Beyond matched pairs and Griliches-type hedonic methods for controlling quality changes in CPI sub-indices

Mathematical considerations and empirical examples on the use of linear and non-linear hedonic models with time-dependent quality parameters

Paper presented at the Sixth Meeting of the International Working Group on Price Indices
Canberra, Australia
2–6 April 2001

Timo Koskimäki[1],
Statistics Finland

Yrjö Vartia[2],
University of Helsinki

[1] timo.koskimaki@stat.fi
[2] yrjo.vartia@helsinki.fi

1. Introduction[3]

The majority of recent research literature dealing with quality adjustment of price indices is largely concerned of rather technical issues. In the realm of using regression models for quality adjustment the frequently occurring topics are the correct choice of price determining factors, functional forms of the regression equation and other aspects of the estimation of a correct regression model.

As it comes to statistical agencies compiling price indices, the dominant form of thinking is the traditional matched model - approach . The academic research on hedonic regression models has so far been used in the official index compilation only in a very conservative manner. As a variety originally chosen in the CPI sample disappears from the market a new variety is selected instead. Hedonic regression modelling then provides a method to estimate a previous period's price for the qualitatively equivalent product. This is done by estimating the previous period forecasted price for the newly selected product. The calculation is performed using pre-established regression coefficients for certain price-determining factors. The basic mechanism, however, is still matching the pairs.

In our opinion, there are more natural ways to use hedonic methods in index compilation, where the quality control is not based on pre-specified matched pairs of observations, see also Hyrkkö – Kinnunen – Vartia (1998). We will elaborate a rather general mathematical framework on hedonic regression models which, we hope, will make the future discussion on the topic more structured. The framework explicates the mathematical relationships between various types of hedonic quality adjustment strategies. The formal considerations will be illustrated with empirical examples. The data we use has been gathered as a part of standard CPI data-collection. The empirical results presented here are thus not only examples but also give indication on the magnitude of bias caused by the use of inappropriate methods when analysing rapidly changing markets.

The structure of the paper is as follows:

In Chapter 2 we will make some remarks on currently used practices of quality adjustment:
(i) quality adjustment using only matched pairs approach,
(ii) overlap prices and
(iii) the attempt to cure the shortcomings of traditional approaches by using regression estimates to patch the data.

In (iii) a disappeared **matched pair** is substituted by a constructed **patched pair**. The remarks we make are based on the results obtained from our test data of computers when attempting to follow the above mentioned practices. Timo Koskimäki wrote the first draft for this part.

In Chapter 3 we expose our formal framework starting with the most general variant of hedonic models, the case where the coefficients of the model are allowed to change in time and the model includes second order terms, or any other non-linear terms. Linear and time-invariant models, including the frequently used Griliches-type models, are considered as special cases of the general model. The relationship between the various types of hedonic approaches will be analysed and proven mathematically. Empirical examples based on our test data will be provided for most the model types we consider. Yrjö Vartia wrote the first draft for this part.

Chapter 4 concludes and suggests topics for further research.

2. The Traditional Approach

2.1 Characteristics of the PC markets - the test data

The data used in this experiment consists of 245 price observations. In addition to price, each observation also contains 5 simple quality characteristics of the computer model: processor type, processor speed, size of the hard disk, size of the memory and size of the display. Material consists only of desktop PCs', laptops are not included in the study material. The data used was collected between May and October 2000. For this study, the originally monthly data sets have been merged to three two-month sets, which we in the following refer to as "spring", "summer" and "fall".

The data was collected as a part of the standard price collection of the Finnish CPI. The only change as compared to standard practice was that the five quality variables used in this analysis were coded into the database. The information itself was already available in the product characteristics that the price collectors are obliged to collect as a part of the standard price collection procedure. The additional cost caused by coding of quality characteristics was negligible.

[3] Ms Mari Suviranta and Mr Kari Manninen from Statistics Finland have participated the work related to our project in many ways. We express to them our most sincere thanks.

Basic characteristics of the data are given in table 1 below:

Table 1: Characteristics of the test data

	Characteristics					
	Number of observations	*Mean Price*	*Mean Processor speed*	*Mean Memory size*	*Mean Hard disk size*	*Share of high-end processors (Pentium III +, AMD Athlo*
Period	*N*		*MHz*	*Mb*	*Gb*	*per cent*
Spring	83	1312	532	73,3	10,6	37,3
Summer	83	1282	549	74,8	10,7	43,4
Fall	79	1326	606	72,8	14	44,3
Total	*245*	*1306*	*561*	*73,6*	*11,7*	*41,6*

Assuming that our sample gives a representative picture of the Finnish PC-markets, the following seems to have happened between spring and fall of year 2000:

- High-end processor types (pure Pentium and AMD Athlon -processors) have increased their market share
- Average processor speed has clearly increased, as well as the size of hard disks
- New computers are equipped with approximately same amount of memory throughout the period
- Unadjusted mean price decreases towards the summer and then rises again in the fall

2.2 Matching the pairs

As stated above, the prevalent strategy in official price index construction is keeping the quality constant by following the **matched pairs strategy**. The following account taken from the draft OECD handbook that has been prepared by Jack Triplet, summarises the pros of the approach in an excellent way:

> "Price indexes, nearly universally, employ one fundamental methodological principle: The price index compiling agency chooses a sample of retail outlets or sellers and of product. It collects an initial period, or base period, price for each of the products selected. It then collects at some later date the price for exactly the same product, from the same seller, that was selected in the initial period. The price index is computed by matching, observation by observation, the price at the later period with the initial price.
>
> The great advantages of this matching methodology are sometimes not explicitly stated, and other times not fully appreciated. The "matched model" methodology holds constant many price-determining factors that are usually not directly observable. Examples are characteristics of the retailer, such as customer service, reputation of the manufacturer etc. Matching the price quotes model by model (and outlet by outlet) is not just a methodology for holding quality change constant in the items selected for pricing. It is also a methodology for holding constant non-observable aspects of the transaction that might bias the measure of price change. "Triplett (2000), pages 3-4).

The problem is, of course, that matching the pairs often fails.

Let us now try to apply the matching of the pairs methodology to our test data. When matching the spring data with the summer data we obtain 55 matching pairs. In the fall data, only 16 computers out of the originally chosen ones are left in the sample. Matching of the summer data with the fall data yields 27 matching pairs. The sample deterioration rates are presented in table 2. The high share of non-matches in summer- fall comparison is somewhat unexpected. From the sample design point of view it should be approximately the same as for spring-summer comparison. Apparently, quite a number of new computer models were introduced into the markets in the fall and reduced the effective sample size more than normally would be expected.

Table 2: Share of successful matches

	Target	*Matches*	*Share of successful matches*
Period	*N*	*N*	*per cent*
Spring - summer matching	83	55	**66,3**
Spring - fall matching	79	16	**20,3**
Summer - fall matching	79	27	**34,2**

It is evident, that reliance on the pure matched pairs approach does not make very effective use of the data collected. In our case, depending on the study period, thirty to eighty percent of the collected price observations can not be used for index

construction due to the failure of finding a matching variety. Although the two-month design of the study material somewhat exaggerates the sample deterioration, the phenomenon is not negligible in a monthly material either.

The wasting of collected data is not the only shortcoming of the matched pairs approach. In table 3 below we compare the characteristics of matching models with the non-matching ones.

Table 3: Characteristics of matching and non-matching computer

	Number of observations	*Mean Price*	*Mean Processor speed*	*Mean Memory size*	*Mean Hard disk size*	*Share of high-end processors (Pentium III +, AMD Ath*
Period	*N*	*•*	*MHz*	*Mb*	*Gb*	*per cent*
Summer, spring matches	55	1282	535	72,1	10,2	43,6
Summer, non-matches	28	1280	577	80	11,9	42,8
Difference, per cent		***0,2***	***-7,3***	***-9,9***	***-14,3***	***1,9***
Fall, spring matches	16	1307	518	68	10,2	25
Fall, non matches (1)	63	1331	628	74	15	49,2
Difference, per cent		***-1,8***	***-17,5***	***-8,1***	***-32,0***	***-49,2***
Fall, Summer matches	27	1296	548	71,1	10,7	37,3
Fall, non matches (2)	52	1342	636	73,7	15,7	48,1
Difference, per cent		***-3,4***	***-13,8***	***-3,5***	***-31,8***	***-22,5***

The samples of matching pairs get - in our test material - rapidly biased. After two months, the computers that were on the market in the spring and have been kept in the sample are clearly of lower quality than the replacements selected into the summer sample. The same holds also when we compare summer data with fall data and, of course, in comparing spring with fall. This is quite understandable on the basis of traditional price-collection practices. Price collectors are advised to follow the price of the one and same model as long as it is possible. In rapidly changing markets this practice apparently leads to non-representative samples in the course of a couple of months.

Table 4 shows the period-on-period changes and price indices based spring 2000 = 100. No explicit quality adjustments have been made. The matched indices show only very moderate price changes. The overlap-index, where the summer-fall matching pairs are all utilised, ends up with 2,4 percent price decrease. The "fixed-base" index that only takes into account the 16 matched pairs for the summer - fall comparison shows, at the end of the period, practically no price change at all.

It is naturally impossible to continue the pure fixed base index for any longer period in time as all the models in the sample will vanish. In practice it is possible, in addition to the overlapping prices, continue the price collection as if nothing had happened, i.e. consider the changes in the models as irrelevant. This practice would in the case of computers lead to rapidly increasing prices. The third possibility, if we exclude the use of hedonic methods, is to apply some form of judgmental quality adjustment. The impact of this procedure to observed price developments is difficult to foresee.

The "patched" index in table 4 has been constructed by replacing the base period missing prices with estimated values using the same type of linear models which will be presented in chapter 2.3. below. The patched index also shows a very moderate price decline for the period. The main difference compared to fixed and overlap indices is that the patched index does not indicate price increase from summer to fall as the other two indices do.

Table 4: Traditional price indices calculated from matched (and patched) samples

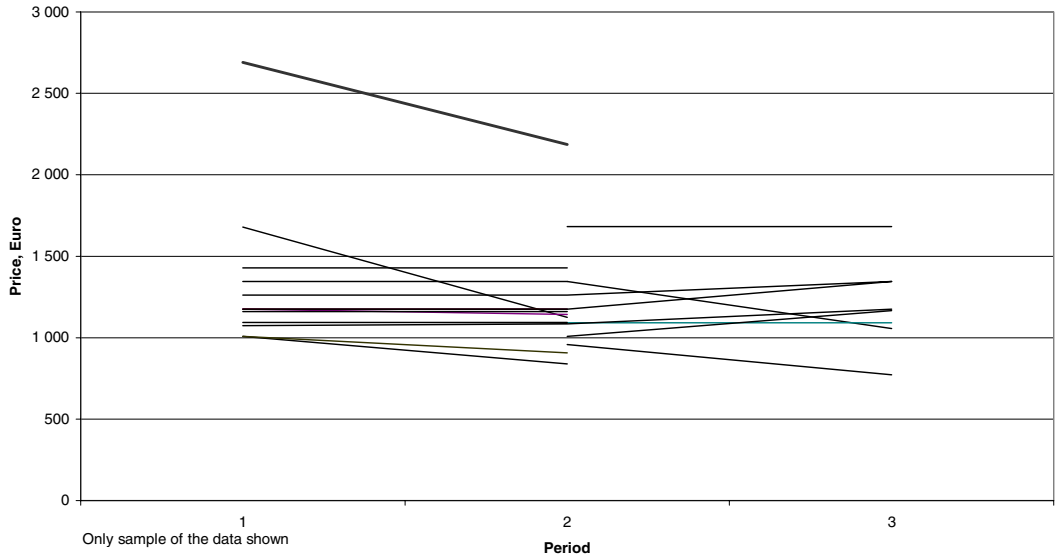
	Fixed-base	Overlap	Patched
Spring	100	100	100
Summer	96,5	96,5	99,8
Fall	99,8	97,6	99,4

One could argue that computer models should show some form of "market life-cycle", i.e. that a new model is introduced into the market at a higher price and that the price of the model then decreases gradually until it disappears from the market see Turvey (1999). . The phenomenon does not seem to be very prevalent in our data. Figure 2.1 shows the price developments in our test material. Instead of "market life cycle " with continuously decreasing price the more common case seem to be a "straight line" indicating that during its market life, the price of a computer model does not change. Almost 70 per-

cent of the computers disappearing from our sample in summer showed no price change at all during the two-month period. Although here again our test-sample design exaggerates the phenomenon, a considerable share of the price changes occur in combination of the introduction of a new model.

