ equals with 1 – SSE/SST, where the SST $= \sum_i (y_{it} - \bar{y}_t)^2$, the OLS estimator is in fact selected to maximize $R^2$. This is the reason for our test – loss of fit.

Now we go back to Table 3.1 and give necessary statistics for testing equality of price models, that is, number of observations ($N_t$), categories ($k$), equations ($J$), restrictions ($R$), decrees of freedom of free model ($Df_t$) and the sum of squared errors ($SSE$). Our tests are based of hierarchic nested price models meaning that the models are nested with each other so that they can be obtained from each other by imposing suitable linear restrictions on parameters. Our test is

$$F \sim \{(SSE_0 - SSE_1)/R\}/(SSE_1/Df_t)$$

where $SSE_0$ is the sum of squared errors of the restricted model, $SSE_1$ is the sum of squared errors of the free model, $Df_t$ is the degrees of freedom of the free model and $R$ is the number of linear restrictions. When the degrees of freedom for free model becomes large the $F$-statistics reduced into $\chi^2_R$-test, where $R$ corresponds number of linear restrictions (see Greene 1997, p. 344 and p. 657). For example, a 1% critical value of $\chi^2_{60} = 1.46$ and becomes closer to one when $R > 60$. Table 3.3 shows necessary statistics for nested price models results for testing the significance of additional partition.

Table 3.3: Testing the hypothesis of the categorial and behavioral homogeneity using hierarchical nested price models in year 2022.

| | **Intercept/categorical heterogeneity** | | | | | |
|---|---|---|---|---|---|---|
| | **Model 1** | **Model 2** | **Model 3** | **Model 4** | **Model 5** | **Model 6** |
| | No categori-zation | Size of a car | Size of a car × Make | Size of a car × Make × Model | Size of a car × Make × Model × Driving Power | Size of a car × Make × Model × Driving Power × Type of a car |
| $N_t$ | 269663 | 269663 | 269663 | 269663 | 269663 | 269663 |
| $k$ | 1 | 4 | 103 | 516 | 1189 | 1691 |
| $J$ | 1 | 1 | 1 | 1 | 1 | 1 |
| *Parameters* | 12 | 12 | 12 | 12 | 12 | 12 |
| *SSE* | 26886 | 22855 | 13545 | 6476 | 5928 | 5812 |
| | | **Model 1 vs 2** | **Model 2 vs 3** | **Model 3 vs 4** | **Model 5 vs 4** | **Model 6 vs 5** |
| *Test statistic* | | 11896 | 1872 | 711 | 36.8 | 10.7 |
| | **Slope heterogeneity categories** | | | | | |
| | **Model 6 'Naïve'** | | **Model 7 'Typical'** | | **Model 8 'Good or Best'** | |
| | No heterogeneity | | Size of a car | | Size of a car × Make | |
| $N_t$ | 269663 | | 269663 | | 269663 | |
| $k$ | 1691 | | 1691 | | 1691 | |
| $J$ | 1 | | 4 | | 74 | |
| *Parameters* | 12 | | 48 | | 888 | |
| *SSE* | 5812 | | 5605 | | 4908 | |

| | | **Model 7 vs 6** | **Model 8 vs 7** |
|---|---|---|---|
| *Test statistic* | | 206.5 | 45 |

Table 3.4: Estimation results for model 7 and 8.

| | Model 8 | Model 8 | Model 7 | Model 7 |
|---|---|---|---|---|
| Year | 2020 | 2021 | 2020 | 2021 |
| Number of observations | 287936 | 269663 | 287936 | 269663 |
| Number of equations | 72 | 74 | 4 | 4 |
| Number of stratums/categories | 1594 | 1691 | 1594 | 1691 |
| Degrees of freedom | 285478 | 267084 | 286294 | 267924 |
| SSE | 5401.6405077 | 4908.43633 | 6096.4446791 | 5604.5913163 |
| R2 | 0.9645034599 | 0.9675392005 | 0.9600517847 | 0.9630515476 |
| RMSE | 0.1375550427 | 0.1355650208 | 0.1459258378 | 0.1446325907 |
| Constant | 9.9126394001 | 9.8211262087 | 9.6028720349 | 9.6497628502 |
| | (0.0125144633) | (0.0118472501) | (0.0132567809) | (0.0126661687) |
| If automatic gearbox $x_1 = 1$, else $x_1 = 0$ | 0.0902673948 | 0.0923941505 | 0.0935819809 | 0.0986927146 |
| | (0.0006280357) | (0.0006591217) | (0.0006661904) | (0.0007021883) |
| If towing hook $x_2 = 1$, else $x_2 = 0$ | 0.0118209506 | 0.0113174535 | 0.0101699236 | 0.010722559 |
| | (0.0005717011) | (0.0005829585) | (0.0006070502) | (0.0006220419) |
| If service history is available $x_3 = 1$, else $x_3 = 0$ | -0.010492392 | -0.008856039 | -0.009808606 | -0.009576151 |
| | (0.0006760757) | (0.0006586455) | (0.0007173066) | (0.0007027478) |
| If cruise control $x_4 = 1$, else $x_4 = 0$ | 0.017682513 | 0.0190084745 | 0.0159907088 | 0.0161078885 |
| | (0.0006925544) | (0.0006978619) | (0.0007368138) | (0.0007456235) |
| Selling time of a car, $x_5$ | -0.000386744 | 0.0036841099 | -0.000090959 | 0.0045121389 |
| | (0.0008966894) | (0.0004936162) | (0.0009512569) | (0.0005266493) |
| $x_6 = x_5^{1/2}$ | 0.0054383443 | -0.012634214 | 0.0047894653 | -0.015270555 |
| | (0.0030867169) | (0.0019822649) | (0.0032745562) | (0.0021148394) |
| Age of a car, $x_7$ | -0.138809764 | -0.135251635 | -0.144926668 | -0.140582936 |
| | (0.0004627876) | (0.0004667166) | (0.0004908363) | (0.0004980448) |
| $x_8 = x_7^{1/2}$ | 0.2915511757 | 0.2950576677 | 0.3143214419 | 0.312142484 |
| | (0.0027085731) | (0.0027962898) | (0.0028720215) | (0.0029842413) |
| Mileage, $x_9$ | -0.033047764 | -0.033221364 | -0.029542445 | -0.03080527 |
| | (0.0001519705) | (0.0001555112) | (0.0001611581) | (0.000165791) |
| $x_{10} = x_9^{1/2}$ | 0.0180405738 | 0.026330353 | -0.001833825 | 0.0129272658 |
| | (0.0011911371) | (0.0012313394) | (0.0012646921) | (0.0013158057) |
| Power/Weight ratio of a car, $x_{11}$ | 12.089654612 | 9.8976375615 | 9.3307834547 | 9.2220969294 |
| | (0.1461774499) | (0.1356128354) | (0.1550967155) | (0.1448799132) |
| $x_{12} = x_{11}^{1/2}$ | -2.549090343 | -1.520907481 | -0.631702081 | -0.671611542 |
| | (0.083681855) | (0.0786039287) | (0.0887347781) | (0.0840297929) |
| HE($c_{kt}$), Categorial heterogeneity | 1 | 1 | 1 | 1 |
| | (0.0009034596) | (0.0009179413) | (0.0009825843) | (0.0010226122) |
| HE($\boldsymbol{b}_{jt}$), Behavioral heterogeneity | 1 | 1 | 1 | 1 |
| | (0.0009001723) | (0.0009163475) | (0.0014366229) | (0.0015525605) |

