5

Hedonic Regression Methods

Hedonic Modeling and Estimation

5.1 The hedonic regression method recognizes that heterogeneous goods can be described by their attributes or characteristics. That is, a good is essentially a bundle of (performance) characteristics.[1] In the housing context, this bundle may contain attributes of both the structure and the location of the properties. There is no market for characteristics, since they cannot be sold separately, so the prices of the characteristics are not independently observed. The demand and supply for the properties implicitly determine the characteristics' marginal contributions to the prices of the properties. Regression techniques can be used to estimate those marginal contributions or shadow prices. One purpose of the hedonic method might be to obtain estimates of the willingness to pay for, or marginal cost of producing, the different characteristics. Here we focus on the second main purpose, the construction of quality-adjusted price indices.

Hedonic Modeling

5.2 The starting point is the assumption that the price p_n^t of property n in period t is a function of a fixed number, say K, characteristics measured by "quantities" z_{nk}^t. With T+1 time periods, going from the base period 0 to period T, we have

$$p_n^t = f(z_{n1}^t, \ldots, z_{nK}^t, \varepsilon_n^t) \qquad (5.1)$$

$$t = 0, \ldots, T$$

where ε_n^t is a random error term (white noise). In order to be able to estimate the marginal contributions of the characteristics using standard regression techniques, equation (5.1) has to be specified as a parametric model. The two best-known hedonic specifications are the fully linear model

$$p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t \qquad (5.2)$$

and the logarithmic-linear model

$$\ln p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t \qquad (5.3)$$

where β_0^t and β_k^t are the intercept term and the characteristics parameters to be estimated. In both specifications the characteristics may be transformations, like logarithms, of continuous variables. In practice, many explanatory variables will be categorical rather than continuous and represented by a set of dummy variables which take the value of 1 if a property belongs to the category in question and the value of 0 otherwise.

5.3 For products such as high-tech goods, the log-linear model (5.3) is usually preferred, among other things because it most likely reduces the problem of heteroskedasticity (non-constant variance of the errors) as prices tend to be log-normally distributed (Diewert, 2003b). In the housing context, on the other hand, the linear model has much to recommend. In Chapter 3, the size of the structure and the size of the land it is built on were mentioned as two important price determining variables. Since the value of a property is generally equal to the *sum* of the price of the structure and the price of land, it can be argued that land and structures should be included in the model in a linear fashion, provided that the data are available. Chapter 8 will discuss this issue in more detail, including a decomposition of the hedonic price index into land and structures components. Unfortunately, not all data sources will contain information on lot and structure size. Lot size in particular may be lacking. When lot (or structure) size is not included as an explanatory variable, many empirical studies have found log-linear models to perform reasonably well.

5.4 The characteristics parameters β_k^t in (5.2) and (5.3) are allowed to change over time. This is in line with the idea that housing market conditions determine the marginal contributions of the characteristics: when demand and supply conditions change, there is no a priori reason to expect that those contributions are constant (Pakes, 2003). Yet, it seems most likely that market conditions change gradually. Therefore, the simplifying assumption can confidently be made, perhaps only for the short term, that the characteristics parameters (but not the intercept term) are constant over time. In the log-linear case this would give rise to the following constrained version of (5.3):

$$\ln p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t \qquad (5.4)$$

As will be seen below, the time dependent intercept terms (the β_0^t) can be converted into a constant quality price index.

5.5 Suppose we have data on selling prices and characteristics for the samples $S(0), S(1), \ldots, S(T)$ of properties sold in periods $t = 0, \ldots, T$ with sizes $N(0), N(1), \ldots, N(T)$. Under the classic error assumptions, in particular a zero mean and constant variance, the parameters of the hedonic models (5.2) and (5.3) can be estimated by Ordinary Least Squares (OLS) regression on the sample data of each time period separately. The constrained version (5.4) can be estimated on the pooled data pertaining to all time periods, provided that dummy variables are included which indicate the time periods (leaving out one dummy to prevent perfect collinearity). The estimating equation for the constrained log-linear model (5.4), which is generally referred to as the *time dummy variable hedonic model*, thus becomes

$$\ln p_n^t = \beta_0 + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t \qquad (5.5)$$

[1] The hedonic regression approach dates back at least to Court (1939) and Griliches (1961). Lancaster (1966) and Rosen (1974) laid down the conceptual foundations of the approach. Colwell and Dilmore (1999) argue that the first published hedonic study was a 1922 University of Minnesota master's thesis on agricultural land values.

where the time dummy variable D_n^τ has the value 1 if the observation comes from period τ and 0 otherwise; a time dummy for the base period 0 is left out. Although unusual, it is also possible to specify a time dummy model with the untransformed price as the dependent variable. This specification will be considered in the empirical example given at the end of this chapter.

Some Practical Issues

5.6 An important issue is the choice of the set of explanatory variables included in the hedonic equation. If some relevant variables – characteristics that can be expected to affect the price of a property (listed in Chapter 3) – are excluded, then the estimated parameters of the included characteristics will suffer from omitted variables bias. The bias carries over to the predicted prices computed from the regression coefficients and to the hedonic indices. Each property can be viewed as a unique good, for a large part due to its location. But detailed information on location and neighbourhood can be hard to obtain (Case, Pollakowski, and Wachter, 1991). Other characteristics may be unavailable also and some could be difficult to measure directly. So it is fair to say that in practice some omitted variables bias will always be present when estimating a hedonic model for housing. [2] The sign and magnitude of the bias, and its impact on the price index, are difficult to predict. The magnitude depends among other things on the correlation between the omitted and included variables.

5.7 The importance of location has led researchers to make use of longitude and latitude data of individual properties in hedonic regressions. This is usually achieved by constructing a matrix of distances between all properties in the data set and then using appropriate (though rather specialized) econometric methods to allow for spatial dependence in the estimated equation. Explicitly accounting for spatial dependence can ameliorate the omitted locational variables problem. Spatial dependence can be captured either in the regressors or the error term. The first approach, i.e., including location as an explanatory variable using geospatial data, is the most straightforward one. This can be done parametrically or nonparametrically, for example by making use of splines, as demonstrated by Hill, Melser and Reid (2010). For an elaborate discussion and a review of the literature on spatial dependence, the use of geospatial data and also on nonparametric estimation, we refer the reader to Hill (2011). [3]

5.8 Multicollinearity is a well-known problem in hedonic regressions. A high correlation between some of the included variables increases the standard errors of the regression coefficients; the coefficients become unstable. Again, it is difficult to say a priori how this will affect hedonic indices. For some purposes, multicollinearity may not be too problematic. For example, if we are not so much interested in the values of the parameters but merely in the predicted prices to be used in the estimation of the overall quality-adjusted house price index, then the problem of multicollinearity should not be exaggerated. In this case it is better to include a relevant variable, even if this would cause multicollinearity, than leaving it out as the latter gives rise to omitted variables bias. But when the parameter values are of interest as such, for example when we are trying to decompose the property prices into land and structures components, then multicollinearity does pose problems. In Chapter 8 it will be shown that this is indeed a problem.