This lack of clear market life cycles for PC models poses real problems to the traditional index construction methods, see Vartia 1976. If the price changes in most cases occur in combination with the model changes, the only reasonable method for describing the price changes is some form of explicit quality adjustment.

Figure 2.1: Computer prices
Test data set, spring - fall 2000, Finland



2.3 Hedonic models of price-determining factors

A necessary step in using hedonic methods for quality adjustment is to construct a valid hedonic model. In case of computers, there exists quite a body of research- and working papers on the choice of relevant variables and other aspects of model construction. As the details of the modelling approach are not our major concern here, we just show here summary estimation results on the two models and the sets of data, which will be used as examples in the remaining part of this paper.

In both of the models the log of price is explained by other log-form variables. In our linear model the explanatory variables are the log of processor speed and the log of the size of memory. In the non-linear model a second-order term, squared log of processor speed (actually the square of its deviation from its mean, which allows easy interpretation of the regression coefficients), is incorporated in the model as well.

Linear model estimated from the whole data set:

Root MSE	0.13326	R-Square	0.5930
Dependent Mean	7.15222	Adj R-Sq	0.5896
Coeff Var	1.86315		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.42550	0.32665	4.36	<.0001
lnmem	log of memory	1	0.29869	0.03715	8.04	<.0001
lnspeed	log of speed	1	0.70499	0.05569	12.66	<.0001

Non - linear model estimated from the whole data set:

Root MSE	0.13008	R-Square	0.6138
Dependent Mean	7.15222	Adj R-Sq	0.6089
Coeff Var	1.81873		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.79074	0.33461	5.35	<.0001
lnspeed2a	2nd order processor speed a	1	0.75094	0.20855	3.60	0.0004
lnmem	log of memory	1	0.27245	0.03699	7.37	<.0001
lnspeed	log of speed	1	0.66164	0.05568	11.88	<.0001

Linear model estimated from the spring data:

Root MSE	0.12390	R-Square	0.6458
Dependent Mean	7.15659	Adj R-Sq	0.6370
Coeff Var	1.73125		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.42706	0.60982	0.70	0.4858
lnmem	log of memory	1	0.31492	0.05977	5.27	<.0001
lnspeed	log of speed	1	0.85979	0.10416	8.25	<.0001

Non- linear model fitted from the spring data:

Root MSE	0.11855	R-Square	0.6798
Dependent Mean	7.15659	Adj R-Sq	0.6677
Coeff Var	1.65645		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.91994	0.60778	1.51	0.1341
lnspeed2b		1	1.20525	0.41614	2.90	0.0049
lnmem	log of memory	1	0.25287	0.06107	4.14	<.0001
lnspeed	log of speed	1	0.81955	0.10062	8.14	<.0001

Linear model estimated from the fall data:

Root MSE	0.13253	R-Square	0.6284
Dependent Mean	7.16682	Adj R-Sq	0.6186
Coeff Var	1.84928		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.39368	0.52989	2.63	0.0103
lnmem	log of memory	1	0.28087	0.06923	4.06	0.0001
lnspeed	log of speed	1	0.71644	0.09126	7.85	<.0001

Non-linear model estimated from the fall data:

Root MSE	0.12647	R-Square	0.6661
Dependent Mean	7.16682	Adj R-Sq	0.6527
Coeff Var	1.76466		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.56040	0.50888	3.07	0.0030
lnspeed2b		1	0.92807	0.31901	2.91	0.0048
lnmem	log of memory	1	0.23795	0.06769	3.52	0.0007
lnspeed	log of speed	1	0.71419	0.08709	8.20	<.0001

3. Mathematical theory of hedonic price functions and of quality corrections in index numbers

3.1 Conceptual background and notation for different hedonic models

Let's start from the general notation for all different variations of **hedonic models (HM)** we are going to consider. We choose to consider instead of the actual price P its logarithm $y = \log p$ as the variable to be explained or forecasted using the vector x of relevant explanatory (quality) variables and the time t. The conceptual background of all regression models is the conditional expectation

(1) $$E(\log p \mid x, t) = g^t(x).$$

The function $g^t(x)$ of time t and x-vector defined by (1), defines the **regression surface** of $\log p$. It is a mapping from $R^K \times T$ to R, where K is the number of explanatory x-variables and T is the set of time periods concerned. The functional form of $g^t(x)$ is determined by the theoretical (usually hypothetical) joint distribution of the random vector $(\log p, x_1, ..., x_K) = (\log p, x)$ indexed by time t. Different models for this joint distribution (which should be carefully adjusted to the actual problem considered, to the data available and to a priori knowledge concerning the dependencies[4].) lead to different functional forms[5]. of $g^t(x)$ We will shortly review some general possibilities, which lead to different variations of HM. We ignore in this treatment the problems of model building and of estimation as these are widely discussed in standard econometric and statistical texts, such as Goldberger (1964) or Spanos (1986).

We denote the **estimated regression function** simply by

(2) $$\begin{aligned} f^t(x) &= est\ g^t(x) \\ &= est\ E(\log p \mid x, t). \end{aligned}$$

This simply gives the sample version of the systematic part or the best forecast of $\log p$ in terms of explanatory variables x and time:

(3) $$\hat{\log p} = \hat{\log p}^t(x) = f^t(x).$$

Note that the variable x should not include any t-specification (an index t related to time) or any observation specification (or a subindex such as i or j), because it is just the freely chosen symbol for the "independent " variable in the functional notation. This comment holds for most forthcoming expressions and is not repeated after this. The function $f^t(x)$ should be applicable to just any values of its argument-vector, not just to their observed values (as in estimation) but also to any hypothetical x-values or to x-values from other periods, as will be done after a while.

3.2 The general hedonic model

These general expressions look very simple because we deliberately leave aside, at this stage, all modelling and estimation problems. For instance, all the different functional forms of the HM are hidden in our notation $f^t(x)$.

Different functional forms will be taken up at a later stage of our presentation. This will be done by specialising the general set-up (which is an easy step), so we use here "from general to specific approach" of (econometric) model building.

Essential features of (3) can be represented using its partial derivatives. Assume for simplicity of presentation that all x-variables are continuous[6]. Denote the partial derivative of any x_k by

---

[4] See Spanos (1986)

[5] See Rao (1965, p. 220-249)

[6] For discrete integer-valued variables partial derivatives are replaced by the effects of changing the variable by one unit, while keeping all other variables constant – the ceteris paribus condition. We also assume that $f^t(x)$ is continuously differentiable everywhere, an apparently innocent assumption from the practical point of view.

$$
\begin{aligned}
f_k^t(x) &= \frac{\partial}{\partial x_k} f^t(x) \\
&\approx \frac{f^t(x_1,..., x_k + \tfrac{1}{2}, ..., x_K) - f^t(x_1,..., x_k - \tfrac{1}{2}, ..., x_K)}{(x_k + \tfrac{1}{2}) - (x_k - \tfrac{1}{2})} \\
&= f^t(x_1,..., x_k + \tfrac{1}{2}, ..., x_K) - f^t(x_1,..., x_k - \tfrac{1}{2}, ..., x_K).
\end{aligned} \tag{4}
$$

Note the symmetric way of defining the change in the k´th argument, which is the standard way of starting to approximate the derivative by the difference quotient in numerical analysis, see Comrie (1966, p. 349) or Kahaner - Moler - Nash (1989, p. 30). This approximation gives the ordinary (economic) interpretation of partial derivatives as **the effect of changing the relevant variable by one unit, ceteris paribus.**

In HM's $f_k^t(x)$ is the (estimated) effect of the unit change of k'th variable on $\log p$. Or even more concretely, 100 $f_k^t(x)$ tells how many **log percent** (also denoted by L% or log-%, see Törnqvist – Vartia – Vartia (1985)) the price increases when the quality variable x_k increases by one unit (a semi-elasticity). As the notation shows the derivative $f_k^t(x)$ depends usually on both x and t.

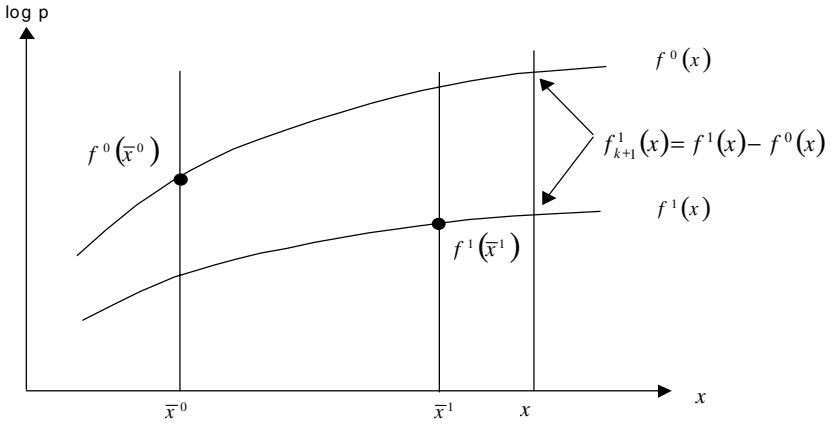
Corollary: We cannot generally assume or restrict the quality effects as independent of time (i.e. to be the same for all time periods) or as independent of the values of quality variables. These are rather restrictive but testable special cases of our general set-up.

For a given value x of the quality vector we denote **the effect of discrete change in time** similarly.

$$
f_{k+1}^t(x) = f^t(x) - f^{t-1}(x) = \log \hat{p}^t(x) - \log \hat{p}^{t-1}(x) = \frac{\Delta f^t}{\Delta t}, \quad \text{where } \bullet t = 1. \tag{5}
$$

Also this 'partial derivative' of time (actually the difference quotient, cf. (4)) is usually a function of both x and t. Note that $f_{k+1}^t(x)$ estimates the change of log-price for a given quality vector[7]. **It is the estimated pure price (PPC) change for that quality vector.** A useful illustration of this fundamental idea of a general hedonic method is given in Figure 1.

Figure 1: Illustration of PPC, the fundamental idea of the general hedonic method



3.3 Standard quality points and the hedonic price index

Here we consider two time periods t = 0 and t = 1 and for the sake of illustration the quantity vector x is taken as one-dimensional. Figure 1 shows also the essential features of HM when x moves in higher dimensions. Although the figure is restricted to one-dimensional x only, we verbalise the results in higher dimensions. These illustrations and later interpretations were developed by Vartia and Kurjenoja (1992) to evaluate wage discrimination between men and women. We have

[7] Differences of time dependent functions are affected by values from both periods and may be "plotted at" or regarded as properties of the latter period (as is usually done) or of the former period (which would be an exceptional interpretation). In numerical analysis these possibilities are distinguished by referring to **backward and forward differences** respectively. Actually the difference should be plotted at a compromise value, namely at the mean value of the periods t and t-1, i.e. at t – ½. Therefore the proper notation for (5) would be $f_{k+1}^{t-1/2}(x)$.

shown also the mean values $\bar{x}^0$ and $\bar{x}^1$ of the quality variables and predictions (or fitted values) of $\log p$, $f^0(\bar{x}^0)$ and $f^1(\bar{x}^1)$ at these mean values. The points $\bar{x}^0$ and $\bar{x}^1$ are referred as **old and new standard quality points (SQP)**, respectively. **The vertical differences of the old and new hedonic functions** $f^0(x)$ and $f^1(x)$ at these SQP´s measure the quality corrected (or pure) price changes (PPC) at these points.