Parameters for heterogeneity components - HE($c_{kt}$) and HE($\boldsymbol{b}_{jt}$), - are presented by unity parameter. This operation is allowed, because all elements of the ($k$ + $E$)-vector of covariates will estimate into ones and linear combinations of $k$- and $E$-vectors of ones may present by single unity.

Some notes about the Table 3.3 and estimation of equations (1) and (3):

1. A typical FE model is inadequate (model with detailed categories, no slope heterogeneity) and leads into biased estimates and biased quality adjusting in hedonic index numbers. Statistical inference for equations (1) to (3) suggest using most detailed categorial heterogeneity (1691 categories) and slope heterogeneity based on categorization of size of a car and make (74 equations). We call this model as *heterogeneously behaving FE model*.
2. All parameters for explanatory variables in estimation of all $j$ equations (1) will not estimate to statistically significant parameters. We do not exclude these variables because insignificant variables have no systematic significant effects on log-prices and on hedonic quality adjusting (estimation efficiency from exclusion of variables is minimal when decrees of freedom in estimation are large).
3. Statistically and mathematically a single equation (3) coincides precisely the set of $J$ equations – simply saying (3) is precise representation of the set of J equations (1), but now we may derive variance-covariance estimator for $\widehat{\boldsymbol{\phi}}_t$, which is a new result.
4. Equation (3) is mathematically equal with (being different representation of (1)) the set of $J$ equations in (1), where $\widehat{\boldsymbol{\phi}}_t = \left(\hat{\alpha}_t, \widehat{\boldsymbol{\beta}}_t', \widehat{\mathbf{1}}_t'\right)'$. This means: First, that parameters for representative behavior $\hat{\alpha}_t$, $\widehat{\boldsymbol{\beta}}_t'$ are necessarily weighted averages (relative shares as weights) of $\hat{\alpha}_{kt}$, $\widehat{\boldsymbol{\beta}}_{jt}'$. Second, that parameters for the covariates ($c_{ikt}$, $\boldsymbol{b}_{ijt}$) must estimate into ($k$ + $E$)-vector of ones.
5. Estimation of (4) enables us to evaluate standard errors for any parameter of $\widehat{\boldsymbol{\phi}}_t$ – we may estimate separate $t$-statistic for each categorial variable (separate 1691 test for the partition) and for each behavioral covariate variable (here 12) to find significance ones. All behavioral covariate variables may be analyzed in isolation to find 'winners' and 'losers' compared with average representative behavior. This is fine property of (3) and (4), but hard to derive otherwise for heterogeneously behaving cross-sections (heterogeneously behaving slopes).
6. According to the variance-covariance estimator (4) – one may, by exclusion of behavioral heterogeneity, lead to more efficient estimation of parameters, but omitting relevant variables (covariates) leads to estimates being efficient but biased.

Interpretation of estimation results in Table 3.4 are familiar to most statisticians but we repeat them here. Estimate of four first indicator-type $x$-variables, accessories, directly itself tells their effect on log-prices. In equation (1) (or (3)) log-prices are specified for rest of the $x$-variables as non-linear with respect to selling time ($x_5$), age ($x_7$), mileage ($x_9$) and power/weight ratio ($x_{11}$) and additional interpretations are needed. We do this applying partial derivates for the equation (3) with respect to $x_e$-variables where $e$ = 5, 7, 9, 11; that is for example for $x_5$ (other $x$-variables similarly)

$$\partial y_{it}/\partial x_{i5t} = \partial \boldsymbol{Z}_t \boldsymbol{\phi}_t/\partial x_{i5t} = \hat{\beta}_{j5t} + 0.5 * \hat{\beta}_{j6t}/x_{i6t}^{1/2}, \text{ for all } i \in j$$

These partial derivates are evaluated for all observations $i$ and variable $x_e$. We sort these partial derivates according to $x_e$-variables and classify them equidistantly into ordered cohorts. Then we average derivatives cohort by cohort and calculate cumulative sums of them. The results are presented in Figures 3.1 to 3.4 for the $x_e$-variables where $e$ = 5, 7, 9, 11. The approach takes account slope heterogeneity of 'size of a car × Make'-categorization and partial derivates are evaluated at realized points of $x_e$-variables so that we have together more than million partial derivates. The method is transparently interpreted and is based on standard economics.

Figure 3.1: The price effect of selling time (months) on the average log-prices in year 2020 and 2021.

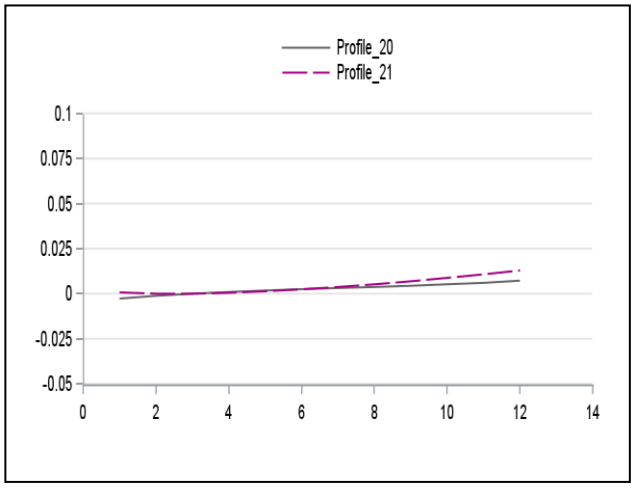



Figure 3.2: The price effect of age (years) on the average log-prices in year 2020 and 2021.