5.9 As with other methods, some data cleaning might be necessary. Obvious entry errors should be deleted. Yet a cautious approach is called for. Deleting outliers from a regression with the aim of producing more stable coefficients (hence, more stable price indices) is often arbitrary and could lead to biased estimates. The use of hedonics requires data on all characteristics included in the model. Unfortunately, partial non-response is present in many data sets. That is, the information on one or more characteristics may be missing for a part of the sample. Procedures have been developed to impute the missing data, but again it is important to avoid arbitrary choices that can have an impact on the results.

5.10 In the next two sections, the two main hedonic approaches, the time dummy approach and the imputations approach, to constructing quality-adjusted house price indices will be discussed. Without denying potential econometric problems, our focus will be on the use of least squares regression to estimate the models.

Time Dummy Variable Method

5.11 The time dummy variable approach to constructing a hedonic house price index has been used frequently in academic studies but not so much by statistical agencies. [4] One advantage of this approach is its simplicity; the price index follows immediately from the estimated

[2] A related point is that the characteristics of each house in the sample should be available in real time. House characteristics can change over time (which is actually the reason why they are given a superscript for time t in the hedonic models above). Keeping the characteristics fixed implies that the hedonic price index would not be adjusted for such quality changes.

[3] Colwell (1998) proposed a nonparametric spatial interpolation method which seems well adapted to model land prices as a function of the property's geographical two-dimensional coordinates.

[4] This method was originally developed by Court (1939; 109-111) as his hedonic suggestion number two. The terminology adopted by us is not uniformly employed in the real estate literature. For example, Crone and Voith (1992) refer to the time dummy method as the "constrained hedonic" method, Gatzlaff and Ling (1994) call it the "explicit time-variable" method, while Knight, Dombrow and Sirmans (1995) name it the "varying parameter" method. Other terms also appear in the literature so that statements about the relative merits of different hedonic methods require careful interpretation.

pooled time dummy regression equation (5.5). Running one overall regression on the pooled data of the samples $S(0), S(1), \ldots, S(T)$ relating to periods $t = 0, \ldots, T$ (with sizes $N(0), N(1), \ldots, N(T)$) yields coefficients $\hat{\beta}^0$, $\hat{\delta}^t$ $(t = 1, \ldots, T)$ and $\hat{\beta}_k$ $(k = 1, \ldots, K)$. The time dummy parameter shifts the hedonic surface upwards or downwards and measures the effect of "time" on the logarithm of price. Exponentiating the time dummy coefficients thus controls for changes in the quantities of the characteristics and provides a measure of quality-adjusted house price change between the base period 0 and each comparison period *t*. In other words, the time dummy index going from period 0 to period *t* is given by [5]

$$P_{TD}^{0t} = \exp(\hat{\delta}^t) \qquad (5.6)$$

5.12 Pooling cross-section data preserves degrees of freedom. The regression coefficients $\hat{\beta}_k$ will therefore generally have lower standard errors than the coefficients $\hat{\beta}_k^t$ that would be obtained by estimating model (5.19) separately on the data of the samples $S(0), S(1), \ldots, S(T)$. Although the increased efficiency can be seen as an advantage, it comes at an expense: the assumption of fixed characteristics parameters is a disadvantage of the time dummy hedonic method.

5.13 When using OLS, the time dummy hedonic index can be written as (see e.g. Diewert, Heravi and Silver, 2009; de Haan, 2010a)

$$P_{TD}^{0t} = \frac{\prod_{n \in S(t)} (p_n^t)^{1/N(t)}}{\prod_{n \in S(0)} (p_n^0)^{1/N(0)}} \exp\left[\sum_{k=1}^{K} \hat{\beta}_k (\bar{z}_k^0 - \bar{z}_k^t)\right] \qquad (5.7)$$

where $\bar{z}_k^s = \sum_{n \in S(s)} z_{nk}^s / N(s)$ is the sample mean of characteristic k in period s $(s = 0, t)$. Equation (5.7) tells us that the time dummy index is essentially the product of two factors. The first factor is the ratio of the geometric mean prices in the periods *t* and 0. The second factor, $\exp[\sum_{k=1}^{K} \hat{\beta}_k (\bar{z}_k^0 - \bar{z}_k^t)]$, adjusts this ratio of raw sample means for differences in the average characteristics $\bar{z}_k^0$ and $\bar{z}_k^t$; it serves as a quality-adjustment factor which accounts for both changes in the quality mix and quality changes of the individual properties (provided that all relevant quality-determining attributes are included in the hedonic model). Notice that the time dummy price index simplifies to the ratio of geometric mean prices if $\bar{z}_k^t = \bar{z}_k^0$, i.e. if the average characteristics in period *t* and period 0 happen to be equal.

5.14 Suppose for simplicity that the housing stock is constant, in the sense that there are no houses entering or exiting, and that the quality of the individual properties does not change. Suppose further that $S(0)$ and $S(t)$ are random or "representative" selections from the housing stock. In that case the time dummy method implicitly aims at a ratio of geometric mean prices for the total stock, which is equal to the geometric mean of the individual price ratios. [6] Although it is true that the target of measurement may be different for different purposes, it is difficult to see what purposes a geometric stock RPPI would meet. Arithmetic target RPPIs, such as an index that tracks the value of the fixed housing stock over time, seem to be more appropriate (see also Chapters 4 and 8).

5.15 The samples of houses traded, $S(0)$ and $S(t)$, may not be representative for the total housing stock (or for the total population of houses sold). A solution could be to weight the samples in order to make them representative. Running an OLS regression on the (pooled) weighted data set is equivalent to running a Weighted Least Squares (WLS) regression on the original data set. Under the assumption of a constant variance of the errors, econometric textbooks do not suggest the use of WLS since this will introduce heteroskedasticity. Note that a WLS time dummy method will still generate a geometric index, in this case a weighted one.

5.16 A better option than using WLS regressions could be to stratify the samples, run separate OLS regressions on the data of the different strata, and then explicitly weight the stratum-specific hedonic indices using stock (or sales) weights to construct an overall RPPI with an arithmetic structure at the upper level of aggregation. This stratified hedonic approach has several other advantages as well, as will be explained later.