(6) $$f^1(\bar{x}^0) - f^0(\bar{x}^0) = \log P_0^1(\bar{x}^0), \qquad \text{PPC at the old SQP}$$

(7) $$f^1(\bar{x}^1) - f^0(\bar{x}^1) = \log P_0^1(\bar{x}^1), \qquad \text{PPC at the new SQP.}$$

Note that PPC is given in log-change form $\Delta \log p$ or $\log P_0^1$ where P_0^1 is the **hedonic price index (HPI)**. In (6) the log-change of the HPI $\log P_0^1(\bar{x}^0)$ is calculated at the old SQP, while in (7) it is calculated in at the new SQP. Of course in both cases arguments and SQP`s are the same, while the function changes from 0 to 1.**This is the essence of standard quality point method**.

We explicate the interpretation of the components of (6) using (3):

(8) $$\begin{aligned} \log P_0^1(\bar{x}^0) &= \log \hat{p}^1(\bar{x}^0) - \log \hat{p}^0(\bar{x}^0) \\ &= f^1(\bar{x}^0) - f^0(\bar{x}^0)\ . \end{aligned}$$

The **hedonic price functions (HPF)** may be also referred to as **quality valuation functions (QVF)**. Thus $f^1(x) = \log \hat{p}^1(x)$ values the different quality points according to period 1 valuations as $f^0(x) = \log \hat{p}^0(x)$ uses period 0 valuations. Quality valuations are allowed to change in time in general set-up and time invariant special valuations are introduced later as a special assumption.

Using this suggestive terminology (8) leads to the following natural interpretation. The HPI $\log P_0^1(\bar{x}^0)$ compares the old and new quality valuations at the same old SQP. The term $\log \hat{p}^0(\bar{x}^0)$ shows how the old SQP $\bar{x}^0$ was valued in period 0 as $\log \hat{p}^1(\bar{x}^0)$ estimates its valuation using period 1 preferences. Their difference $\log P_0^1(\bar{x})$ measures the pure change in log prices for a constant quality point, as it should. Similarly

(9) $$\begin{aligned} \log P_0^1(\bar{x}^1) &= \log \hat{p}^1(\bar{x}^1) - \log \hat{p}^0(\bar{x}^1) \\ &= f^1(\bar{x}^1) - f^0(\bar{x}^1) \end{aligned}$$

measures also the effect of changing valuations (or log prices) but for another constant quality point, namely SQP $\bar{x}^1$. Thus (8) and (9) are the natural versions of HPI where in both cases, the effects of quality changes have been eliminated or controlled using standard quality points.

These basic ideas are further elaborated using OAXACA-type decompositions and their generalisations, to be presented later.

These most important results (8) - (9) can be easily inferred and remembered using figure 1.

A. If $\bar{x}^0 = \bar{x}^1$ or if there is no (or only little) quality change on the average $\log P^1(\bar{x}^0) = \log P^1(\bar{x}^1)$, the choice of SQP does not matter, rather surprisingly. In this case hedonic modelling does not affect the observed prices as compared to simple comparison because

(10) $$\log P^1(\bar{x}^0) = \log P^1(\bar{x}^1) \approx \log \tilde{P}^1 - \log \tilde{P}^0\ ,$$

although changing qualities may have considerable effect on micro level. In (10) $\tilde{P}^1$ and $\tilde{P}^0$ may be defined either as unit values or other mean prices (preferably geometric means, in which case $\log \tilde{P}^t = \frac{1}{n^t} \sum_{i=1}^{n^t} \log p_i^t$). This result is easily proven when QVF's are linear functions in quality variables.

B. If there are systematic quality changes (i.e. $\bar{x}^0 \neq \bar{x}^1$) but the QVF's (HPF's) are roughly horizontal, quality corrections have only minimal effects, because (10) applies approximately. This may at a first glance seem trivial, because quality variables having only minimal effects are not usually included in hedonic models as they are not considered or called as quality variables. However, if x is K-dimensional, it may contain some quality variables that affect the price considerably whereas other quality variables only have minor effects on the price. If the change $\bar{x}^0 \to \bar{x}^1$ happens to realise only in the direction of quality variables having only minor effects, then essentially we have the same situation as in A and (10) applies.

Combining cases A and B shows, that even if there are strong quality effects on the micro level but either the average quality does not change or the average quality change is realised only for quality variables having minor price effect, then quality corrections are actually not needed. Of course it doesn't make any harm to use HM's even in these cases.

3.4 The linear time-dependent hedonic model

We may restrict our treatment to the standard linear case by 'assuming' or treating HPF's as linear functions of quality variables. This is not as innocent as it is usually regarded, because now a plane approximates a regression surface (1), which in reality is non-linear. This may cause bias of unknown magnitude, which in some cases may not be negligible.

The linear regression models are usually considered as easier to estimate, handle and understand than non-linear ones. This is more like a general attitude towards model building than an established fact. Of course, if the regression surface (1) happens to be a plane, then it is naturally to estimate it by a corresponding linear functional form

$$f^t(x) = a^t + \sum_{k=1}^{K} b_k^t x_k^t \,. \tag{11}$$

Now the partial derivatives have especially simple forms,

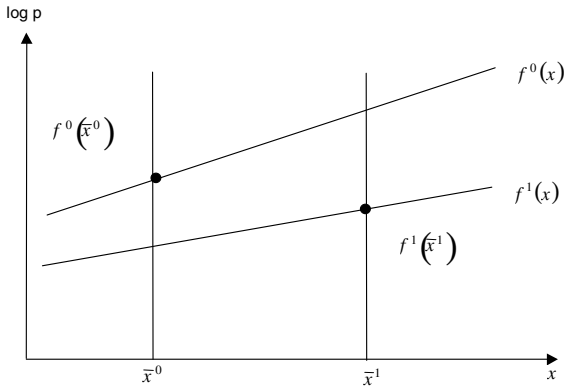
$$f_k^t(x) = b_k^t \,, \tag{12}$$

which are time dependent constants. On the other hand the pure price change (PPC) for a given quality point x, namely

$$f_k^t(x) = f^t(x) - f^{t-1}(x) = (a^t - a^{t-1}) + \sum_{k=1}^{K} \left(b_k^t - b_k^{t-1}\right) x_k \tag{13}$$

which depends on both x and t. This means that pure price changes for a given x are allowed to depend on x. The schematic representation in the previous Figure 1 simplifies in the case of **linear hedonic model (LHM)** as follows:

Figure 2: Illustration of a linear hedonic model



For linear LHM we easily derive the following identities:

Theorem 1. For LHM given in (11) for any set of observations (not necessarily a sample) $x_i = (x_{1i}, ..., x_{Ki}),\ i = 1, ..., n^t$ we have for all t:

(14) $$\frac{1}{n^t}\sum_{i=1}^{n^t} f^t(x_i) = f^t\left(\frac{1}{n^t}\sum x_i\right) \text{ or } \overline{f^t(x)} = f^t(\bar{x}).$$

Proof.

$$\begin{aligned}\overline{f^t(x)} &= \frac{1}{n^t}\sum_{i=1}^{n^t}\left[a^t + \sum_{k=1}^{K} b_k^t x_{ki}\right] \\ &= a^t + \frac{1}{n^t}\sum_{i=1}^{n^t}\sum_{k=1}^{K} b_k^t x_{ki} \\ &= a^t + \frac{1}{n^t}\sum_{k=1}^{K}\sum_{i=1}^{n^t} b_k^t x_{ki} \\ &= a^t + \sum_{k=1}^{K} b_k^t \bar{x}_k = f^t(\bar{x}).\end{aligned}$$

Simply an arithmetic average (denoted by a long bar) of a linear function is the corresponding value o the function at the arithmetic mean value (denoted by a short bar in argument). This is easily remembered by moving the long bar above the function above its argument and making it short. This result holds necessarily for linear functions, but usually fails for non-linear functions. Only by luck (or in very special circumstances) 14 hold for non-linear functions, as will be demonstrated later. The case of non-linear functions is closely related to Jensen´s inequality (see Rao (1968, p.46) or Chung (1968, p. 45, 281)) and Itô´s formula, see Björk (1998, pp. 38-40, 43-48). Jensen´s inequality hods for any convex function $f^t(x)$ of any real random variable having an arbitrary distribution and it states that

(14b) $$Ef^t(x) \geq f^t(Ex)$$

provided that the expectations exist. For any finite set of x-values this implies

(14c) $$\frac{1}{n^t}\sum_{i=1}^{n^t} f^t(x_i) \geq f^t\left(\frac{1}{n^t}\sum x_i\right) \text{ i.e. } \overline{f^t(x)} \geq f^t(\bar{x}).$$

For concave functions the inequality is reversed.

Theorem 2. If LHM is fitted by OLS (or by any other estimation method that forces the sum of residual to zero) to a sample $(\log p_i^t, x_i^t),\ i = 1, ..., n$ from the period t. Then in addition to (14) also

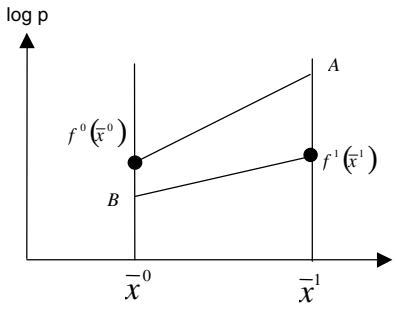
(15) $$\begin{aligned}\overline{\log p^t} &= \frac{1}{n^t}\sum \log p_i^t \\ &= \frac{1}{n^t}\sum \log \hat{p}_i^t\end{aligned}$$

Therefore both $\overline{logp}^t = \overline{\log \hat{p}^t}$ and $\overline{f^t(x^t)} = f^t(\bar{x}^t)$.

Proof. (15) is just the condition that the sum of residuals in $\log p_i^t = \log \hat{p}^t + e_i = f^t(x_i^t) + e_i$ is zero.

Theorem 3. Decomposing the change $f^1(\bar{x}^1) - f^0(\bar{x}^0)$ into HPI and the **effect of changing qualities ECQ**. In Figure 3 we can move from point $(f^0(\bar{x}^0), \bar{x}^0)$ to $(f^1(\bar{x}^1), \bar{x}^1)$ first via A giving (16) and then via B giving 17.

Figure 3:



(16) $$f^1(\bar{x}^1) = f^0(\bar{x}^0) + [f^1(\bar{x}^0) - f^0(\bar{x}^0)] + [f^1(\bar{x}^1) - f^1(\bar{x}^0)]$$

(17) $$\begin{aligned} f^1(\bar{x}^1) - f^0(\bar{x}^0) &= [f^1(\bar{x}^0) - f^0(\bar{x}^0)] + [f^1(\bar{x}^1) - f^1(\bar{x}^0)] \\ &= \log P_0^1(\bar{x}^0) + LQCF^1(\bar{x}^0 \to \bar{x}^1) \end{aligned}$$

where **LQCF** is the **Logarithmic Quality Correction Factor.** Note that (16) and (17) are actually algebraic identities because the "cross terms" – where upper indexes in the HPF and in its argument are different – can be cancelled out. The interpretation to the pair (HPI, ECQ)= $\left(\log P_0^1(\bar{x}^0), LQCF^1(\bar{x}^0 \to \bar{x}^1)\right)$ needs first some effort to be fully understood and remembered.

(18) $$\begin{aligned} f^1(\bar{x}^1) - f^0(\bar{x}^0) &= [f^1(\bar{x}^1) - f^0(\bar{x}^1)] + [f^0(\bar{x}^1) - f^0(\bar{x}^0)] \\ &= \log P_0^1(\bar{x}^1) + LQCF^0(\bar{x}^0 \to \bar{x}^1) \end{aligned}$$

with its obvious interpretation into another pair of (HPI, ECQ) namely $\left(\log P_0^1(\bar{x}^1), LQCF^1(\bar{x}^0 \to \bar{x}^1)\right)$. Note that there is a similar asymmetry or variation in the superscripts as in the well-known decompositions of value ratios into Laspeyres and Paasche indices, where Laspeyres may be applied either as a price or a volume index and Paasche appears always as the other pair.