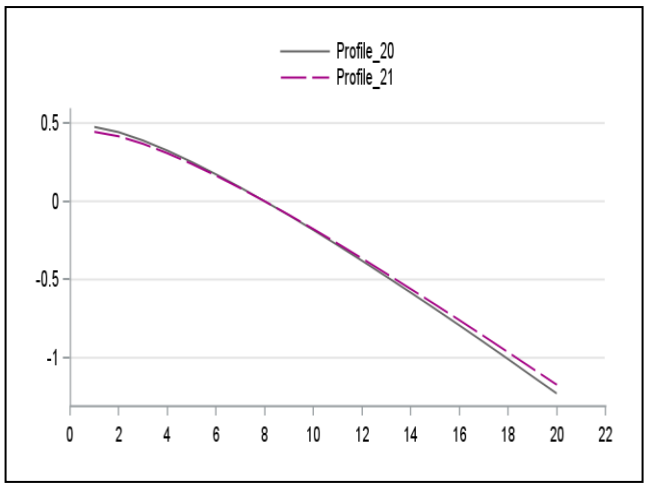



Figure 3.3: The price effect of mileage (ten thousand) on the average log-prices in year 2020 and 2021.

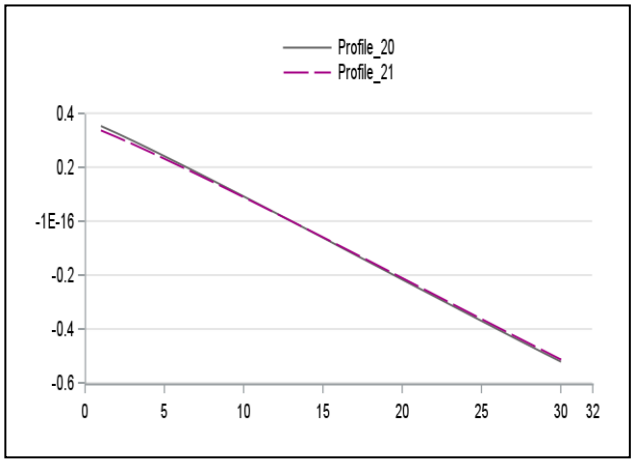
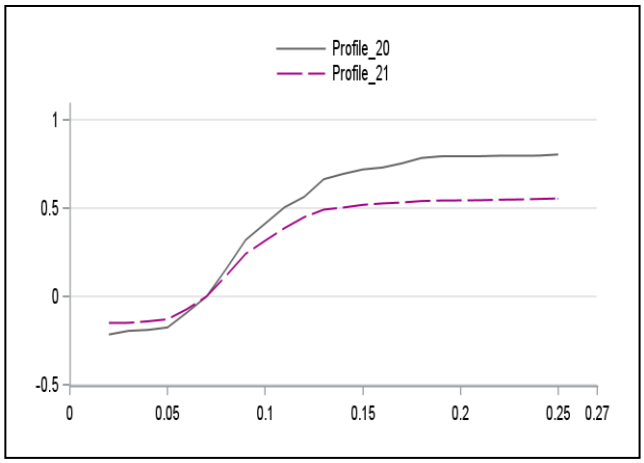



Figure 3.4: The price effect of power/weight ratio (kW/kg) on the average log-prices in year 2020 and 2021.




Figures tell us: Selling time ($x_5$), age ($x_7$) and mileage ($x_9$) behave almost similarly for the years 2020 and 2021 but power/weight ratio ($x_{11}$) not. This is caused by new markets for "plug hybrids" and fully electric cars that are still developing and find more stable practices – it seems that the price effects from high power/weight cars will be declined in time.

We have analyzed the first part of hedonic method – the data generating process of log-prices in heterogeneously behaving gross-sections. Next step in this study continues into the hedonic quality adjusting.

## 4 Combining Regression Analysis and Index Numbers

Classical index calculation is based on bilateral price-links between commodities being comparable in quality – prices and quantities are measured for the same set of commodities and outlets. This means that the price modelling in chapter 3 is unnecessary for bilateral price-links because measured quality characteristics $\boldsymbol{x}_{i0} = \boldsymbol{x}_{it}$ for all 0, $t$ and quality adjusting is not needed. In our case of used cars $\boldsymbol{x}_{i0} \neq \boldsymbol{x}_{it}$ and quality adjusting is necessary. Some notes about our price modelling in Chapter 3 combined with quality adjusting must be done. First, our price modelling is based on optimal solution, the best linear unbiased estimator

(BLUE) under homoscedastic errors for heterogeneously behaving cross-sections. Second, this optimal solution does not only include slope heterogeneity but also optimal solution for partition of observations. The optimal OLS solution does not restrict into the correct size BLU estimates, but other optimal solutions may produce aggregating observations into category/stratum level. These optimal algebraic properties of the OLS are

1. The residuals sum up to zero for all category/stratum.
2. The conditional average equals with unconditional average for all category/stratum.
3. The regression hyperplane passes through the means of dependent and independent variables.

These three properties lead us into the optimal unbiased estimates of unconditional and conditional averages meaning that they both are estimated into the correct size without systematic errors. In our empirical analysis we use two averages – unweighted geometric and arithmetic averages. The aggregation rule for unweighted geometric average is trivial and is presented in most statistical and econometric textbooks. The conditional arithmetic average is more complicated and is presented first in Suoperä (2006, Annex 5, pp.31) and later in Vartia, Suoperä and Vuorio (2019), Suoperä and Vuorio (2019): Suoperä and Auno (2021) and Kaila, Luomaranta and Suoperä (2023). Both averages are unbiased and based on transparent algebra being consistent in aggregation, even aggregation for arithmetic averages are not independent of units of measurement. Our hedonic quality adjusting is based on these conditional and unconditional averages together with a well-known decomposition developed by Oaxaca (1973). Because our price modelling is applied for previous year data, our construction of hedonic index numbers, based on the Oaxaca decomposition, is based on the base strategy which is free of chain drift.