5.17 A problem with the time dummy method is the revision that goes with it. If the time series is extended to $T+1$ and new sample data is added, the characteristics coefficients will change. Consequently, the newly computed price index numbers for the periods $t = 1, \ldots, T$ will differ from those previously computed. [7] When additional data become available, the efficiency due to the pooling of data increases and better estimates can be made. This can actually be seen as a strength rather than a weakness of the method. On the other hand, statistical agencies and their users will most likely be reluctant to accept continuous revisions of previously published figures.

5.18 The multiperiod time dummy method therefore appears to be of limited use for the production of official house price indices although there are ways to deal with the problem of revisions. One way would be to estimate time dummy indices for adjacent periods *t*-1 and *t* and then multiply them to obtain a time series which is free of revisions. This high-frequency chaining has the additional advantage of relaxing the assumption of fixed parameters.

[5] The expected value of the exponential of the time dummy coefficient is not exactly equal to the exponential of the time dummy parameter. The associated bias is often referred to as small sample bias: it diminishes when the sample size grows. Unless the sample size is extraordinary small, the bias will be small compared to the standard error and can usually be neglected in practice.

[6] In index number theory such an index is referred to as a Jevons index.

[7] In the words of Hill (2004), the time dummy approach violates time fixity.

It is, however, not entirely without problems. Drift in the index can occur when the data exhibit systematic fluctuations such as seasonal fluctuations.[8]

Characteristics Prices and Imputation Methods

5.19 In the second main approach to compiling a hedonic price index, separate regressions are run for all time periods and the index is constructed by making use of the predicted prices based on the regression coefficients. Because the implicit characteristics prices are allowed to vary over time, this method is more flexible than the time dummy variable method. Two variants can be distinguished: the *characteristics prices approach* and the *imputations approach*. It will be shown that, under certain circumstances, both approaches are equivalent. We will first discuss the characteristics prices approach.[9]

Characteristics Prices Approach

5.20 To illustrate this approach, suppose as before that sample data are available on prices and relevant characteristics of houses sold in the base period 0 and each comparison period *t*. We will first assume that the linear hedonic model (5.2) holds true and is estimated on the data of period 0 and period *t* separately. This yields regression coefficients $\hat{\beta}_0^s$ and $\hat{\beta}_k^s$ $(k=1,\ldots,K)$ for $s=0,t$. The predicted prices for each individual property are $\hat{p}_n^0 = \hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 z_{nk}^0$ and $\hat{p}_n^t = \hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t z_{nk}^t$. It is also possible to compute predicted period 0 and period *t* prices for a "standardized" property with fixed (quantities of) characteristics z_k^*. The resulting estimated price relative is

$$\frac{\hat{p}^t}{\hat{p}^0} = \frac{\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t z_k^*}{\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 z_k^*} \tag{5.8}$$

Expression (5.8) is a quality-adjusted price index because the characteristics are kept fixed. But different values of z_k^* will give rise to different index numbers. So what would be the preferred choice?

5.21 Suppose that we were aiming at a sales-based RPPI. There are two natural choices for z_k^* in (5.8): the sample average characteristics of the base period, $\bar{z}_k^0$, and the sample averages of the comparison period t $(t=1,\ldots,T)$, $\bar{z}_k^t$. The usual solution in index number theory is to treat the resulting price indices – which are equally valid – in a symmetric manner by taking the geometric mean. Setting $z_k^* = \bar{z}_k^0$ in (5.8) generates a Laspeyres-type characteristics prices (CP) index:

$$P_{CPL}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t \bar{z}_k^0}{\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 \bar{z}_k^0} \tag{5.9}$$

Setting $z_k^* = \bar{z}_k^t$ in (5.8) yields a Paasche-type index:

$$P_{CPP}^{0t} = \frac{\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t \bar{z}_k^t}{\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 \bar{z}_k^t} \tag{5.10}$$

By taking the geometric mean of (5.9) and (5.10) the Fisher-type characteristics prices index is obtained:

$$P_{CPF}^{0t} = \left[P_{CPL}^{0t} P_{CPP}^{0t}\right]^{1/2} \tag{5.11}$$

5.22 The characteristics prices method can also applied in combination with the log-linear model given by (5.3). Running separate regressions of this model on the sample data for periods 0 and *t* yields predicted prices (after exponentiating) $\hat{p}_n^0 = \exp(\hat{\beta}_0^0)\exp[\sum_{k=1}^{K} \hat{\beta}_k^0 z_{nk}^0]$ and $\hat{p}_n^t = \exp(\hat{\beta}_0^t)\exp[\sum_{k=1}^{K} \hat{\beta}_k^t z_{nk}^t]$. Similar to what was done in (5.8) for the linear model, prices can be predicted for a standardized house. Using the sample averages of the characteristics in the base period to define the standardized house, the geometric counterpart to the Laspeyres-type characteristics prices index (5.9) is found:

$$P_{CPGL}^{0t} = \frac{\exp(\hat{\beta}_0^t)\exp\left[\sum_{k=1}^{K} \hat{\beta}_k^t \bar{z}_k^0\right]}{\exp(\hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} \hat{\beta}_k^0 \bar{z}_k^0\right]}$$

$$= \exp(\hat{\beta}_0^t - \hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} (\hat{\beta}_k^t - \hat{\beta}_k^0)\bar{z}_k^0\right] \tag{5.12}$$

The geometric counterpart to the Paasche-type hedonic index (5.10) is obtained by using the sample averages of the characteristics in the comparison period:

$$P_{CPGP}^{0t} = \frac{\exp(\hat{\beta}_0^t)\exp\left[\sum_{k=1}^{K} \hat{\beta}_k^t \bar{z}_k^t\right]}{\exp(\hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} \hat{\beta}_k^0 \bar{z}_k^t\right]}$$

$$= \exp(\hat{\beta}_0^t - \hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} (\hat{\beta}_k^t - \hat{\beta}_k^0)\bar{z}_k^t\right] \tag{5.13}$$

[8] An alternative approach would be the use of a moving window. For example, suppose we initially estimated a time dummy index on the data of twelve months. Next, we delete the data of the first month and add the data of the thirteenth month and estimate a time dummy index on this data set, and so on. By multiplying (chaining) the last month-to-month changes a non-revised time series is obtained. For an application, see Shimizu, Nishimura and Watanabe (2010). In the example for the town of "A", given at the end of this chapter, drift does not seem to be a major problem; the moving window method gives much the same results as the multiperiod time dummy regression.

[9] Again, the terminology differs between authors. For example, Crone and Voith (1992) and Knight, Dombrow and Sirmans (1995) refer to this approach as the "hedonic method" (as opposed to the "constrained hedonic" or "varying parameter" method, what we have called the time dummy variable approach), while Gatzlaff and Ling (1994) refer to it as the "strictly cross-sectional" method.