Methodological comment. Recognising the validity of (17) and (18) is very different from realising the relevance of them. Actually the algebraic "triviality" of (18) is revealed by the following derivation of it. Take any real number z and force zero in the form 0 = z – z in the expression

(19) $$f^1(\bar{x}^1) - f^0(\bar{x}^0) = [f^1(\bar{x}^1) - z] + [z - f^0(\bar{x}^0)]$$

and group terms as indicated. This is an identity for any z. Choose $z = f^0(\bar{x}^1)$ to get (18), while $z = f^1(\bar{x}^0)$ gives (17).

Hence, something important arises from a mathematical triviality. Note that the interpretation does not hold for an arbitrary z in (19), but require the carefully adjusted choice of z giving (17) and (18).

This is an easy formal proof of (17) and (18), which has very little to do with the actual meaning and relevance of the results[8]. Their relevance stems from the decomposition of $f^1(\bar{x}^1) - f^0(\bar{x}^1)$ in terms of the two different (HPI, ECQ) –pairs $\left(\log P_0^1(\bar{x}^0), LQCF^1(\bar{x}^0 \to \bar{x}^1)\right)$ and $\left(\log P_0^1(\bar{x}^1), LQCF^0(\bar{x}^0 \to \bar{x}^1)\right)$ included in the illustrations. This simple identity is a generalisation of OAXACA-decomposition that is traditionally given for linear HPF's.

Theorem 4. Generalisation of theorem 3 to non-linear hedonic price function HPF. Decompositions (17) and (18) and their proofs are valid as such **also for non-linear HPF's** as illustrated in Figure 1.

Note, however, that Theorems 1, 2 and 5 do not hold for non-linear HPF´s, see Lemma 3 and Theorem 6.

[8] Cf. Wittgenstein (1967, p.92e): "Each proof proves not merely the truth of the proposition proved, but also that it can be proved *in this way.*"

We have decided to postpone the derivation of more familiar and in practice important linear HPF's so that the reader may better appreciate the relevance and meaning of the decompositions above.

Theorem 5. If linear HPF:s $f^0(x)$ and $f^1(x)$ are fitted separately to t-specific samples $(\log p_i^t, x_i^t)$, t = 0 and 1, $i = 1, ..., n_t$ using OLS (or any other estimation method which forces the sum of residuals to zero) then the following (HPI, ECP) –decompositions hold as identities

$$\begin{aligned} \overline{\log p}^1 - \overline{\log p}^0 &= f^1(\bar{x}^1) - f^0(\bar{x}^0) \\ &= \log P_0^1(\bar{x}^0) + LQCF^1(\bar{x}^0 \to \bar{x}^1) \end{aligned} \tag{20}$$

$$\begin{aligned} \overline{\log p}^1 - \overline{\log p}^0 &= f^1(\bar{x}^1) - f^0(\bar{x}^0) \\ &= \log P_0^1(\bar{x}^1) + LQCF^0(\bar{x}^0 \to \bar{x}^1) \end{aligned} \tag{21}$$

Proof. Combine theorems 1, 2 and 3.

Before going any further, we present here empirical results based on the above considerations:

Empirical example 1: Standard Quality point approach, linear time dependent model

Estimated average log prices, HPI´s and LQCF´s

		HPF or QVF based on:		Hedonic Price Index, HPI	
Standard Quality point, SQP		Spring valuations	Fall valuations	Pure change in log-units	in prices in log-percents
Spring (old SQP)		**7,157**	7,080	-0,077	-7,7 log-%
Fall (new SQP)		7,261	**7,167**	-0,094	-9,4 log-%
LQCF (=ECQ)	in log-units	0,105	0,087		
	in log-percents	10,5 log-%	8,7 log-%		

Actual observed mean values (cf. Figure 2) are given in **bold**.

Estimated geometric mean prices, HPI´s and QCF´s
In original units (€)

		HPF or QVF based on:		Hedonic Price Index, HPI	
Standard Quality point		Spring valuations	Fall valuations	Pure change in €	in prices in per cent
Spring (old SQP)		**1 283**	1 188	-95	-7,4 %
Fall (new SQP)		1 424	**1 296**	-128	-9,0 %
QCF (=ECQ)	in €	141	108		
	in per cent	11,0 %	9,1 %		

Actual observed mean values (cf. Figure 2) are given in **bold**.

We present also the results showing exact correspondence with our decompositions and notations:

Logarithmic Standard Quality Point		**Logarithmic Pure Price Index**		**Logarithmic Quality Correction Factor**		**Sum of these**	
Linear time dependent HPF´s or QVF´s							
	SQP	**PPI**		**LQCF**		**Sum of these**	
Name	Notation	Notation	Value	Notation	Value	Value	Equations
Old SQP	$\bar{x}^0$	$\log P_0^1\left(\bar{x}^0\right)$	-0,077	$LQCF^1\left(\bar{x}^0 \to \bar{x}^1\right)$	0,087	0,010	(17) & (53)
New SQP	$\bar{x}^1$	$\log P_0^1\left(\bar{x}^1\right)$	-0,094	$LQCF^0\left(\bar{x}^0 \to \bar{x}^1\right)$	0,105	0,010	(18) & (55)
Weighted average of decompositions (17) and (18):							
Overall SQP	$\bar{x}$	$\log P_0^1\left(\bar{x}\right)$	**-0,086**	$LQCF^{1/2}\left(\bar{x}^0 \to \bar{x}^1\right)$	**0,096**	**0,010**	(61)

3.5 The non-linear time-dependent hedonic model

Theorem 6. If non-linear HPF's $f^0(x)$ and $f^1(x)$ are fitted to t-specific samples using any estimation method, which forces the sum of residuals to zero, then (20) and (21) are not usually identities but hold approximately. The approximation error is related to the first equality (while the latter holds still as an identity) and depends on the degree of curvature (the second order partial derivatives of $f^0(x)$ and $f^1(x)$), at $\bar{x}^0$ and $\bar{x}^1$ and on the variances and covariances of the x-variables.

In fact the approximation error depends on the differences (or changes) of the second partial derivatives and second central moments (variances and covariances) between the periods considered.

This technique makes use of some rather straightforward properties on non-linear approximation theory, that are applicable in much wider contexts. As they seem to be not well known we will demonstrate their usefulness in an elementary way.

Lemma 1. Let us start from the simple case of two-dimensional x-vector $x=(x_1,x_2)$ and consider a function $f^t(x)=f^t(x_1,x_2)$ that is quadratic in these two arguments. Take any fixed point $\bar{x}=(\bar{x}_1,\bar{x}_2)$, not necessarily the mean of the variables at this stage, and consider the following **special representation** of $f^t:\Re^2\to\Re$

$$
\begin{aligned}
f^t(x) &= a^t + b_1^t\left(x_1-\bar{x}_1\right)+b_2^t\left(x_2-\bar{x}_2\right)\\
&\quad +\tfrac{1}{2}c_1^t\left(x_1-\bar{x}_1\right)^2+\tfrac{1}{2}c_2^t\left(x_2-\bar{x}_2\right)^2\\
&\quad +\tfrac{1}{2}d^t\left(x_1-\bar{x}_1\right)\left(x_2-\bar{x}_2\right)
\end{aligned}
\qquad (22)
$$

This representation[9] of $f^t(x)$ is unique and has unique coefficients, which depend, of course, of the point $\bar{x}=(\bar{x}_1,\bar{x}_2)$ chosen. Consider two **t-specific samples** $\left(y_i^t, x_{1i}^t, x_{2i}^t\right)$, t = 0 or 1, $i=1,\ldots,n^t$ and suppose that $f^t(x_1,x_2)$ is fitted to them separately using any estimation method, that forces the sum of residuals to zero so that assumptions of the theorem 6 apply.

If we now specify the point above $\bar{x}=(\bar{x}_1,\bar{x}_2)$ as the **t-specific mean** $\bar{x}=\bar{x}^t=\left(\bar{x}_1^t,\bar{x}_2^t\right)=\left(\frac{1}{n_t}\sum x_{1i}^1,\frac{1}{n_t}\sum x_{2i}^1\right)$ we derive in this simple special case quadratic function of two variables a most important result, which generalises to arbitrary twice continuously differentiable functions of any number of arguments:

$$
\begin{aligned}
&\overline{f^t\left(x_1^t,x_2^t\right)}-f^t\left(\bar{x}_1^1,\bar{x}_2^1\right)=\bar{y}^t-f^t\left(\bar{x}_1^1,\bar{x}_2^1\right)\\
&=\tfrac{1}{2}\left[c_1^t Var\left(x_1^t\right)+c_2^t Var\left(x_2^t\right)+d^t Cov\left(x_1^t,x_2^t\right)\right]=\Delta^t
\end{aligned}
\qquad (23)
$$

[9] Note that any valid representation of any function f describes and defines the function completely and therefore any of its (infinitely many) representations reveals all the properties of f. The function f as a mapping should not be mixed with any of its representations, because it's something they all describe in a specific way. Much confusion arises when this is not understood.

This **functional equation** is the basic mathematical result leading to both univariate and multidimensional **Itô's lemma**, which is the basic tool in continuous time finance models, see Björk (1997). It shows, **how the mean of values of a function** of several (here two) variables is expressed in terms of **value of the function calculated at mean value of its arguments**. These values are not the same (unless the function is linear, or the arguments have no variation) although our intuition sometimes fails to notice this. The difference (23) of these two depends on Δ^t or of the second order derivatives or the curvatures of the function (at the mean point as will be seen later) and on the corresponding variances and covariances of the sample of argument points.

We repeat (23) in a more intuitive form for an easy reference

$$\begin{aligned} &\overline{f^t\left(x_1^t, x_2^t\right)} = f^t\left(\bar{x}_1^t, \bar{x}_2^t\right) + \Delta^t, \text{ where} \\ &\Delta^t = \tfrac{1}{2}\left[c_1^t Var\left(x_1^t\right) + c_2^t Var\left(x_2^t\right) + d^t Cov\left(x_1^t, x_2^t\right)\right] \end{aligned} \tag{23b}$$

This is easy to visualise using the concepts of **the long upper bar** (which refers to the mean of the fitted values) and the **short upper bar** (which refers to the mean of the arguments of the function). If the long upper bar is substituted by a short upper bar for its arguments, the covariance - curvature-term (of the Itô-type) Δ^t must be added as a correction.

Proof. Consider

$$\begin{aligned} n^t \bar{y}^t &= \sum y_i^t = \sum\left[f^t\left(x_{1i}^t, x_{2i}^t\right) + e_i^t\right] = \sum f^t\left(x_{1i}^t, x_{2i}^t\right) + 0 \\ &= \sum\left[a^t + b_1^t\left(x_1^t - \bar{x}_1^t\right) + b_2^t\left(x_2^t - \bar{x}_2^t\right) + \tfrac{1}{2}c_1^t\left(x_1^t - \bar{x}_1^t\right)^2 + \tfrac{1}{2}c_2^t\left(x_2^t - \bar{x}_2^t\right)^2 + \tfrac{1}{2}d^t\left(x_1^t - \bar{x}_1^t\right)\left(x_2^t - \bar{x}_2^t\right)\right] \\ &= n^t\left[a^t + \tfrac{1}{2}c_1^t Var\left(x_1^t\right) + \tfrac{1}{2}c_2^t Var\left(x_2^t\right) + \tfrac{1}{2}d^t \operatorname{cov}\left(x_1^t, x_2^t\right)\right] = n^t \Delta^t, \end{aligned}$$

where $Var\left(x_k^t\right) = \frac{1}{n^t}\sum\left(x_k^t - \bar{x}_k^t\right)^2$, k = 1, 2.[10] Dividing by n^t and noting that $a^t = f^t\left(\bar{x}_1^t, \bar{x}_2^t\right)$ gives (23).