We rely on: *First* the BLU estimates decided by statistical inference, *second*, unbiased conditional and unconditional averages, *third*, mathematically consistent and transparent Oaxaca decomposition even it is not unambiguous, *four*, consistent aggregation rules, *fifth*, drift free construction strategy of indices that are based on hedonic quality adjusting. A well-known time-dummy hedonic regression (see Summers (1973); Rao (2004)) or its weighted version in the sense of Diewert and Fox (2018) have little to do with above mentioned properties – first their link with the traditional index number theory is missing and second the weighted version of Diewert and Fox (2018) leads to parameter estimates whose statistical properties are unknown. We show transparently how these shortcomings may be corrected using well-known basic statistics, consistent aggregation clauses, some algebra, hedonic quality adjusting and several index number formulas and of course unbiased estimators. In our view these are preferable for statistical offices, since the methods are transparent, minimizes modeling assumptions, and are consistent with index number tradition. Our analysis herein follows the tradition of Koev (2003); Suoperä (2004, 2006); Vartia, Suoperä & Vuorio (2021): Suoperä & Auno (2021); Kaila, Luomaranta and Suoperä (2023).

Our focus in the study is three-fold: In the first step, we aggregate estimated equations from observations into category level, stratums. In the second step we make for category/stratum aggregates and their econometric relations a well-known decomposition introduced by Oaxaca (1973). The last step is similar as traditional index numbers – the averaged category/stratum-level price decompositions are summed up using weights of index number formulas, that is '*weights-by-economic-importance*'-variable. We analyze two sets of index number formulae. The first set is based on formulas using old or new weights (asymmetrical weights) and are called as a basic set of index numbers (old weights: Laspeyres (*L*), Log-Laspeyres (*l*) and new weights: Log-Paasche (*p*) and Paasche (*P*)). The second set of index numbers include four formulae using symmetrical weights: Montgomery-Vartia (*MV*), Törnqvist (*T*), Fisher (*F*) and Sato-Vartia (*SV*). We call these index number formulae as *excellent*. For the fundamental analysis of these index number formulae see Vartia & Suoperä, 2018. The analysis therein is in logarithmic form.

## 4.1 Algebra of Price-Ratio Decompositions

We simplify our analysis into two-time case, the base period ($t$ = 0, a previous year) and the observation month of a current year ($t$) analyzing only one stratum $A_k$ belonging into equation $j$. We use vector notations for our conditional and unconditional average prices and calculate the difference between two price models (0, $t$) in spirit of Oaxaca. The algebra for unweighted arithmetic average is based on logarithmic mean, $L$, developed by Leo Törnqvist (1935, p. 35) (see also Y. Vartia, 1976; L. Törnqvist, P. Vartia and Y. Vartia, 1985, p. 44). We use logarithmic mean for aggregation of observations for unweighted arithmetic average (see Suoperä, 2006, pp.31). The algebra is presented here only for unweighted geometric average and its difference but is analogously presented also for unweighted arithmetic average in log-form (see Suoperä (2006, pp.31). We show first necessary weights in aggregation of unconditional and conditional averages and then their Oaxaca decompositions for estimated price models, that is ($n_k$ is number of observations in stratum k)

**Table 4.1**: Important statistics for hedonic quality adjusting for category/stratum *k*.

| Statistics | **Unweighted geometric average** | **Unweighted arithmetic average** |
|---|---|---|
| **Weights** | $w_{ikt} = \frac{1}{n_k}, \forall i \in A_k$ | $w_{ikt} = \frac{L(p_{ikt},1)}{L(\sum_i p_{ikt}, n_k)}, \forall i \in A_k,$ <br> $L$ means logarithmic mean |
| **Unconditional** | $\bar{p}_{kt} = \prod p_{ikt}^{w_{ikt}} = exp\{\sum_i w_{ikt} log(p_{ikt})\}$ | $\bar{p}_{kt} = exp\{\sum_i w_{ikt} log(p_{ikt})\} \equiv \frac{1}{n_k}\sum_i p_{ikt}$ |
| **Conditional** | $log(\bar{p}_{kt}) = \hat{\alpha}_{kt} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{jt}$ | $log(\bar{p}_{kt}) = \hat{\alpha}^*_{kt} + \overline{\overline{\boldsymbol{x}}}'_{kt}\widehat{\boldsymbol{\beta}}_{jt},$ <br> where $\overline{\overline{\boldsymbol{x}}}'_{kt} = \sum_i w_{ikt}\boldsymbol{x}'_{ikt}$ |

Oaxaca decomposition:

(5a) $log(\bar{p}_{kt}) - log(\bar{p}_{ko}) = \hat{\alpha}_{kt} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{jt} - \hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{k0}\widehat{\boldsymbol{\beta}}_{j0} \leftrightarrow$

(5b) $log(\bar{p}_{kt}/\bar{p}_{k0}) = \{(\hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{j0}) - (\hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{k0}\widehat{\boldsymbol{\beta}}_{j0})\} + \{(\hat{\alpha}_{kt} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{jt}) - (\hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{j0})\} \leftrightarrow$

(5c) Price-ratio = {Quality Corrections } + {Quality Adjusted Price Change conditional on $\overline{\boldsymbol{x}}'_{kt}$}.

Table 4.1 and equations (5a) to (5c) reveals what we have spoken about - our transparent simple algebra using optimal unbiased statistics. First, both averages satisfy three basic algebraic properties of the OLS method without systematic errors. Second, the slope estimates are BLUE under homoscedastic errors. Third, both averages are unbiased and consistent in aggregation. Fourth, the Oaxaca decomposition in (5b) is consistent and surprisingly the most optimal for our empirical application. Fifth, true price-ratio of averaged prices is decomposed into two parts: quality corrections and quality adjusted price change with comparable in quality, that is $\overline{\boldsymbol{x}}'_{kt}$. Sixth, the Oaxaca decomposition in (5b) tell that the OLS estimation is necessary to apply only for time period 0 because unconditional and conditional averages equal for any category/stratum *k* because of algebraic property of OLS.

Using unconditional and conditional averages in suitable manner, the equations (5) may represent by simple logarithmic price ratios as

(6) $log(\bar{p}_{kt}/\bar{p}_{k0}) = log(\tilde{p}_{kt}/\bar{p}_{k0}) + log(\bar{p}_{kt}/\tilde{p}_{kt}), \forall\, k, 0, t$

It is very simple and holds as an identity. On the left, we have the price-ratio of actual average prices. On the right, the first term is quality correction (QC) estimated using the base period valuation of characteristics (i.e., $log(\tilde{p}_{kt}) = \hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{j0}$ and $log(\bar{p}_{k0}) = \hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{k0}\widehat{\boldsymbol{\beta}}_{j0}$) and the second term is quality adjusted (QA) price change (i.e., $log(\bar{p}_{kt}) = \hat{\alpha}_{kt} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{jt}$ and ($log(\tilde{p}_{kt}) = \hat{\alpha}_{k0} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{j0}$) estimated using the base period valuation of characteristics ($\widehat{\boldsymbol{\beta}}_{j0}$) with characteristics being comparable in quality (i.e., $\overline{\boldsymbol{x}}'_{kt}$, for all *k* and *t*). We construct the equation (6) for unweighted arithmetic and geometric averages.