Taking the geometric mean of (5.12) and (5.13) yields

$$\begin{aligned} P^{0t}_{CPGF} &= \left[P^{0t}_{HGL} P^{0t}_{HGP}\right]^{1/2} \\ &= \exp(\hat{\beta}^{t}_{0} - \hat{\beta}^{0}_{0}) \exp\left[\sum_{k=1}^{K} (\hat{\beta}^{t}_{k} - \hat{\beta}^{0}_{k}) \bar{z}^{0t}_{k}\right] \qquad (5.14) \end{aligned}$$

where $\bar{z}^{0t}_{k} = (\bar{z}^{0}_{k} + \bar{z}^{t}_{k})/2$ in (5.14) denotes the mean of the average characteristics in the base and comparison period.

5.23 If the target index is a stock-based rather than a sales-based RPPI, the two natural choices for the characteristics z^{*}_{k} in equation (5.8) would be the average *stock* characteristics of the base period and those of the comparison period. The first choice would produce a Laspeyres-type stock RPPI, the second choice a Paasche-type stock RPPI. Both indices measure the quality-adjusted value change of the housing stock, but the results will usually differ. Not only does the average quality of the housing stock change over time, the Laspeyres-type index ignores new properties that entered the housing market whereas the Paasche-type index does not take into account disappearing properties.

5.24 Of course the assumption of known stock averages for all property characteristics included in the hedonic model is unrealistic. In most situations we have to rely on estimates, i.e. on the sample averages $\bar{z}^{0}_{k}$ and $\bar{z}^{t}_{k}$ which are based on the same characteristics data that is used to estimate the hedonic equations. This leads to formulae (5.9) and (5.10), or the geometric mean (5.11), which describe sales-based RPPIs. Once again we are reminded that sales RPPIs can be seen as estimators of stock RPPIs, provided that the samples are representative of the total stock. The latter is rather doubtful, however, and the usual approach is to stratify the samples and weight the estimated stratum indices using stock weights.

Hedonic Imputation Approach

5.25 The question arises how the characteristics prices method described above relates to the standard (matched-model) methodology to construct price indices. From an index number point of view we can look at the issue in the following way. The period *t* prices of properties sold in period 0 cannot be observed and are "missing" because those properties, or at least the greater part, will not be resold in period *t*. Similarly, the period 0 prices of the properties sold in period *t* are unobservable. To apply standard index number formulae these "missing prices" must be imputed.[10] Hedonic imputation indices do this by using predicted prices, evaluated at fixed characteristics, based on the hedonic regressions for all time periods.

Arithmetic Imputation Indices

5.26 The Laspeyres imputation index imputes period *t* prices for the properties belonging to the base period sample $S(0)$, evaluated at base period characteristics to control for quality changes. Using the linear model (5.1), the imputed prices are $\hat{p}^{t}_{n}(0) = \hat{\beta}^{t}_{0} + \sum_{k=1}^{K} \hat{\beta}^{t}_{k} z^{0}_{nk}$, and the hedonic imputation Laspeyres index becomes

$$\begin{aligned} P^{0t}_{HIL} &= \frac{\sum_{n \in S(0)} 1 \hat{p}^{t}_{n}(0)}{\sum_{n \in S(0)} 1 p^{0}_{n}} = \frac{\sum_{n \in S(0)} \left[\hat{\beta}^{t}_{0} + \sum_{k=1}^{K} \hat{\beta}^{t}_{k} z^{0}_{nk}\right]}{\sum_{n \in S(0)} p^{0}_{n}} \\ &= \frac{\hat{\beta}^{t}_{0} + \sum_{k=1}^{K} \hat{\beta}^{t}_{k} \bar{z}^{0}_{k}}{\sum_{n \in S(0)} p^{0}_{n} / N(0)} \qquad (5.15) \end{aligned}$$

Notice that the quantity associated with each price is 1; basically, every house is unique and cannot be matched except through the use of a model.

5.27 The hedonic imputation Laspeyres index (5.15) is an example of a *single imputation* index in which the observed prices are left unchanged. It can be argued that it would be better to use a *double imputation* approach, where the observed prices are replaced by the predicted values. This is because biases in the period 0 and period *t* estimates resulting from omitted variables are likely to offset each other, at least to some degree; see e.g. Hill, 2011. Using $\hat{p}^{0}_{n} = \hat{\beta}^{0}_{0} + \sum_{k=1}^{K} \hat{\beta}^{0}_{k} z^{0}_{nk}$, the hedonic double imputation (DI) Laspeyres price index is

$$\begin{aligned} P^{0t}_{HDIL} &= \frac{\sum_{n \in S(0)} 1 \hat{p}^{t}_{n}(0)}{\sum_{n \in S(0)} 1 \hat{p}^{0}_{n}} = \frac{\sum_{n \in S(0)} \left[\hat{\beta}^{t}_{0} + \sum_{k=1}^{K} \hat{\beta}^{t}_{k} z^{0}_{nk}\right]}{\sum_{n \in S(0)} \left[\hat{\beta}^{0}_{0} + \sum_{k=1}^{K} \hat{\beta}^{0}_{k} z^{0}_{nk}\right]} \\ &= \frac{\hat{\beta}^{t}_{0} + \sum_{k=1}^{K} \hat{\beta}^{t}_{k} \bar{z}^{0}_{k}}{\hat{\beta}^{0}_{0} + \sum_{k=1}^{K} \hat{\beta}^{0}_{k} \bar{z}^{0}_{k}} = P^{0t}_{CPL} \qquad (5.16) \end{aligned}$$

A comparison with equation (5.12) shows that, using the linear model, the double imputation index equals the Laspeyres-type characteristics prices index. This result does not depend on the estimation method. If we would use OLS regression to estimate the linear model, then the single imputation index would be equal to the double imputation index and also coincide with the characteristics prices index as in this case $\sum_{n \in S(0)} p^{0}_{n} = \sum_{n \in S(0)} \hat{p}^{0}_{n}$, due to the fact that the hedonic model includes an intercept term so that the OLS regression residuals sum to zero.