A straightforward consequence of Lemma 1 is

Lemma 2. Assume the same as in Lemma 1.

$$\begin{aligned} f^{\overline{1}}\left(x_1^t, x_2^t\right) - f^{\overline{0}}\left(x_1^t, x_2^t\right) &= \left[f^1\left(\bar{x}_1^1, \bar{x}_2^1\right) + \Delta^1\right] - \left[f^0\left(\bar{x}_1^0, \bar{x}_2^0\right) + \Delta^0\right] \\ &= \left[f^1\left(\bar{x}_1^1, \bar{x}_2^1\right) - f^0\left(\bar{x}_1^0, \bar{x}_2^0\right)\right] + \left[\Delta^1 - \Delta^0\right] \\ &\approx f^1\left(\bar{x}_1^1, \bar{x}_2^1\right) - f^0\left(\bar{x}_1^0, \bar{x}_2^0\right), \end{aligned} \tag{24}$$

where the difference of the covariance-type terms is

$$\begin{aligned} \left[\Delta^1 - \Delta^0\right] = &\tfrac{1}{2}\left[c_1^1 Var\left(x_1^1\right) + c_2^1 Var\left(x_2^1\right) + 2d^1 \operatorname{cov}\left(x_1^1, x_2^1\right)\right] \\ &- \tfrac{1}{2}\left[c_1^0 Var\left(x_1^0\right) + c_2^0 Var\left(x_2^0\right) + 2d^0 \operatorname{cov}\left(x_1^0, x_2^0\right)\right]. \end{aligned} \tag{24b}$$

Note from the first line of (24), how Δ^t move the short upper bar versions from their long upper bar averages. However, when terms are grouped as in the second line, differences between long and short bar versions will be good approximations for each other, because the difference (24b) is usually nearly zero. This approximation must be good if $c_k^0 \approx c_k^1$, $d^0 \approx d^1$ and $Var\left(x_k^0\right) \approx Var\left(x_k^1\right)$, $\operatorname{cov}\left(x_1^0, x_2^0\right) \approx \operatorname{cov}\left(x_1^1, x_2^1\right)$, (k = 1, 2), i.e., the curvatures and the sample covariances are roughly time invariant. This usually holds in practice. Equation (24) is even closer to Itô's lemma than Lemma 1.

Next we generalise the Lemmas 1 and 2 for arbitrary quadratic functions in K-dimensional spaces. Let's explain first why the representation (22) was chosen for $f^t : \Re^2 \to \Re$.
Calculate the first two partial derivatives of it:

[10] Here we have n^t and not $n^t - 1$ as a divisor here and in the sample covariance. First degree terms vanish because $\sum\left(x_k^t - \bar{x}_k^t\right) = \sum x_k^t - n^t \bar{x}_k^t = 0$.

(25) $$f_k^t(\bar{x}_1,\bar{x}_2)=\frac{\partial}{\partial x_k}f^t(\bar{x}_1,\bar{x}_2)=b_k^t \quad (k=1,2)$$

(26) $$f_{k,l}^t(\bar{x}_1,\bar{x}_2)=\frac{\partial^2}{\partial x_k\partial x_l}f^t(\bar{x}_1,\bar{x}_2)=d^t \quad \text{(if k and l are different)}$$

$$f_{k,k}^t(\bar{x}_1,\bar{x}_2)=\frac{\partial^2}{\partial x_k\partial x_k}f^t(\bar{x}_1,\bar{x}_2)=c_k^t$$

We see e.g. that parameters $(b_1^t,b_2^t,c_1^t,c_2^t,d^t)$ are exactly the first and second partial derivatives of $f^t(x_1,x_2)$ calculated at the chosen point $\bar{x}=(\bar{x}_1,\bar{x}_2)$, not necessarily the mean point. This explains why also the multiplier ½ appears in (22).

Now return to the general case where $x=(x_1,...,x_K)$ and $\bar{x}=(\bar{x}_1,...,\bar{x}_K)$. A straightforward generalisation (which is formally simpler than lemma 1) is:

Lemma 3. Let $f^t:\Re^K\to\Re$ be any continuously twice differentiable function fitted in such a way to the samples $(y_i^t,x_{1i}^t\ ...,\ x_{Ki}^t)$, i = 1, …, n^t, that the sum of residuals equals zero. Then we have

(27) $$\bar{y}^t-f^t(\bar{x}^t)=\overline{f^t(x^t)}-f^t(\bar{x}^t)\approx\Delta^t$$

where the Itô-type covariance-curvature term equals

(27b) $$\begin{aligned}\Delta^t&=\sum_{k=1}^K\frac{1}{2}f_{k,k}^t(\bar{x}^t)Var(x_k^t)+\sum_{k=1}^K\sum_{l<k}f_{k,l}^t(\bar{x}^t)\mathrm{cov}(x_k^t,x_l^t)\\&=\tfrac{1}{2}\sum_{k=1}^K\sum_{l=1}^K f_{k,l}^t(\bar{x}^t)\mathrm{cov}(x_k^t,x_l^t).\end{aligned}$$

Note, that (27) is even simpler in its notation than its two-dimensional special version. Now we see in a very compact form, how an average of fitted values $\overline{f^t(x^t)}$ is approximated by the value of the function $f^t(\bar{x}^t)$ at the average point of its argument vector. Note that for linear functions these must be equal. For non-linear functions Δ^t appears and by Jensen´s inequality it must be positive (negative) for convex (concave) functions.

Equation becomes an identity for all quadratic functions $f^t:\Re^K\to\Re$ and a second degree Taylor-approximations for any function (having continuous second partial derivatives in the neighbourhood of x, see Apostol (1957, p. 124)). Note that covariances $\mathrm{cov}(x_k^t,x_{li}^t)$ are calculated 'only once' in the middle expression (where l > k and multiplier ½ appears asymmetrically) but appears two times in the last expression.[11]

The last representation is the most symmetric one and corresponds to the Taylor-expression of several variables.
Lemma 3 contains all previous cases and other interesting results as its special cases. We review these shortly

1. If $f^t(x)$ is (at most) quadratic but second order derivatives $f_{k,l}^t(x)$ vanish, we are back to the linear case. Theorems 1 – 5 are derived as a special case because all Σ-expressions vanish and $\bar{y}^t-f^t(\bar{x}^t)=\overline{f^t(x^t)}-f^t(\bar{x}^t)\equiv 0$. Also

(28) $$\bar{y}^t=\frac{1}{n^1}\sum_{i=1}^{n^1}\log\ p_i^t=\overline{\log\ p}^t$$

2. We rewrite (27) for any $f^t(x)$ well approximated by a quadratic Taylor-approximation

[11] These problems stem from the identity $\frac{1}{2}(x_1+x_2)^2=\frac{1}{2}x_1^2+\frac{1}{2}x_2^2+x_1x_2=\frac{1}{2}(x_1^2+x_1x_2+x_2x_1+x_2^2)$, where $\frac{1}{2}x_2x_1$ is calculated 'only once' in the middle expression but 'appears twice' in the last expression. These three different representations define the same quadratic function $f(x_1,x_2)$. We must always decide which of them (or perhaps something else) we are using and not get confused.

(29) $$\overline{\log p^t} = \overline{f^t(x^t)} = f^t(\bar{x}^t) + \Delta^t \quad (t = 0,1)$$

Here the first equality holds strictly because sum of residuals was forced to zero in estimation but second holds only as an approximation. Subtracting equations for different time periods from each other and arranging terms we get

(30) $$\begin{aligned} \overline{\log p^1} - \overline{\log p^0} &= \frac{1}{n^1}\sum_{i=1}^{n^1} \log p_i^1 - \frac{1}{n^0}\sum_{i=1}^{n^0} \log p_i^0 \\ &= \overline{f^1(x^1)} - \overline{f^0(x^0)} \end{aligned}$$

These are strict equalities because sums of residuals vanished. This rather surprising result holding for arbitrary $f^t(x)$´s ,which seems to be unknown in hedonic literature. Because $\overline{\log p}^t = \log G(p^t)$, the logarithm of geometric mean of prices[12] we also have an interesting result

(31) $$\log \frac{G(p^1)}{G(p^0)} = \overline{f^1(x^1)} - \overline{f^0(x^0)}$$

which holds without error for any non-linear HPF's fitted separately to respective samples (when residuals sum to zero). Using the approximation part of (29) we get finally

(32) $$\begin{aligned} \overline{f^1(x^1)} - \overline{f^0(x^0)} &= \left[f^1(\bar{x}^1) + \Delta^1\right] - \left[f^0(\bar{x}^0) + \Delta^0\right] \\ &= \left[f^1(\bar{x}^1) - f^0(\bar{x}^0)\right] + \left[\Delta^1 - \Delta^0\right] \\ &\approx \left[f^1(\bar{x}^1) - f^0(\bar{x}^0)\right] \end{aligned}$$

(33) $$\delta_0^1 = \Delta^1 - \Delta^0 = \tfrac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} f_{k,l}^1(\bar{x}^1)\mathrm{cov}(x_k^1, x_l^1) - \tfrac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} f_{k,l}^0(\bar{x}^0)\mathrm{cov}(x_k^0, x_l^0)$$

Look at figure 4 for an illustration of (32). The vertical lines at old and new SQP´s meet the functions at **black** crossing points, namely at "the short bar points" $f^0(\bar{x}^0)$ and $f^1(\bar{x}^1)$.The hollow points below them have drifted away because of (29), but anyhow the **differences** between long and short bar points remain approximately equal. This is a rather involved deduction, although all its parts are almost self-evident. In practice expression $\delta_0^1 = \left[\Delta^1 - \Delta^0\right]$involving covariance terms will cancel away (because they are differences of similar terms of period 1 minus period 0). Therefore a simple but powerful theorem results.

Theorem 6. For any non-linear HPF's fitted separately to respective samples we have (32). Usually even δ_0^1 may be neglected in which case

(34) $$\begin{aligned} \log \frac{G(p^1)}{G(p^0)} = \overline{f^1(x^1)} - \overline{f^0(x^0)} &= \frac{1}{n^1}\sum f(x_i^1) - \frac{1}{n^0}\sum f(x_i^0) \\ \approx f^1(\bar{x}^1) - f^0(\bar{x}^0) &= f^1\left(\frac{1}{n^1}\sum x_i^1\right) - f^1\left(\frac{1}{n^0}\sum x_i^0\right), \end{aligned}$$

where $x = (x_1, ..., x_K)$ is an arbitrary vector of quality variables. Furthermore, the difference $f^1(\bar{x}^1) - f^0(\bar{x}^0)$ is decomposed in the natural way e.g. as follows

(35) $$\begin{aligned} f^1(\bar{x}^1) - f^0(\bar{x}^0) &= \log P_0^1(\bar{x}^1) + LQCF^0(\bar{x}^0 \to \bar{x}^1) \\ &= \left[f^1(\bar{x}^1) - f^0(\bar{x}^1)\right] + \left[f^0(\bar{x}^1) - f^0(\bar{x}^0)\right] \end{aligned}$$

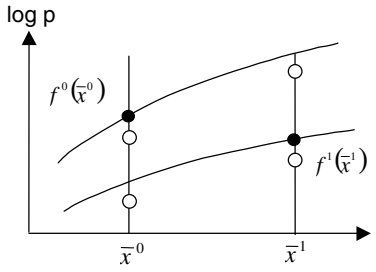
[12] Equations (30) and (31) together explain, why exactly the unweighted geometric mean should be used to aggregate prices (or, equivalently, price ratios) on a micro index level, as is generally suggested

(36)
$$f^1(\bar{x}^1) - f^0(\bar{x}^0) = \log P_0^1(\bar{x}^0) + LQCF^1(\bar{x}^0 \to \bar{x}^1)$$
$$= \left[f^1(\bar{x}^0) - f^0(\bar{x}^0)\right] + \left[f^1(\bar{x}^1) - f^1(\bar{x}^0)\right]$$

These decompositions focus at particular parts of HPF´s namely at the old and new SQP´s (or better at old and new observations) and forget almost everything else. In a more symmetric treatment a weighted mean of these decompositions (weighted by numbers of observations) is taken. This leads to a similar decomposition, where the terms in the resulting (HPI, ECQ)-pair are weighted means of corresponding terms in (34) and (35). It may be shown by direct calculations or by imputations in both directions to be considered later, that this leads to (HPI, ECQ) = ($\log P_0^1(\bar{x}), LQCF^{1/2}(\bar{x}^0 \to \bar{x}^1)$), where $\bar{x}$ is the overall mean of quality variables or the weighted mean of old and new SQP´s.