## 4.2 Index Number Formulas

In price modelling all used cars are grouped together to form *K* categories, $A_k, k = 1, \dots, K$, which define our partition of observations, that is $A = A_1 \cup A_2 \cup \dots A_K$, where different $A_k$ categories are disjoint. Previous chapter ends our analysis into equation (6), where logarithmic price ratio of true actual averages (*A*) is decomposed into log-price ratios for quality corrections (QC) and quality adjusted (QA) price change. This is done for all categories, for which we define an index number formulas. We use a simple notation here for an index number

$$P_f^{t/0} = P_f(\overline{\boldsymbol{p}}_0, \boldsymbol{q}_0, \overline{\boldsymbol{p}}_t, \boldsymbol{q}_t),$$

where $\overline{\boldsymbol{p}}_0$ and $\overline{\boldsymbol{p}}_t$ are *K*-vector of average prices (geometric or arithmetic) and $\boldsymbol{q}_0$ and $\boldsymbol{q}_t$ *K*-vector of corresponding quantities of sold cars. We define above price index for equation (6), that is

$$\text{(7a)} \quad exp\{\textstyle\sum_k w_{k,f}\, log(\bar{p}_{kt}/\bar{p}_{k0})\} = exp\{\sum_k w_{k,f}\, log(\tilde{p}_{kt}/\bar{p}_{k0}) + \sum_k w_{k,f}\, log(\bar{p}_{kt}/\tilde{p}_{kt})\} \leftrightarrow$$

$$\text{(7b)} \quad P_{f,A}^{t/0} = P_{f,QC}^{t/0} \cdot P_{f,QA}^{t/0}$$

The left side is the price index for average prices (*A*) for formula *f* for price-link from base period 0 to the period *t*. The first term in the right side is the price index for quality corrections (*QC*) and the last term price index for quality adjusted price changes (*QA*). Weights in equation (7a) for formulas are presented in Table 4.2.

Table 4.2: Weights for index number formulae (logarithmic forms).

| Basic formulae, see Vartia & Suoperä, 2017, 2018, *L* means logarithmic mean, see Vartia, 1976a, p. 128 | |
|---|---|
| Symbol and name of formula | Weights of the formula |
| *Laspeyres, f = L* | $w_{k,f} = w_{k,L}^0 = \frac{L(\bar{p}_{kt}q_{k0}, \bar{p}_{k0}q_{k0})}{L(\sum_k \bar{p}_{kt}q_{k0}, \sum_k \bar{p}_{k0}q_{k0})}$ |
| *log-Laspeyres, f = LL* | $w_{k,f} = w_{k,l}^0 = v_k^0/V^0$ |
| *log-Paasche, f = LP* | $w_{k,f} = w_{k,p}^t = v_k^t/V^t$ |
| *Paasche, f = P* | $w_{k,f} = w_{k,P}^t = \frac{L(\bar{p}_{kt}q_{kt}, \bar{p}_{k0}q_{kt})}{L(\sum_k \bar{p}_{kt}q_{kt}, \sum_k \bar{p}_{k0}q_{kt})}$ |
| Excellent formula, see Vartia & Suoperä, 2017, 2018), *L* means logarithmic mean, see Vartia, 1976 | |
| *Törnqvist, f = T* | $w_{k,f} = \bar{w}_{k,T} = 0.5 \cdot (w_{k,l}^0 + w_{k,p}^t)$ |
| *Sato-Vartia, f = SV* | $w_{k,f} = \bar{w}_{k,SV} = \frac{L(w_k^t, w_k^0)}{\sum L(w_k^t, w_k^0)}$ |
| *Montgomery-Vartia, f = MV* | $w_{k,f} = \bar{w}_{k,MV} = \frac{L(\bar{p}_{kt}q_{kt}, \bar{p}_{k0}q_{k0})}{L(\sum_k \bar{p}_{kt}q_{kt}, \sum_k \bar{p}_{k0}q_{k0})}$ |
| *Fisher, f = F* | $w_{k,f} = \bar{w}_{k,F} = 0.5 \cdot (w_{k,L}^0 + w_{k,P}^t)$ |

Some notes are necessary:

1. We define price-link form $0 \rightarrow t$ meaning that we use the base strategy that is free of the chain drift. The base period is a previous year normalized as an average month and *t* a month of a current year.

2. Our aggregation means here always '*a weighted-by-economic-importance*'-variable familiar to index numbers, i.e., weighting by $w_{k,f}$.
3. Price index is based on transparent and familiar traditional theory of index numbers.
4. Quality corrections can be decomposed for *E* dimensional *x variable-by-variable* such that $P_{f,QC}^{t/0} = P_{f,QC,x_1}^{t/0} \cdot P_{f,QC,x_2}^{t/0} \cdot \ldots \cdot P_{f,QC,x_E}^{t/0}$ holds as an identity.
5. We may construct index series not only for average prices (true averages and quality adjusted) but also for any single quality corrections or any combinations of them consistently.
6. We use '*a flexible basket*'-approach that states '*when the expenditure on a category tends to zero, then its effect on the index should vanish*' (Pursiainen, 2006, pp32). *We make comparison's only for categories having expenditures for both 0 and t*.

In Table 4.2 we gather all information that is necessary for calculation of hedonic price indices for equations (7). We analyze all index number formulae in logarithmic form, including Laspeyres, Paasche and Fisher (see Vartia, 1976, p.128). The aggregation of price changes or their decompositions in (6) and (7) are much simpler in additive form using 'log's' – as in (7), they may simply transform back to indices. In empirical part we use two set of formulas – basic and excellent.

# 5 Empirical Results for Category Averages and Hedonic Index Numbers

The empirical results for price models are presented in chapter three. Now we proceed into empirical analyze of elementary aggregates, unweighted geometric and arithmetic averages, and their index number solutions based on Oaxaca decompositions. *First*, we show which average (arithmetic or geometric) should be selected as average statistics of relative change and *second*, does the formula matter.

## 5.1 Arithmetic or Geometric Average as Mean Statistic

Table 5.1 shows how much arithmetic and geometric averages deviate in aggregate level.