5.28 The hedonic single imputation Paasche index imputes base period prices for the properties belonging to the period *t* sample $S(t)$, evaluated at period *t* characteristics. Using again the linear model (5.1), these imputed prices

[10] As noted earlier, the hedonic theory dates back at least to Court (1939; 108). Imputation was his hedonic suggestion number one. His suggestion was followed up by Griliches (1971a; 59-60) (1971b; 6) and Triplett and McDonald (1977; 144). More recent contributions to the hedonic imputations literature include Diewert (2003b), de Haan (2004) (2009) (2010a), Triplett (2004) and Diewert, Heravi and Silver (2009). In a housing context the hedonic imputation method is discussed in detail by Hill and Melser (2008) and Hill (2011).

are given by $\hat{p}_n^0(t) = \hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 z_{nk}^t$. To save space we will only show the double imputation variant. Here, the observed (period *t*) prices are replaced by their model-based predictions $\hat{p}_n^t = \hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t z_{nk}^t$. Thus, the hedonic double imputation Paasche price index is

$$P_{HDIP}^{0t} = \frac{\sum_{n \in S(t)} 1 \hat{p}_n^t}{\sum_{n \in S(t)} 1 \hat{p}_n^0(t)} = \frac{\sum_{n \in S(t)} \left[\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t z_{nk}^t \right]}{\sum_{n \in S(t)} \left[\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 z_{nk}^t \right]}$$

$$= \frac{\hat{\beta}_0^t + \sum_{k=1}^{K} \hat{\beta}_k^t \bar{z}_k^t}{\hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\beta}_k^0 \bar{z}_k^t} = P_{CPP}^{0t} \qquad (5.17)$$

which coincides with the Paasche-type characteristics prices index. If OLS regression is used, then (5.17) is equal to the single imputation Paasche index because in this particular case the numerator equals $\sum_{n \in S(t)} p_n^t$. It will then be unnecessary to estimate the hedonic equations for the comparison periods $t = 1,\ldots,T$; estimating the base period hedonic equation to obtain the base period imputed values will suffice.

5.29 The hedonic double imputation Fisher index is found by taking the geometric mean of (5.16) and (5.17):

$$P_{HDIF}^{0t} = \left[P_{HDIL}^{0t} P_{HDIP}^{0t} \right]^{1/2} \qquad (5.18)$$

The above imputation indices can be given two interpretations. They can be viewed either as estimators of the quality-adjusted value change of the entire housing stock, i.e., as stock-based RPPIs, or as estimators of quality-adjusted sales-based RPPIs. Under the first interpretation, to produce approximately unbiased results, each sample should be a random or representative selection from the housing stock. Sample selection bias problems could be less severe under the second interpretation, although this depends on the sampling design.[11]

Geometric Imputation Indices

5.30 The imputation approach can also be applied to geometric price index number formulae. Let us start with what might be called the geometric counterpart to the imputation Laspeyres price index (5.15). For reasons of "consistency" the imputations will now be computed using the log-linear hedonic model (5.3) instead of the linear model. The imputed period *t* prices for the properties belonging to the base period sample $S(0)$, evaluated at base period characteristics, are $\hat{p}_n^t(0) = \exp(\hat{\beta}_0^t)\exp[\sum_{k=1}^{K} \hat{\beta}_k^t z_{nk}^0]$. Hence, the double imputation unweighted geometric index, in which the base period prices are replaced by predicted values $\hat{p}_n^0 = \exp(\hat{\beta}_0^0)\exp[\sum_{k=1}^{K} \hat{\beta}_k^0 z_{nk}^0]$, is

$$P_{HDIGL}^{0t} = \frac{\prod_{n \in S(0)} (\hat{p}_n^t(0))^{1/N(0)}}{\prod_{n \in S(0)} (\hat{p}_n^0)^{1/N(0)}}$$

$$= \exp(\hat{\beta}_0^t - \hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} (\hat{\beta}_k^t - \hat{\beta}_k^0)\bar{z}_k^0 \right] = P_{CPGL}^{0t} \qquad (5.19)$$

Similarly, the geometric counterpart to the imputation Paasche price index (5.16) is obtained by imputing period 0 prices for the properties belonging to the period *t* sample $S(t)$, which are given by $\hat{p}_n^0(t) = \exp(\hat{\beta}_0^0)\exp[\sum_{k=1}^{K} \hat{\beta}_k^0 z_{nk}^t]$, and replacing the observed period *t* prices by the predictions $\hat{p}_n^t = \exp(\hat{\beta}_0^t)\exp[\sum_{k=1}^{K} \hat{\beta}_k^t z_{nk}^t]$. So we have

$$P_{HDIGP}^{0t} = \frac{\prod_{n \in S(t)} (\hat{p}_n^t)^{1/N(t)}}{\prod_{n \in S(t)} (\hat{p}_n^0(t))^{1/N(t)}}$$

$$= \exp(\hat{\beta}_0^t - \hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} (\hat{\beta}_k^t - \hat{\beta}_k^0)\bar{z}_k^t \right] = P_{CPGP}^{0t} \qquad (5.20)$$

5.31 When OLS is used to estimate the log-linear regression equations, the denominator of (5.19) and the numerator of (5.20) will equal the geometric sample means of the prices in period 0 and period *t*, respectively, and the double imputation indices coincide with single imputation indices. Taking the geometric mean of (5.19) and (5.20) yields

$$P_{HDIGF}^{0t} = \left[P_{HDIGL}^{0t} P_{HDIGP}^{0t} \right]^{1/2}$$

$$= \exp(\hat{\beta}_0^t - \hat{\beta}_0^0)\exp\left[\sum_{k=1}^{K} (\hat{\beta}_k^t - \hat{\beta}_k^0)\bar{z}_k^{0t} \right] = P_{CPGF}^{0t} \qquad (5.21)$$

where $\bar{z}_k^{0t} = (\bar{z}_k^0 + \bar{z}_k^t)/2$ denotes the mean of the average characteristics in periods 0 and *t*, as before.

5.32 The symmetric imputation index equation (5.21) can be rewritten in a way that is surprisingly similar to equation (5.7) for the time dummy index when OLS is used to estimate the hedonic equations (see Diewert, Heravi and Silver, 2009, and de Haan, 2010a):

$$P_{HDIGF}^{0t} = \frac{\prod_{n \in S(t)} (p_n^t)^{1/N(t)}}{\prod_{n \in S(0)} (p_n^0)^{1/N(0)}} \exp\left[\sum_{k=1}^{K} \hat{\beta}_k^{0t} (\bar{z}_k^0 - \bar{z}_k^t) \right] \qquad (5.22)$$

where $\hat{\beta}_k^{0t} = (\hat{\beta}_k^0 + \hat{\beta}_k^t)/2$ denotes the average value of the *k*-th coefficient in periods 0 and *t*. Equation (5.22) adjusts the ratio of observed geometric mean prices for any differences in the average sample characteristics. Triplett (2006) refers to this as "hedonic quality adjustment". A comparison with equation (5.7) shows that if the sample averages of

[11] If all property transactions are observed, there is no sampling involved from a sales point of view, and sample selection bias is not an issue. In many countries the Land Registry records all transactions, at least for resold houses. However, such data sets usually have limited information on characteristics; see e.g. Lim and Pavlou (2007) or Academetrics (2009).

all characteristics stay the same $(\bar{z}_k^0 = \bar{z}_k^t)$, then the symmetric hedonic imputation index and the time dummy index coincide and equal the ratio of observed geometric mean prices, but this will obviously, rarely happen. Both types of hedonic indices also coincide if, for each characteristic, the average coefficient $\hat{\beta}_k^{0t}$ from the two separate regressions would be equal to the coefficient $\hat{\beta}_k$ from the time dummy regression. This is rare as well, but it suggests that both approaches generate similar results if the characteristics parameters are approximately constant over time.