Now we have derived decompositions for three natural choices of SQP´s. Of course, these are only three possibilities. In hedonic imputation all observed quality points are used as SQP´s to produce PPC´s for all prices, which actually means comparing distances of old and new HPF´s for all data points and taking the average of these as the HPI. We return to imputation techniques in Chapter 3.7.

Figure 4: Illustration of (34)- (36)



Equations (35) and (36)are identities and their interpretations as quality controlled price indices $\log P_0^1(\bar{x}^1)$ and $\log P_0^1(\bar{x}^0)$ and as **Logarithmic Quality Correction Function** $LQCF^0(\bar{x}^0 \to \bar{x}^1)$ and $LQCF^1(\bar{x}^0 \to \bar{x}^1)$ are evident from the figure. For arbitrary linear HPF's with time dependent coefficients approximation in (34) turns out to an equality and all equations (34)-(35) are identities. Note that in all these equations quality variables are actually K-dimensional vectors and figure 4 is intended to reveal only their essential features.

To calculate SQP-estimates based in terms like $f^1(\bar{x}^1)$and $f^0(\bar{x}^1)$, one has to be careful to able to concentrate on its correct interpretation, which differs from common intuitive interpretations. We have based our treatment to a rather abstract application of general mathematical concepts, although the expressions may not communicate this and may appear at the first glance even trivial. To give a concrete example we consider our simple quadratic function of our second empirical example, where $x = (x_1, x_2)$ is two-dimensional and therefore e.g. $\bar{x}^0 = (\bar{x}_1^0, \bar{x}_2^0)$ is the old SQP. This $f^1(x) = f^1(x_1, x_2)$ was estimated using the standard OLS-method in the form

(36b)
$$f^1(x) = f^1(x_1, x_2) = a^1 + b_1^1 x_1 + b_2^1 x_2 + c^1 (x_2 - \bar{x}_2^1)^2,$$

where x_2 = lnspeed and its squared deviation from its new mean (see above) was treated as the third (pseudo)variable, as it should **in OLS estimation**. However, $f^1(x) = f^1(x_1, x_2)$ **must be treated as a function of two independent arguments**. It should not be considered as a function of three independent variables, as it appears in (36b) especially if it is written using a new (pseudo)variable $z = (x_2 - \bar{x}_2^1)^2$, if we "specify" our regression equation in the following way:

(36c)
$$f^1(x, z) = f^1(x_1, x_2, z) = a^1 + b_1^1 x_1 + b_2^1 x_2 + c^1 z\,.$$

This is the usual way of representing non-linear regression functions **when they are estimated**, but some serious problems may arise from this convention. Here z cannot be treated "as an independent variable", because its values are determined exactly by x_2 . If this is not taken into account (which may happen by accident, if some standard routines in statistical programs like SAS are used) serious errors occur without us noticing anything.

Simply we have

(36d) $$f^1(\bar{x}^1) = f^1(\bar{x}_1^1, \bar{x}_2^1) = a^1 + b_1^1\bar{x}_1^1 + b_2^1\bar{x}_2^1 + c^1(\bar{x}_2^1 - \bar{x}_2^1)^2 = a^1 + b_1^1\bar{x}_1^1 + b_2^1\bar{x}_2^1 + 0\,,$$

where we have an interesting representation of zero in the last component of the middle expression!

On the other hand, we get a perhaps oddly looking formula for the "cross term", where upper indices of the function and the SQP differ:

(36e) $$f^1(\bar{x}^0) = f^1(\bar{x}_1^0, \bar{x}_2^0) = a^1 + b_1^1\bar{x}_1^0 + b_2^1\bar{x}_2^0 + c^1(\bar{x}_2^0 - \bar{x}_2^1)^2\,.$$

Note that coefficients are from period 1 function and the quadratic term includes, of course, the new mean of the second quality variable, but the mean point for which the value of the function is calculated is the old SQP. Therefore (36e) includes this quadratic term, which is not included in (36d).

We want to emphasise , that a particular representation of a function (which is a **concept totally independent** of the way it is expressed, i.e. independent of the particular representation chosen e.g. to estimate it in regression analysis) should not be allowed to confuse the treatment. Ordinarily, economists and statisticians seem to totally unaware of the need of this distinction, as they manipulate long and complicated mathematical expressions in a mechanical way. Therefore, a function and (its infinitely many possible) representations or mathematical expressions must be carefully distinguished from each other whenever possible, by sticking firmly to **the general mathematical notation of a function** without even mentioning the complicated and partly arbitrary expressions used to specify it. This is an interesting case of **the identification problem**: a function can never identify its representation, because there are infinitely many representations even for the simplest functions, say for a function f(x) = 3 + 6x defined for all real numbers. We have e.g. f(x) = 3 + 6a +6(x – a) = 6(x + ½) = x(6 + 1/(2x)) etc., which may be especially suitable for particular purposes in revealing different properties of this f(x). Especially non-linear transformations of variables (say in non-linear co-ordinate transformations) easily confuse careless researchers.

Empirical example 2: Standard Quality point approach, non-linear time dependent model

Estimated average log prices, HPI´s and LQCF´s

		HPF or QVF based on:		Hedonic Price Index, HPI	
Standard Quality point, SQP		Spring valuations	Fall valuations	Pure change in log-units	in prices in log-percents
Spring (old SQP)		**7,133**	7,064	-0,069	-6,9 log-%
Fall (new SQP)		7,251	**7,137**	-0,115	-11,5 log-%
LQCF (=ECQ)	in log-units	0,118	0,073		
	in log-percents	11,8 log-%	7,3 log-%		

Estimated geometric mean prices, HPI´s and QCF´s
In original units (€)

		HPF or QVF based on:		Hedonic Price Index, HPI	
Standard Quality point		Spring valuations	Fall valuations	Pure change in €	in prices in per cent
Spring (old SQP)		**1 253**	1 169	-84	-6,7 %
Fall (new SQP)		1 410	**1 257**	-153	-10,9 %
QCF (=ECQ)	in €	157	88		
	in per cent	12,5 %	7,5 %		

These results and their average decomposition are shown below in a more accurate notation.

Logarithmic Standard Quality Point	**Logarithmic Pure Price Index**	**Logarithmic Quality Correction Factor**	**Sum of these**

Non-linear time dependent HPF´s or QVF´s

	SQP	**PPI**		**LQCF**		**Sum of these**	
Name	Notation	Notation	Value	Notation	Value	Value	Equations
Old SQP	$\bar{x}^0$	$\log P_0^1\left(\bar{x}^0\right)$	-0,069	$LQCF^1\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0,073	0,004	(35)
New SQP	$\bar{x}^1$	$\log P_0^1\left(\bar{x}^1\right)$	-0,115	$LQCF^0\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0,118	0,003	(36)
Weighted average of decompositions (35) and (36):							
Overall SQP	$\bar{x}$	$\log P_0^1\left(\bar{x}\right)$	**-0,092**	$LQCF^{1/2}\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	**0,096**	**0,004**	(61)

3.6 Decompositions for time-invariant quality valuations

We are ready to pass on to more popular and restricted HM's where quality valuations are time invariant. In this case HPF's (or QVF's) are **separable in time and quality dimensions**, which is represented by the following property of $f^t(x)$:

(37) $$f^t(x) = a^t + f(x),$$

where a^t does not depend on x and $f(x)$ does not depend on t. Now the partial derivatives of $f^t(x)$ satisfy

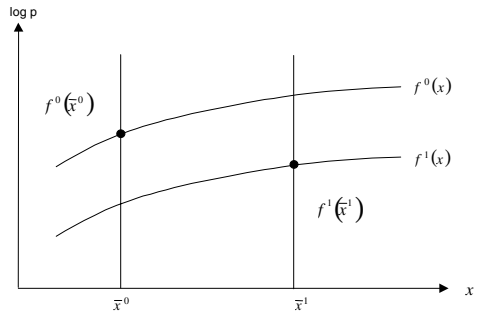
(38) $$f_k^t(x) = f_k(x),$$ independent of t

(39) $$f_{k+1}^t(x) = f^t(x) - f^{t-1}(x) = a^t - a^{t-1},$$ independent of both t and x.

The pure price chance PPC for a given x is now independent of x, because HPF's for different periods are just shifted versions of a time invariant QVF $f(x)$ because of (37).

This is a very intuitive, easy and beautiful case, but may not adequately represent actual quality valuations. Therefore, **its validity cannot be decided a priori**, but should be tested against the previous more general HM's, where quality valuations were allowed to change in time. Such tests are standard procedures in regression analysis.

Figure 5: Illustration of non-linear Griliches-type hedonic models (non-linear GTHM)



If we specify further the model (37) by restricting $f(x)$ to a linear function of the quality variables x we get the Griliches type HM (GTHM):

(40) $$f^t(x) = a^t + \sum_{k=1}^{K} b_k x_k \ .$$

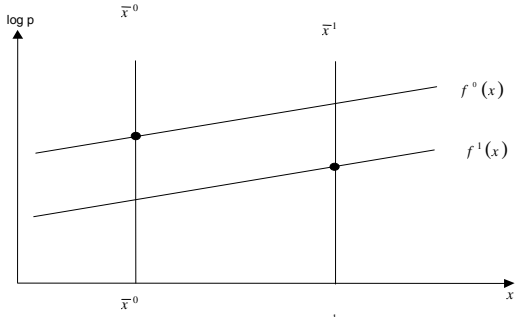
Here the derivatives are very simple

(41) $$f_k^t(x) = b_k \qquad \text{(independent of x and time)}$$

This means that the valuation of quality components is constant in cross-section and is time-invariant. I. e: For all possible quality points, the **pure price change** is the same constant (the assumption of "constant price change").