Table 5.1: Arithmetic and geometric average prices (Euro) in year 2020, 2021 and 2022.

| Year | Arithmetic average | Geometric average |
|---|---|---|
| 2020 | 15416 | 11990 |
| 2021 | 17214 | 13622 |
| 2022 | 18742 | 14280 |

Average prices are estimated from category averages using their frequencies as weights (i.e., relative shares). Averages deviate substantially being about 30 log-%. For more expensive makes and models the difference become even bigger indicating that geometric average is poor as official statistic as averages.

## 5.2 Arithmetic or Geometric Average as Statistic of Relative Change

We get back to equation (6) and show how closely relative changes of arithmetic and geometric averages are related. First, we regress relative change of arithmetic averages on relative changes of geometric averages (left side of eq. (6)). Second, we do the same for relative changes of quality adjusted average prices (second term right hand in eq. (6)). The model is the simplest regression

$$y_{kt} = \rho \cdot x_{kt} + \varepsilon_{kt},$$

where $y_{kt}$stands for relative changes of arithmetic averages and $x$ for relative changes of geometric averages for price-links $0 \rightarrow t$ and categories $k = 1,\dots, K$. Similar equation are applied also for corresponding relative changes of quality adjusted price changes. The estimator for $\rho$ is also nicely interpreted as

$$\hat{\rho} = r(y,x) \cdot \frac{s_y}{s_x}$$

When the standard deviations of *x* and *y* are closely related, the estimator $\hat{\rho}$ practically equals to correlation coefficient between *x* and *y*. In both OLS estimation we have 17935 observations (total number of categories in years 2020, 2021 and 2022) from price ratios and Table 5.2 presents the results.

Table 5.2: Linear relation between price ratios of arithmetic and geometric averages.

| | $s_y$ | $s_x$ | $\hat{\rho}$ | $r_{xy}$ | $R^2$ |
|---|---|---|---|---|---|
| Actual price ratio left side of (6) | 0,103 | 0,106 | 0,966 | 0,995 | 0,991 |
| Quality adjusted price ratio, second right term of (6) | 0,182 | 0,179 | 0,9998 | 0,986 | 0,973 |

Empirical results show that price ratios using unweighted arithmetic or geometric average prices are very closely related. Both 95 % fit plots for y include complete linear dependence meaning that statistically the choice between arithmetic or geometric average have no matter. The correlation coefficient tells the same story – they are close to one. Quite amazingly, although the arithmetic and geometric average prices deviate largely (see Table 5.1), their price ratios go 'hand-to-hand' – at least *statistically*. Next, we analyze differences between these averages using index numbers.

## 5.3 Does Formula and Average matter in Index Compilation?

All index numbers and index series are based on base strategy, where the base period is a previous year normalized as an average month and the observation period is a month of a current year. The strategy is free of chain drift. Our empirical analyze turns into two questions - 'Does the formula matter in index compilation?' and 'Does the average matter in index compilation?'. We compare two sets of formulas, the basic and excellent (Vartia and Suoperä, 2017, see Table 4.2 and eq. (7)). All formulas are examined in log-form. In this study our basic formulas are Laspeyres (*L*), log-Laspeyres (*LL*), Paasche (*P*) and log-Paasche (*LP*). *L* and *LL* formulas use asymmetric old weights and formulas *P* and *LP* new ones. The second set of formulas – excellent ones – uses symmetrical weights and are Fisher (*F*), Törnqvist (*T*), Montgomery-Vartia (*MV*) and Sato-Vartia (*SV*) (see Vartia & Suoperä, 2017, 2018). The following graphs show why they are excellent.

The Figures 5.1-5.4 present all that is needed to make decisions about the formula and the average used in index compilation. Index series in Figures 5.2 and 5.4 are made using arithmetic and geometric average price ratios. Index series based on arithmetic and geometric averages deviate seriously but excellent formulas go 'hand-in-hand' for both index series (both index series includes four excellent formulas). Our empirical results in previous chapter show that price changes based on arithmetic and geometric average prices are statistically almost 'equal' (95 % fit plots for y includes complete linear dependence) and correlation between them was $r_{xy}$ = 0.995. Simple econometric modelling concludes: 'statistically the choice between arithmetic or geometric average have no matter'.

Figure 5.1: Index series for actual average prices for ‘Small Cars’ make ‘Honda’. Basic formulas: indices based on geometric are dotted and arithmetic solid lines.

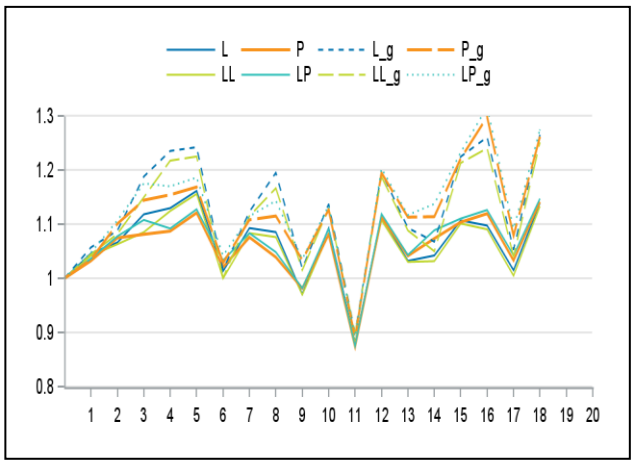



Figure 5.2: Index series for actual average prices for ‘Small Cars’ make ‘Honda’. Excellent formulas: indices based on geometric are dotted and arithmetic solid lines.

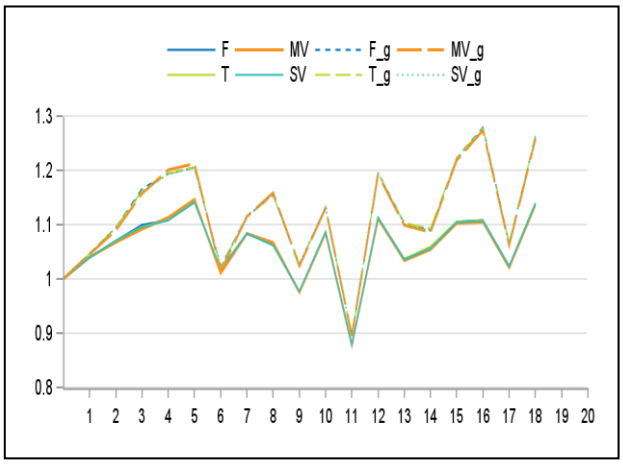



Figure 5.3: Index series for actual average prices for ‘Small Cars’ make ‘MB’. Basic formulas: indices based on geometric are dotted and arithmetic solid lines.