5.33 If the characteristics parameters can be assumed constant over time, the average coefficients $\hat{\beta}_k^{0t}$ in equation (5.22) can be replaced by the base period coefficients $\hat{\beta}_k^0$. In that case there would be no need to run a regression in each time period, and we would in fact be using the non-symmetric imputation price index given by equation (5.13).[12] The base period regression could be run on a bigger data set to increase the stability of the coefficients. It is advisable to regularly check if the coefficients have significantly changed and to update them when necessary.

5.34 As mentioned earlier, geometric price indices are less suitable as estimators of quality-adjusted RPPIs. This is not to say that they should never be used. In conjunction with stratification, the use of (5.21) could produce satisfactory results since this would combine quality adjustment (using a log-linear hedonic regression model) and a symmetric index number formula within the different strata with mix adjustment across strata. The stratified hedonic approach will be discussed in the next section.

Stratified Hedonic Indices

5.35 Chapter 4 dealt with stratification or mix adjustment. Stratification is a simple and powerful tool to adjust for changes in the quality mix of the properties sold. However, some quality mix changes within the strata are likely to remain, as essentially every property is a unique good, and some unit value bias could therefore occur. A more detailed stratification scheme may be unfeasible, especially when the number of observations is relatively small. Provided that the necessary data on characteristics are available, it could be worthwhile to work with a less fine stratification scheme and use hedonic regression at the stratum level to adjust for quality mix changes. This two-stage approach combines hedonics at the lower (stratum) level and explicit weighting at the upper level to form an overall RPPI.

5.36 Two advantages of stratification have been mentioned earlier. First, stratification enables the statistical agency to publish different RPPIs for different market segments. Users will benefit from this because it is well known that different types of houses, different regions, etc. can exhibit quite different price trends. Second, stratification can be helpful for reducing sample selection bias, including bias due to non-response, in particular for a stock-based RPPI.

5.37 When using hedonic regression techniques to adjust for quality (mix) changes, stratification is highly recommended. It is very unlikely that a single hedonic model holds true for all market segments, hence separate regressions should be run for different types of properties, different locations, etc. There are in fact two issues involved. Perhaps the biggest issue is that different sets of property characteristics will be needed for different market segments. For example, the characteristics that are relevant for detached dwelling units differ from those that are relevant for high rise apartments, if only because the floor of the apartment seems an important price determining variable. The second, though probably less important, issue is that the parameter values for the same characteristics can differ across housing market segments. Statistical tests for differences in parameter values between sub-samples can be found in any econometrics textbook.

5.38 The stratified hedonic approach can be illustrated most easily with reference to the imputation method, especially in combination with the Laspeyres index formula. Recall the third expression on the right-hand side of the hedonic single imputation Laspeyres price index (5.15), where the period *t* prices for the houses in the base period sample $S(0)$ are "missing" and imputed (using the estimated hedonic regression model for period *t*) by $\hat{p}_n^t(0)$. Suppose, as in Chapter 4, that the total sample is (post) stratified into *M* sub-samples $S_m(0)$. Equation (5.15) can then be rewritten as

$$P_{HIL}^{0t} = \frac{\sum_{n\in S(0)} \hat{p}_n^t(0)}{\sum_{n\in S(0)} p_n^0} = \frac{\sum_{m=1}^{M}\sum_{n\in S_m(0)} \hat{p}_n^t(0)}{\sum_{m=1}^{M}\sum_{n\in S_m(0)} p_n^0}$$

$$= \frac{\sum_{m=1}^{M}\sum_{n\in S_m(0)} p_n^0 \left[\sum_{n\in S_m(0)} \hat{p}_n^t(0) \Big/ \sum_{n\in S_m(0)} p_n^0 \right]}{\sum_{m=1}^{M}\sum_{n\in S_m(0)} p_n^0} = \sum_{m=1}^{M} s_m^0 P_{HIL,m}^{0t} \qquad (5.23)$$

where $P_{HIL,m}^{0t} = \sum_{n\in S_m(0)} \hat{p}_n^t(0) / \sum_{n\in S_m(0)} p_n^0$ denotes the hedonic (single) imputation Laspeyres price index between the base period and period *t* for cell *m*; $s_m^0 = \sum_{n\in S_m(0)} p_n^0 / \sum_{n\in S(0)} p_n^0$ is the corresponding sales value share, which serves as the weight for $P_{HIL,m}^{0t}$. Note that the last expression of (5.23) has a similar structure as the mix-adjusted index given by equation (4.1), but in the present case the cell indices are hedonic imputation indices rather than unit value indices.

[12] In Europe this type of hedonic quality adjustment is called "hedonic re-pricing", especially in case the sample size is fixed (Destatis, 2009).

5.39 Equation (5.23) shows that if the imputed prices $\hat{p}_n^t(0)$ for all houses in the sample $S(0)$ are based on one overall hedonic regression, then the aggregate hedonic imputation Laspeyres index can be written in the form of a stratified index. But this is just another way of writing things, not what is meant by a stratified hedonic approach. Also, as argued above, the use of a common model is very unrealistic. So instead of running one big hedonic regression, separate regressions should be performed on the data of the sub-samples in each time period to obtain imputed (period *t*) prices and imputation cell indices. That would lead to a stratified Laspeyres-type hedonic imputation index.

5.40 It would be preferable to estimate a stratified Fisher hedonic index rather than a Laspeyres one. This is perfectly feasible for a sales RPPI but may not be feasible for a stock RPPI, as was already mentioned in Chapter 3, since up-to-date census data on the number of properties is often lacking.

Main Advantages and Disadvantages

5.41 This section summarizes the advantages and disadvantages of hedonic regression methods to construct an RPPI. The main advantages are:

- If the list of available property characteristics is sufficiently detailed, hedonic methods can in principle adjust for both sample mix changes and quality changes of the individual properties.
- Price indices can be constructed for different types of dwellings and locations through a proper stratification of the sample. Stratification has a number of other advantages as well.
- The hedonic method is probably the most efficient method for making use of the available data.
- The imputation variant of the hedonic regression method is analogous to the matched model methodology that is widely used in order to construct price indices.