(42) $$f_{k+1}^t(x) = a^t - a^{t-1} \qquad \text{(independent of x)}$$

Figure 6: Illustration of the linear Griliches-type hedonic model (linear GHTM)



In GTHM HPF's are non-linear functions of x or planes. They are furthermore shifted versions of one time invariant QVF $f(x) = \sum b_k x_k$, which is a linear function of all quality variables x_k .
Of course, all our previous results hold for GTHM or its non-linear generalisations (37) referred to as non-linear GTHM. We present the main results for GTHM's as

Theorem 7. **Main theorem for time invariant quality valuations.**

Consider any non-linear HPF's having time invariant quality valuations satisfying (37). When these have been fitted to respective time specific samples using any estimation method forcing the sum of residuals to zero, results of theorem 6 apply. Especially have here

(43) $$\begin{aligned} \log\frac{G(p^1)}{G(p^0)} &= \overline{f^1(x^1)} - \overline{f^0(x^0)} \\ &= \left[a_1 + \overline{f(x^1)}\right] - \left[a_0 + \overline{f(x^0)}\right] \\ &= [a_1 + a_0] + \left[\overline{f(x^1)} - \overline{f(x^0)}\right] \end{aligned}$$

The long bars can be transformed to short bars using (36) and (37) giving

(44) $$\begin{aligned} \overline{f^1(x^1)} - \overline{f^0(x^0)} &= [a_1 + a_0] + \left[\overline{f(x^1)} - \overline{f(x^0)}\right] \\ &\approx \left[a^1 + a^0\right] + \left[f(\bar{x}^1) - f(\bar{x}^0) + d_0^1\right], \end{aligned}$$

where $d_0^1 = \Delta^1 - \Delta^0$ is a usually negligible difference of the covariance terms (or the Itô-term) satisfying

(45) $$\begin{aligned} d_0^1 = \Delta^1 - \Delta^0 &= \tfrac{1}{2}\sum_{k=1}^K\sum_{l=1}^K f_{k,l}^1(\bar{x}^1)\operatorname{cov}(x_k^1, x_l^1) - \tfrac{1}{2}\sum_{k=1}^K\sum_{l=1}^K f_{k,l}^0(\bar{x}^0)\operatorname{cov}(x_k^0, x_l^0) \\ &= \tfrac{1}{2}\sum_{k=1}^K\sum_{l=1}^K f_{k,l}(\bar{x}^1)\operatorname{cov}(x_k^1, x_l^1) - \tfrac{1}{2}\sum_{k=1}^K\sum_{l=1}^K f_{k,l}(\bar{x}^0)\operatorname{cov}(x_k^0, x_l^0). \end{aligned}$$

Here $f_{k,l}(x) = \dfrac{\partial^2}{\partial x_k \partial x_l} f(x)$ are the second derivatives of the time invariant QVF $f(x)$. If $f(x)$ is **quadratic in x** (44) holds as an **identity**, and generally it is a second degree Taylor-approximation for an arbitrary time invariant non-linear QVF $f(x)$.

Neglecting the usually minor difference of the covariance terms $d_0^1 = \Delta^1 - \Delta^0$, we have

(46)
$$\begin{aligned}\overline{f^1(x^1)} - \overline{f^0(x^0)} &\approx \left[a^1 + a^0\right] + \left[f(\bar{x}^1) - f(\bar{x}^0)\right] \\ &\equiv \log P_0^1 + LQCF(\bar{x}^0 \to \bar{x}^1)\end{aligned}$$

where the quality corrected price index $\log P_0^1 = \left[a^1 + a^0\right]$ is now **independent of the standard quality point** (e.g. $\log P_0^1(\bar{x}^1) = \log P_0^1(\bar{x}^0) = \left[a^1 + a^0\right]$) and the logarithmic quality correction factor

(47)
$$f(\bar{x}^1) - f(\bar{x}^0) = LQCF(\bar{x}^0 \to \bar{x}^1)$$

is **time invariant** (e.g. $LQCF^0(\bar{x}^0 \to \bar{x}^1) = LQCF^1(\bar{x}^0 \to \bar{x}^1) = f(\bar{x}^1) - f(\bar{x}^0)$).

Empirical example 3: Estimation results, non-linear Griliches-type model.

Root MSE	0.12459	R-Square	0.6550
Dependent Mean	7.16158	Adj R-Sq	0.6462
Coeff Var	1.73970		

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.40513	0.39743	3.54	0.0005
lnspeed2c	2nd order speed c	1	0.73947	0.23409	3.16	0.0019
lnmem	log of memory	1	0.26552	0.04551	5.83	<.0001
lnspeed	log of speed	1	0.73456	0.06694	10.97	<.0001
fall		1	-0.08900	0.02132	-4.17	<.0001

This leads to the following summary

Logarithmic Standard Quality Point	**Logarithmic Pure Price Index**	**Logarithmic Quality Correction Factor**	**Sum of these**

Non-linear Griliches-type hedonic model, GTHM

	SQP	**PPI**		**LQCF**		**Sum of these**	
Name	Notation	Notation	Value	Notation	Value	Value	Equations
any SQP	lacking	$\log P_0^1$	**-0,089**	$LQCF(\bar{x}^0 \to \bar{x}^1)$	0,079	0,010	(46)

For easy reference we state the very specific results for linear GTHM $f^t(x) = a_t + \sum b_k x_k$ in a separate theorem 8. Here the covariance term d_0^1 vanishes identically and even (46) becomes an identity.

Theorem 8. Main theorem for linear time invariant quality valuations or linear GTHM's.
Consider any linear HPF's having time invariant quality valuations or satisfying (40). These have been fitted to respective samples using any estimation method forcing the sum of residuals to zero. Now theorem 7 applies with $d_0^1 \equiv 0$ and (46) as an identity. We have

(48)
$$\log \frac{G(p^1)}{G(p^0)} \equiv \overline{f^1(x^1)} - \overline{f^0(x^0)} = \left[a_1 - a_0\right] + \left[\overline{f(x^1)} - \overline{f(x^0)}\right]$$

(49)
$$\begin{aligned}\overline{f(x^1)} - \overline{f(x^0)} &= f(\bar{x}^1) - f(\bar{x}^1) \\ &= LQVC^1(\bar{x}^0 \to \bar{x}^1)\end{aligned}$$

Therefore the log-change of geometric mean prices is decomposed in a unique way containing no time or quality point dependent choices into two factors, a quality corrected price index and a quality correction, as follows

$$\log\frac{G(p^1)}{G(p^0)} \equiv [a_1 - a_0] + [f(\bar{x}^1) - f(\bar{x}^1)] = \log P_0^1 + LQCF(\bar{x}^0 \to \bar{x}^1). \tag{50}$$

All equations (48) - (50) hold as identities.

The special conditions leading to these simple and beautiful results cannot be assumed a priori, because they are usually refused by tests based on empirical observations. But theorem 8 gives an ideally simple and beautiful case, against which more general (and more realistic) models should be contrasted. These cases are considered in our preceding theorems.

Empirical example 4: estimation results , linear Griliches-type model

Root MSE	0.12808	R-Square	0.6331
Dependent Mean	7.16158	Adj R-Sq	0.6261
Coeff Var	1.78844		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.05999	0.39282	2.70	0.0077
lnmem	log of memory	1	0.30056	0.04538	6.62	<.0001
lnspeed	log of speed	1	0.76856	0.06792	11.32	<.0001
fall		1	-0.08310	0.02183	-3.81	0.0002

This leads to the following summary

Logarithmic Standard Quality Point	**Logarithmic Pure Price Index**	**Logarithmic Quality Correction Factor**	**Sum of these**

	SQP	**PPI**		**LQCF**		**Sum of these**	
Name	Notation	Notation	Value	Notation	Value	Value	Equations
Linear Griliches-type hedonic model, GTHM							
any SQP	lacking	$\log P_0^1$	**-0,083**	$LQCF(\bar{x}^0 \to \bar{x}^1)$	**0,093**	**0,010**	(50)

3.7 Hedonic imputation (or interpolation) and its connection to standard quality point methods

In the so called Hedonic Imputation or Interpolation (HI for short) we may estimate (”impute”) for all old log-price-quality points $(\log p_i^0, x_i^0)$ new log-prices $\log p^1(x_i^0)$ corresponding to exactly the same old characteristics using the new version of our HPF, namely

$$\log p^1(x_i^0) = f^1(x_i^0) \qquad \text{for all i = 1, …, } n^0 . \tag{51}$$

Let us consider first the simpler case of linear HPF´s, which are time dependent. Using (51) we are ready to calculate the pure (or quality controlled) log-price changes as follows

$$\log p^1(x_i^0) - \log p_i^0 = f^1(x_i^0) - \log p_i^0 , \tag{52}$$

which in geometric terms is a comparison between observed price points and their imputed prices lying on the HPF below (or above) it. This is a vertical movement in the figure underlined by the term *imputation*. *Interpolation* refers to a computa-

tionally different but equivalent horizontal movement, when we correct for quality changes in the case of matched pairs of prices. Averaging over all old observations gives

(53) $$\frac{1}{n^0}\sum(\log p^1(x_i^0)-\log p_i^0) = \overline{f^1(x^0)}-\log G(p^0)$$

$$= \overline{f^1(x^0)}-\overline{f^0(x^0)} \qquad \text{by (15)}$$

$$= f^1(\bar{x}^0)-f^0(\bar{x}^0) \qquad \text{by linearity}$$

$$= \log P_0^1(\bar{x}^0) \qquad \text{by (17).}$$

Note that the asymmetry of the first two expressions vanishes if we start from estimated old prices $\log p^0(x_i^0)$ instead of the observed ones. The rest of the equation holds also for these because the sum of the residuals is zero. Also this procedure is included in Hedonic Imputation. We have derived a very important and intuitively significant connection between HI and SQP-method stated as

Theorem 9: Consider any time dependent linear HPF´s. If old log-prices or their estimates are imputed using the new HPF and the resulting quality controlled log-price changes are averaged over all observations, the HPI calculated at the old quality point $\log P_0^1(\bar{x}^0)$ arises as the result. In this sense hedonic imputation of old prices and SQP-method at the old SQP are equivalent.

We derive a similar result if new log-price-quality points $(\log p_j^1, x_j^1)$ are imputed "backwards" and the resulting pure log-price changes (note the change in the order in the difference)

(54) $$\log p_j^1 - \log p^0(x_j^1) = \log p_j^1 - f^0(x_j^0) \qquad \text{for all j = 1, …, } n^1$$

are averaged:

(55) $$\frac{1}{n^1}\sum(\log p_j^1-\log p^0(x_j^1)) = \log G(p^1)-\overline{f^0(x^1)}$$

$$= \overline{f^1(x^1)}-\overline{f^0(x^1)}$$

$$= f^1(\bar{x}^1)-f^0(\bar{x}^1) \qquad \text{by linearity}$$

$$= \log P_0^1(\bar{x}^1) \qquad \text{by (18).}$$

As above, more symmetric expressions arise if instead of actual new log-prices their estimates are taken as a starting point in (55). We have proven

Theorem 10: Consider any time dependent linear HPF´s. If new log-prices or their estimates are imputed using the old HPF and the resulting quality controlled log-price changes are averaged over all observations, the HPI calculated at the new quality point $\log P_0^1(\bar{x}^1)$ arises as the result. In this sense hedonic imputation of new prices and SQP-method at the new SQP are equivalent.

But we are ready to attack more specific problems of HM, because the conceptual and mathematical setup has been clarified without any assumptions, how we have arrived at our time-specific HPF's or QVF's $f^t(x) = f^t(x_1,...,x_K)$.
If we pool or combine imputations "forward" (53) and "backwards" (54), i.e. use both old and new observations in hedonic imputation, we in fact calculate the weighted average of the resulting pure price indices with number of observations n^0 and n^1 as weights. A straightforward calculation shows that this weighted average of pure price indices coincides with a pure price index calculated at a similarly weighted quality point, referred as **overall mean or SQP**

(56) $$\bar{x} = w^0\bar{x}^0 + w^1\bar{x}^1, \text{ where } w^t = n^t/(n^0+n^1)$$

We have

(57) $$w^0 \log P_0^1(\bar{x}^0) + w^1 \log P_0^1(\bar{x}^1) = \log P_0^1(\bar{x})$$

because $\log P_0^1(x)$ is a linear function as a difference of two linear functions of the same arguments.
This result was shown already in Vartia and Kurjenoja (1992) in the context of wage discrimination between men and women. We state the result shown above as

Theorem 11: Imputation in both directions,
or a linear time dependent HPI if both old and new observations (or estimated values of log-prices) are imputed and the pure log-price changes are averaged over all observations, the log of the HPI (57) $\log P_0^1(\bar{x})$ calculated at the overall mean (56) of the quality variables arises as the result. In this sense hedonic imputation of both old and new prices and the SQP-method at the overall mean are equivalent.

Theorems 9, 10 and 11 may be **generalised for non-linear HPF´s**. Proofs follow from the Theorem 6 and are omitted here. Therefore we have

Theorem 12: Consider any non-linear time dependent HPF´s, which are estimated in such a way that residuals sum to zero. **Then Theorems 9, 10 and 11 hold approximately**. The approximation error is estimated by d_0^1 given in (33); this d_0^1 is negligible in most applications.