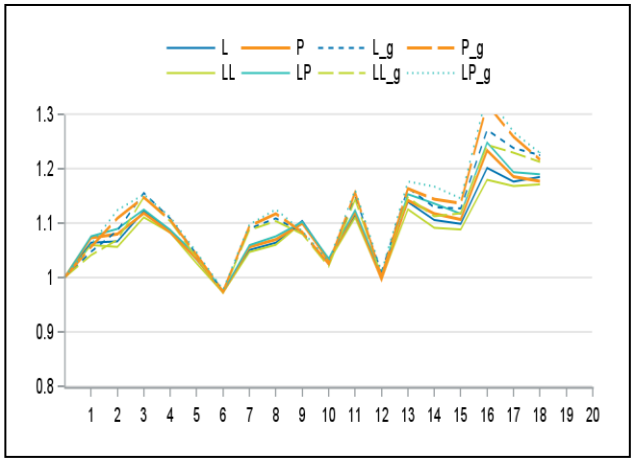
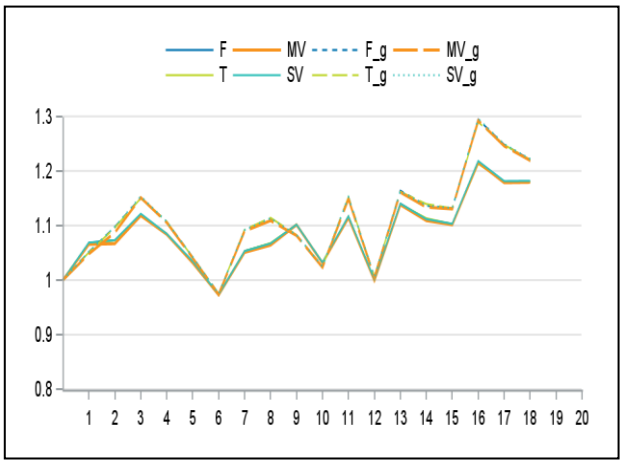



Figure 5.4: Index series for actual average prices for ‘Small Cars’ make ‘MB’. Excellent formulas: indices based on geometric are dotted and arithmetic solid lines.




Figures 5.2 and 5.4 tell that because of *contingent nature of data*, the index series based on arithmetic and geometric averages may occasionally seriously deviate. Statistically they are almost equal but not mathematically. Our selection for price concept of average statistic and index compilation is more interpretable using arithmetic average (see also Table 5.1). In Figure 5.1 and 5.3 we see that basic formulas are contingently biased (see Vartia and Suoperä, 2017, 2018) deviating seriously from each other. Basic formulas for complete data should never be used.

## 5.4 Hedonic Index Numbers for Used Cars in Finland

Next, we aggregate decomposition in equation (7) from *K*-category into total using only excellent formulas. In our empirical analysis excellent formulas are very closely related. This happens because all excellent formulas are quadratic approximations of *Fisher* for small changes in log-prices and log-quantities (Vartia and Suoperä, 2017, 2018, pp. 17-21). This seems to happen here also quite closely for moderate changes of

log-prices and log-quantities. The same happens extremely closely for quality adjusted indices (solid lines in Figure 5.5).

Figure 5.5: Hedonic index series for actual average prices (arithmetic) and quality adjusted prices for excellent formulas *F*, *T*, *MV* and *SV* (solid lines).

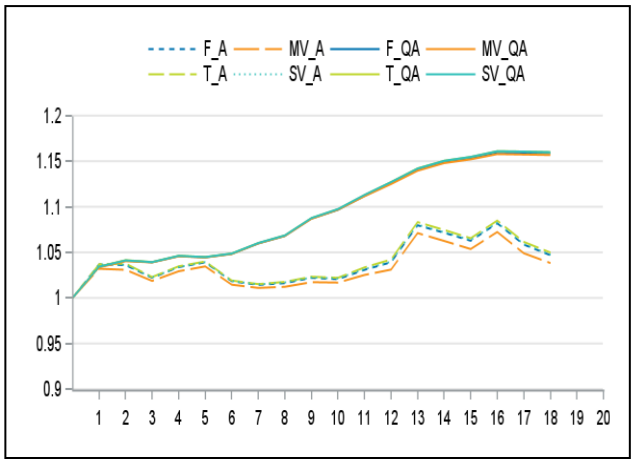
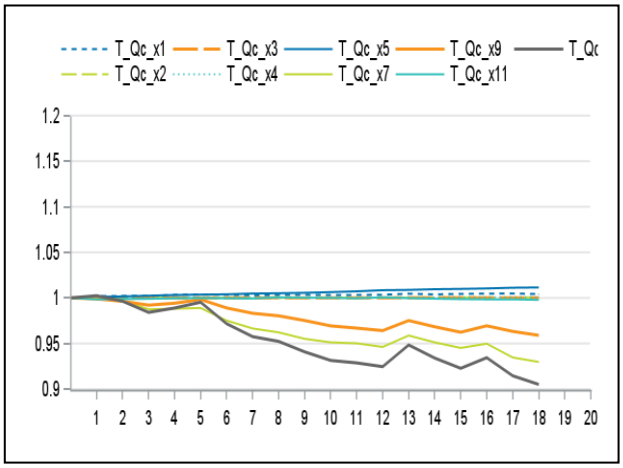



Figure 5.6: Hedonic index series for quality corrections for quality characteristics *x* (T_Qc = all) by Törnqvist formula.




Figures 5.5 and 5.6 must be looked at together: For any excellent formula (*F*, *T*, *MV* and *SV*) difference between index series for actual average prices and quality adjusted prices equals with total quality correction. The most part the difference is explained by quality corrections of age of a car ($x_7$) and mileage ($x_9$) – sold cars are simply older and more driven at observation period. Other quality corrections ($x_1, x_2, x_3, x_4, x_5$ and $x_{11}$) have minor role (index series close to one in Figure 5.6). The Figures 5.5 and 5.6 together are graphical presentation of equation (7b) for Törnqvist ideal formula, that is $P_{T,A}^{t/0} = P_{T,QC}^{t/0} \cdot P_{T,QA}^{t/0}$.

## 6 Conclusion

We show, using statistical inference, how two sources of heterogeneity – categorial and behavioral – may be chosen hierarchically for the best price models for hedonic quality adjusting. By this statistical inference we empirically decide first the 'best' partition of observations and second the 'best' categorization of behavioral 'beta' heterogeneity. The decision-making leads us into the optimal best linear unbiased estimates, BLUE, for fixed categorical and beta effects.