5.42 The main disadvantages of hedonic regression are:

- It may be difficult to control sufficiently for location if property prices and price trends differ across detailed regions. However, a stratified approach to hedonic regressions will help overcome this problem to some extent.
- The method is data intensive since it requires data on all relevant property characteristics, so it is relatively expensive to implement.[13]
- While the method is essentially reproducible, different choices can be made regarding the set of characteristics included in the model, the functional form, possible transformations of the dependent variable[14], the stochastic specification, etc., which could lead to varying estimates of overall price change. Thus, a lot of metadata may be required.
- The general idea of the hedonic method is easily understood but some of the technicalities may not be easy to explain to users.

5.43 The overall evaluation of the hedonic regression method is that it is probably the best method that could be used in order to construct constant quality RPPIs for various types of property.[15] We are in favor of the (double) imputation variant because this is the most flexible hedonic approach and because this approach is analogous to the standard matched-model methodology to construct price indices.

5.44 In the next three sections, the various hedonic regression methods will be illustrated using the data for the town of "A" that was described at the end of Chapter 4. The following two sections show the results of time dummy hedonic regressions, using the log of the selling price as the dependent variable and using the untransformed selling price, respectively. The last section illustrates the hedonic imputation method. All of the resulting price indices are for the *sales* of detached houses; some results using the data for the town of "A" for indices of the *stock* of houses will be postponed until Chapter 8.

Time Dummy Models Using the Logarithm of Price as the Dependent Variable

The Log Linear Time Dummy Model

5.45 Recall the description of the data for the Dutch town of "A" on sales of detached houses. In quarter *t*, there were $N(t)$ sales of detached houses in "A" where p_n^t is the selling price of house *n* sold during quarter *t*. There is information on three characteristics of house *n* sold in period *t*: L_n^t is the area of the plot in square meters (m^2); S_n^t is the floor space area of the structure in m^2 and A_n^t is the age in decades of house *n* in period *t*. Using these variables, the

[13] However, as will be seen from the Dutch example given below, just having information on location, type of property, its age, its floor space area and the plot area may explain most of the variation in the selling price.

[14] For example, the dependent variable could be the sales price of the property or its logarithm or the sales price divided by the area of the structure and so on.

[15] This evaluation agrees with that of Hoffmann and Lorenz (2006; 15): "As far as quality adjustment is concerned, the future will certainly belong to hedonic methods." Gouriéroux and Laferrère (2009) have shown that it is possible to construct an official nationwide credible hedonic regression model for real estate properties.

standard *log linear time dummy hedonic regression model* is defined by the following system of regression equations: [16]

$$\ln p_n^t = \alpha + \beta L_n^t + \gamma S_n^t + \delta A_n^t + \tau^t + \varepsilon_n^t \qquad (5.24)$$

$$t = 1,\ldots,14;\ n = 1,\ldots,N(t);\ \tau^1 \equiv 0$$

where τ^t is a parameter which shifts the hedonic surface in quarter t upwards or downwards as compared to the surface in quarter 1. [17]

5.46 It is easy to construct a price index using the log linear time dummy hedonic model (5.24). Exponentiating both sides of equation (5.24), and neglecting the error term, yields $p_n^t = \exp(\alpha)[\exp(L_n^t)]^\beta [\exp(S_n^t)]^\gamma [\exp(A_n^t)]^\delta \exp(\tau^t)$. If we could observe a property with the *same characteristics* in the base period 1 and in some comparison period $t(>1)$, then the corresponding price relative (again neglecting error terms) would simply be equal to $\exp(\tau^t)$. For two consecutive periods t and t+1, the price relative (again neglecting error terms) would equal $\exp(\tau^{t+1})/\exp(\tau^t)$, and this can serve as the chain link in a price index. Figure 5.1 shows the resulting index, labeled as P_{H1} (hedonic index no. 1), and Table 5.1 lists the index numbers. The R^2 for this model was .8420, which is quite satisfactory for a hedonic regression model with only three explanatory variables. [18] For later comparison purposes, note that the log likelihood was 1407.6.

5.47 A problem with this model is that the underlying price formation model seems implausible: S and L interact multiplicatively in order to determine the overall house price whereas it seems most likely that lot size L and house size S interact in an approximately additive fashion to determine the overall house price.

5.48 Another problem with the regression model (5.24) is that age is entered in an additive fashion. The problem is that we would expect age to interact directly with the structures variable S as a (net) depreciation variable and not interact directly with the land variable L, because land does not depreciate. In the following model, this direct interaction of age with structures will be made.

The Log Linear Time Dummy Model with Quality Adjustment of Structures

5.49 If age A interacts with the quantity of structures S in a multiplicative manner, an appropriate explanatory variable for the selling price of a house would be $\gamma(1-\delta)^A S$ (i.e., geometric depreciation where δ is the decade geometric depreciation rate) or $\gamma(1-\delta A)S$ (straight line depreciation where δ is the decade straight line depreciation rate) instead of the additive specification $\gamma S + \delta A$. In what follows, the straight line variant of this class of models will be estimated [19]. Thus, the *log linear time dummy hedonic regression model with quality adjusted structures* becomes

$$\ln p_n^t = \alpha + \beta L_n^t + \gamma(1-\delta A_n^t)S_n^t + \tau^t + \varepsilon_n^t \qquad (5.25)$$

$$t = 1,\ldots,14;\ n = 1,\ldots,N(t);\ \tau^1 \equiv 0$$

5.50 Regression model (5.25) was run using the 14 quarters of sales data for the town of "A". Note that a single common straight line depreciation rate δ is estimated. The estimated decade (net) depreciation rate [20] was $\hat{\delta} = 11.94\%$ (or around 1.2 % per year), which is very reasonable. As was the case with model (5.24), if a house with the *same characteristics* in two consecutive periods t and t+1 could be observed, the corresponding price relative (neglecting error terms) $\exp(\tau^{t+1})/\exp(\tau^t)$ can serve as the chain link in a price index; see Figure 5.1 and Table 5.1 for the resulting index, labeled P_{H2}. The R^2 for this model was .8345, a bit lower than the previous model and the log likelihood was 1354.9, which is quite a drop from the previous log likelihood of 1407.6. [21]

5.51 It appears that the imposition of more theory – with respect to the treatment of the age of the house – has led to a drop in the empirical fit of the model. However, it is likely that this model and the previous one are misspecified [22]: they both multiply together land area times structure area to determine the price of the house while it is likely that an additive interaction between L and S is more appropriate than a multiplicative one.