For instance, Theorem 9 generalises to

(58) $$\begin{aligned} \log \frac{G(p^1)}{G(p^0)} &= \overline{f^1(x^1)} - \overline{f^0(x^0)} \\ &= [\overline{f^1(x^1)} - \overline{f^0(x^1)}] + [\overline{f^0(x^1)} - \overline{f^0(x^0)}] \\ &= \overline{\log P_0^1(x^0)} + \overline{LQCF^1(x^0 \to x^1)} \end{aligned}$$

This is clearly an identity, where the (HPI, ECQ)-pair is calculated by aggregating **micro-level** PPC´s and ECQ´s according to forward imputation instead of SQP-method. By theorem 12 HPI´s and ECQ´s calculated by imputation and SQP-methods satisfy

(59) $$\overline{\log P_0^1(x^0)} \approx \log P_0^1(\bar{x}^0)$$

(60) $$\overline{LQCF^1(x^0 \to x^1)} \approx LQCF^1(\bar{x}^0 \to \bar{x}^1)$$

Note that in (59) and (60) means are calculated for the fitted values in the left hand side, which is typical in **hedonic imputation**, while they appear as mean values of the arguments in (HPI, ECQ)-pairs, when SQP-methods or interpretations are used. In the case of **linear** time dependent HPF´s and QVF´s these two interpretations coincide, In the linear case it does not matter whether we start from the (HPI, ECQ)-pair $(\log P_0^1(\bar{x}^0), LQCF^1(\bar{x}^0 \to \bar{x}^1))$ of the SQP-type or of the pair $(\overline{\log P_0^1(x^0)}, \overline{LQCF^1(x^0 \to x^1)})$ of the hedonic imputation type. They give exactly the same results for linear HM´s by Theorem 9 and only the practical computations are different.

But in the case of non-linear time dependent HM`s they give different results and provide therefore different generalisations. The first of them generalises the SQP-method for non-linear models, while the latter leads to hedonic imputation. They allocate the $d_0^1 = \Delta^1 - \Delta^0$ terms (plus other possible higher order terms caused by non-linearity) in different ways into (HPI, ECQ)-pairs. Further research is needed to evaluate these problems.

Results of our empirical examples are collected in the following summary table, where they are easily compared.

Summary Table of empirical examples

Logarithmic Standard Quality Point		**Logarithmic Pure Price Index**		**Logarithmic Quality Correction Factor**		**Sum of these**	
	SQP	**PPI**		**LQCF**		**Sum of these**	
Name	Notation	Notation	Value	Notation	Value	Value	Equations
Linear time dependent HPF´s or QVF´s							
Old SQP	$\bar{x}^0$	$\log P_0^1\left(\bar{x}^0\right)$	-0,077	$LQCF^1\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0,087	0,010	(17) & (53)
New SQP	$\bar{x}^1$	$\log P_0^1\left(\bar{x}^1\right)$	-0,094	$LQCF^0\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0,105	0,010	(18) & (55)
Weighted average of decompositions (17) and (18):							
Overall SQP	$\bar{x}$	$\log P_0^1\left(\bar{x}\right)$	**-0,086**	$LQCF^{1/2}\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	**0,096**	**0,010**	(61)
Non-linear time dependent HPF´s or QVF´s							
Old SQP	$\bar{x}^0$	$\log P_0^1\left(\bar{x}^0\right)$	-0,069	$LQCF^1\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0,073	0,004	(35)
New SQP	$\bar{x}^1$	$\log P_0^1\left(\bar{x}^1\right)$	-0,115	$LQCF^0\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0,118	0,003	(36)
Weighted average of decompositions (35) and (36):							
Overall SQP	$\bar{x}$	$\log P_0^1\left(\bar{x}\right)$	**-0,092**	$LQCF^{1/2}\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	**0,096**	**0,004**	(61)
Non-linear Griliches-type hedonic model, GTHM							
any SQP	lacking	$\log P_0^1$	**-0,089**	$LQCF\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	0.079	0.010	(46)
Linear Griliches-type hedonic model, GTHM							
any SQP	lacking	$\log P_0^1$	**-0,083**	$LQCF\left(\bar{x}^0 \rightarrow \bar{x}^1\right)$	**0,093**	**0,010**	(50)

4 Conclusions

We have referred to **the dual nature of hedonic models** by giving two different interpretations of it, namely its **HPF-interpretation** (as comparing the time specific Hedonic Price Functions *for a given the quality point*. In this interpretation we consider pure prices changes when the quality point is fixed). The other is its **QVF-interpretation** (as a Quality Valuation Function *for a given time period*. In this interpretation we are interested, how changes in qualities affect the price when time period is fixed). We have concentrated in the paper mostly on HPF-interpretation of the hedonic models but the numerical examples illustrate both these view.

We have derived generalisations of the popular (but restricted) OAXACA-type decompositions of the log-change in geometric mean prices into **the log-change of pure prices** (or the log of the pure price index, $\log P_0^1$) and **logarithmic quality correction factor** (or the effect of changing qualities on log-prices, LQCF). The most symmetric form of these generalised OAXACA-decompositions is an identity for linear time dependent hedonic models (and an approximation for non-linear ones)

$$\log\frac{G(p^1)}{G(p^0)} = \log P_0^1(\bar{x}) + LQCF^{1/2}(\bar{x}^0 \to \bar{x}^1) \tag{61}$$

where the LQCF is a weighed mean of LQCF´s based on old and new QVF´s given in (17) and (18):

$$LQCF^{1/2}(\bar{x}^0 \to \bar{x}^1) = w^1 LQCF^1(\bar{x}^0 \to \bar{x}^1) + w^0 LQCF^0(\bar{x}^0 \to \bar{x}^1)\,. \tag{62}$$

In the case of linear time-dependent QVF´s (62) reduces to a single ”average QVF”, whose coefficients are weighted averages of time specific QVF´s, which explains our upper index ½ in these expressions. (Proofs of these statements based on Theorem 11 are omitted here for brevity.) Therefore, we may duplicate our results by **concentrating** on QVF´s and LQCF´s instead of HPF´s and $\log P_0^1$´s in eliminating effects of quality changes on the left side expression of (58).

All these important results are either identities (or their approximations), i.e. functional equations, see Eichhorn (1978) arising from the basic functional setup, where log-prices are described using a HPF or QVF of type $f^t(x)$, where x is an arbitrary vector of quality variables. Very little has been said of the choice of these quality variables, the number of them or of the other aspects of hedonic model building such as the specification of the functional form of $f^t(x) = f^t(x_1, x_2, ..., x_k)$, or its estimation and testing. Anything reasonable can be done concerning these aspects of the hedonic modelling, and nothing more has been assumed than that the residuals sum to unity. This an important point in understanding our results. We have been able to **separate** general mathematical aspects of the hedonic modelling and making quality corrections from more specific problems related to model building and estimation of these models. We regard this as our major contribution in the field.

After this we are ready to handle these more specific and technical problems which vary from one situation to another and are essentially problem specific and data dependent.

The methodological comments explain also why we have given very few references to a wide literature on hedonic modelling. In HM either too many of its problems are considered at the same time or some specific aspects of hedonic modelling (perhaps strongly connected to a special situation, application or to some "assumed" functional form such as linear GTHM, etc) has been attacked. Therefore, most of the literature is either **too general** (or confused) to lead to any useful results or **too specific** to be very interesting.

In some instances, rather modest progress has emerged despite of the considerable resources invested in the projects on quality adjustment. In our view and interpretation, **general and specific problems may have been mixed up** in these efforts in such a way that progress has slowed down and less than expected common understanding has emerged. The heterogeneity of the problems and the problem solvers (accompanied by varying commentators and practitioners on the field of price indices) defines a too complex social environment to produce a general agreement on what should be done.

To be more specific, there seems to prevail a subconscious commitment to rather unnecessary but widely utilised conventions (see Leamer (1983)) of sticking to Laspeyres type price indices and to "matched models" approach. This has effectively hindered the analysis of wider problems and more general techniques, e.g. in the realm of hedonic modelling.

It seems unnecessary to stick to linear or non-linear versions of Griliches type of hedonic models GTHM and to avoid time dependent linear or non-linear HPF´s, where either SQP-method or imputation has to be applied. Evidently, trying to take into account time effects in the valuation of quality variables has been (mistakenly) regarded as too difficult. Our results show, that this is no problem in the general setup. Common wisdom has also stressed **estimation problems of HPF's**, but

on a rather practical level without properly referring to statistical or econometric literature. Practitioners seem to be rather reluctant to use modern knowledge in the econometrics or sampling theory and are too tied to their traditional solutions, the very solutions that cause their problems.

Many possible extensions of hedonic modelling are omitted here. Several topics such as

- possibilities to approximate non-linear HPF´s or QVF´s by linear ones, or
- time dependent functions using time invariant ones or
- other practical simplifications of actual HPF´s or QVF´s
- for instance by choosing the quality variables carefully or
- otherwise lowering the dimension of the x-vector

are interesting areas of further research.

We hope that we have succeeded in clarifying the foundations of HPF´s and QVF´s. We have been able to separate those aspects of the quality correction problem that can be considered **in isolation from other more practical difficulties** (say from estimation problems) and also from the understandable reluctance of practitioners to implement new methods the foundations and versatility of which is not yet generally understood.

REFERENCES

Apostol, Tom M (1963): Mathematical Analysis. A Modern Approach to Advanced Calculus, Addison-Wesley Publishing Company, Reading, Massachusetts - Palo Alto - London

Björk, Tomas (1998): Arbitrage Theory in Continuous Time, Oxford University Press

Chung, Kai Lai (1968): A Course in Probability Theory, Harcourt , Brace & World, Inc., New York - Chicago - San Francisco - Atlanta

Comrie, L. J. (1966): Chambers´s Shorter Six-Figure Mathematical Tables, Chambers Ltd

Eichhorn, Wolfgang (1978): Functional Equations in Economics, Addison-Wesley

Goldberger, Arthur S (1964): Econometric Theory, New York, Wiley

Hyrkkö, Jarmo – Kinnunen, Arja – Vartia, Yrjö: Quality corrections in measuring price changes: Implementation of Hedonic Methods in Statistics Finland, Discussion Paper no 450: 1998, Department of Economics, University of Helsinki

Kahaner, D. – Moler, C. – Nash, S. (1989): Numerical Methods and Software, Prentice Hall

Leamer, Edward E (1983): Let´s Take the Con out of Econometrics, The American Economic Review, 73/1, 31 – 43

OECD (2000): Draft handbook on quality adjustment of price indexes for Information and Communication Technology Products. Prepare for Industry Committee, OECD Directorate for Science, Technology and Industry, May 2000. Prepared by Jack Triplett.

Rao, C. Radhakrishna (1968): Linear Statistical Inference and Its Applications, John Wiley & Sons, Inc., New York - London - Sydney

Spanos, Aris (1986): Statistical Foundations of Econometric Modelling, Cambridge University Press, Cambridge - London - New York - New Rochelle - Melbourne - Sydney

Turvey, Ralph: Incorporating New Models into A CPI: PC:s as an example. Proceedings of the Measurement of Inflation Conference, Cardiff, August 31. – September, 1. 1999.

Törnqvist, L - Vartia, P - Vartia.Y (1985): How Should Relative Changes Be Measured, The American Statistician, February 1985, Vol. 39, No. 1

Vartia, Yrjö - Kurjenoja, Jaana (1992): Wage Discrimination (in Finnish: Palkkadiskriminaatio. Naisten ja miesten palkkaero samasta työstä metalli- ja metsäteollisuuden suuryrityksissä v. 1990, Research Reports 60:1992, Department of Economics, University of Helsinki, Finland

Vartia, Yrjö (1976): Relative Changes and Index Numbers, The Research Institute of the Finnish Economy (ETLA), Series A4

Wittgenstein, Ludwig (1967): Remarks on the Foundations of Mathematics, edited by Von Wright, Rhees and Anscombe, Basil Blackwell, Oxford