We combine the BLU estimates with consistent aggregation rules and get unbiased parametric presentations for categorical averages. These *K*-categorical averages - arithmetic and geometric – both satisfy the well-known algebraic properties the OLS method being also unbiased and optimal for making of hedonic index numbers. The price modelling ends to aggregation of relations from observations into *K*-category level with these averages.

Oaxaca decomposition divides changes of actual average (arithmetic or geometric) price ratios into two parts: first, quality correction of quality characteristics and second, quality adjusted price changes. In the Oaxaca decomposition the base period is the previous year normalized as an average month. This enables us to use base strategy which is free of chain drift.

For the base strategy we select 'flexible basket approach' to verify the principle of Pursiainen that states '*when the expenditure on a category tends to zero, then its effect on the index should vanish* (Pursiainen, 2006, pp32). In we combine heterogeneously behaving cross-sections with classical index number theory.

This representation of ‘index numbers’ makes it possible to control quality changes of quality characteristics and remove quality differences from unbiased actual average price ratios.

The making of hedonic index numbers, we use two set of formulas, the basic and excellent ones. We show that basic formulas using asymmetric weighting, are contingently biased and should not be used. Excellent formulas in the study uses symmetrical weighting giving excellent results. Using symmetric weights of these excellent formulas satisfies the principle of ‘*a weighted-by-economic-importance*’-variable optimally being mathematically transparent. According to the study, any excellent formula with arithmetic average can be selected for official statistics.

## References:


**Bailey M. J., Muth, R. F. and Nourse, H. O. '**A Regression Model for Real Estate Price Index Construction'., JASA, vol. 58, 933-942, 1963.

**Case, K. E. and Shiller, R. J. '**Efficiency of the Market for Single Family Homes', American Economic Review, vol. 79, 125-137, 1989.

**Davidson & MacKinnon '**Estimation and Inference in Econometrics', New York, Oxford University Press, 1993.

**Diewert E. and Fox K.** 'Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data' 2018

**Greene, W.** "Econometric Analysis", Prentice-Hall International, Inc. (third ed.), 1997.

**Griliches Z. '**Hedonic Price Indices for Automobiles: An Econometric Analysis of Quality Change', *Zvi Griliches (ed.) Price Indexes ad Quality Changes, 55-97,* 1971.

**Hsiao, C. '**Analysis of Panel Data'., Cambridge University Press, 1986.

**Kaila, J., Luomaranta, H. and Suoperä, A.** Hedonic Price Index Number for Blocks of Flats and Terraced Houses in Finland', 2023 (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

**Koev, E. '**Combining Classification and Hedonic Quality Adjustment in Constructing a House Price Index', Licentiate thesis, Helsinki, 2003.

**Koev, E. & Suoperä A. '**Pientalokiinteistöjen (omakotitalojen ja rakentamattomien pientalotonttien) hintaindeksit 1985=100', Helsinki, 2002. (in Finnish, Statistics Finland).

**Matyás, L. and Sevestre, P.**, Eds. "The Econometrics of Panel Data: Handbook of Theory and Applications, 2nd ed. Dordrecht: Kluwer-Nijoff, 1996.

**Oaxaca, R. '**Male-Female Wage Differentials in Urban Labour Markets', International Economic Review, 14, pp. 693-709, 1973.

**Practical Guide on Multilateral Methods in the HICP (2020, WTPD-model), EuroStat.**

**Pursiainen, H.** 'Consistent Aggregation Methods and Index Number Theory', 2005.

**Quigley, R. '**A Simple Hybrid Model for Estimating Real Estate Price Indexes', Journal of Housing Economics vol. 4, p. 1-12, 1995.

**Rao, D.S. P.** 'On the Equivalence of the Weighted Country Product Dummy (CPD) Method and the Rao System for Multilateral Price Comparisons', Review of Income and Wealth 51:4, 2005, 571-580.

**Summers R.** 'International Comparisons with Incomplete Data", Review of Income and Wealth 29:1, 1973, pp. 1-16.

**Suoperä, A. '**Some new perspectives on price aggregation and hedonic index methods: Empirical application to rents of office and shop premises', 2004, 2006 (unpublished, Statistics Finland).

**Suoperä A. & Auno V.** 'Hedonic Index Numbers for Rents of Office and Shop Premises in Finland', 2021 (https://www.researchgate.net/publication/350460207_Hedonic_Index_Numbers_for_Rents_of_Office_and_Shop_Premises_in_Finland).

**Suoperä, A., Nieminen, K., Montonen, S. and Markkanen H.** "Comparing Basic Averages, Index Numbers and Hedonic Methods as Price Change Statistic", 2021 (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

**Suoperä, A. & Vartia, Y.** 'Analysis and Synthesis of Wage Determination in Heterogeneous Cross-sections', Discussion Paper No. 331, 2011.

**Vartia, Y., Suoperä, A. and Vuorio, J.** 'Hedonic Price Index Number for New Blocks of Flats and Terraced Houses in Finland', 2021 (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

**Vartia, Y.** 'Relative Changes and Index Numbers', Ser. A4, Helsinki, Research Institute of Finnish Economy, 1976.

**Vartia, Y. '**Ideal Log-Change Index Numbers', Scandinavian Journal of Statistics., 3, pp. 121-126,1976.

**Vartia, Y. '**Kvadraattisten mikroyhtälöiden aggregoinnista', ETLA, Discussion Papers no. 25,1979.

**Vartia, Y. & Suoperä, A.** "Index number theory and construction of CPI for complete micro data", 2017 (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

**Vartia, Y. & Suoperä, A.** "Contingently biased, permanently biased and excellent index numbers for complete micro data", 2018 (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

**Vartia, Y. and Vartia, P. '**Descriptive Index Number Theory and the Bank of Finland Currency Index', Scandinavian Journal of Economics, vol. 3, pp. 352 . 364, 1985.

**Törnqvist, L. '**A Memorandum Concerning the Calculation of Bank of Finland Consumption Price Index', unpublished memo, Bank of Finland, 1935.

**Törnqvist, L.** 'Levnadskostnadsindexerna i Finland och Sverige, Deras Tillförlitlighet och Jämförbarhet', Ekonomiska Samfundets Tidskrift, vol. 37, 1–35, 1936.

**Törnqvist, L. & Vartia, P. & Vartia, Y.** 'How Should Relative Changes be Measured'? The American Statistician, Vol. 39, No. 1. pp. 43 - 46, 1985.