[16] The estimating equation for the pooled data set will include time dummy variables to indicate the quarters. For all the models estimated for the town of "A", it is assumed that the error terms e_n^t are independently distributed normal variables with mean 0 and constant variance. Maximum likelihood estimation is used in order to estimate the unknown parameters in each regression model. The nonlinear option in Shazam was used for the actual estimation.

[17] The 15 parameters α, $\tau^1,\ldots,\tau^{14}$ correspond to variables that are exactly collinear in the regression (5.24) and thus the restriction $t^1 = 0$ is imposed in order to identify the remaining parameters.

[18] Later on in this chapter and in Chapter 8, some hedonic regressions will be run that use prices p_n^t as the dependent variables rather than the logs of the prices. To facilitate comparisons of goodness of fit across models, we will transform the predicted values for the log price models into predicted price levels by exponentiating the predicted prices and then calculating the correlation coefficient between these predicted price levels and the actual prices. Squaring this correlation coefficient gives us a *levels type measure of goodness of fit* for the log price models which is denoted by R^{*2}. For this particular model, $R^{*2} = .8061$.

[19] This regression is essentially linear in the unknown parameters and hence it is very easy to estimate.

[20] It is a net depreciation rate because we have no information on renovation expenditures, i.e., δ is equal to gross wear and tear depreciation of the house less average expenditures on renovations and repairs.

[21] The levels type R^2 for this model was $R^{*2} = .7647$, which again is quite a drop from the corresponding levels R^2 for the previous log price model.

[22] If the variation in the independent variables is relatively small, the difference in indexes generated by the various hedonic regression models considered in this section and the following two sections is likely to be small since virtually all of the models considered can offer roughly a linear approximation to the "truth". But when the variation in the independent variables is large, as it is in the present housing context, the choice of functional form can have a substantial effect. Thus a priori reasoning should be applied to both the choice of independent variables in the regression as well as to the choice of functional form. For additional discussion on functional form issues, see Diewert (2003a).

5.52 Note that, given the depreciation rate δ, *quality adjusted structures* (adjusted for the aging of the structure) for each house n in each quarter t can be defined as follows:

$$S_n^{t*} \equiv (1-\delta A_n^t)S_n^t \qquad (5.26)$$

$$t = 1,...,14;\ n = 1,...,N(t)$$

The Log Log Time Dummy Model with Quality Adjustment of Structures for Age

5.53 In the remainder of this section, quality adjusted (for age) structures, $(1-\delta A)S$, will be used as an explanatory variable, rather than the unadjusted structures area, S. The log log model is similar to the previous log linear model, except that now, instead of using L and $(1-\delta A)S$ as explanatory variables in the regression model, the logarithms of the land and quality adjusted structures areas are used as independent variables. Thus the *log log time dummy hedonic regression model with quality adjusted structures* is the following: ([23])

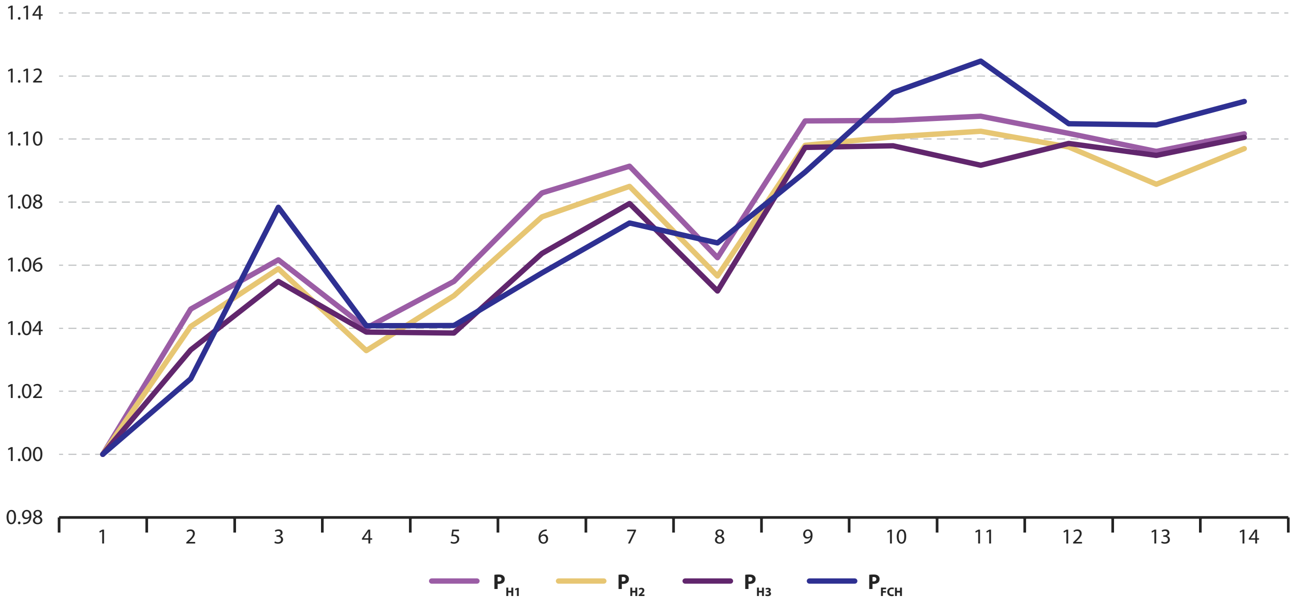
$$\ln p_n^t = \alpha + \beta \ln L_n^t + \gamma \ln[(1-\delta A_n^t)S_n^t] + \tau^t + \varepsilon_n^t \quad (5.27)$$

$$t = 1,...,14;\ n = 1,...,N(t);\ \tau^1 \equiv 0$$

5.54 Using the data for the Dutch town of "A", the estimated decade (net) depreciation rate was $\hat{\delta} = 0.1050$ (standard error 0.00374). If both sides of (5.27) were exponentiated and the error terms were neglected, the house price p_n^t would equal $\exp(\alpha)[L_n^t]^{\beta}[S_n^{t*}]^{\gamma}\exp(\tau^t)$, where S_n^{t*} denotes quality adjusted structures as defined by (5.26). So if we could observe a house with the *same characteristics* in two consecutive periods t and t+1, the corresponding price relative (neglecting error terms) would be equal to $\exp(\tau^{t+1})/\exp(\tau^t)$ and this again can serve as the chain link in a price index; see Figure 5.1 and Table 5.1 for the resulting index, labeled P_{H3}. The R^2 for this model was .8599 (the levels measure of fit was $R^{*2} = .8880$), which is an increase over models (5.25) and (5.26); the log likelihood was 1545.4, a big increase over the log likelihoods for the other two models (1407.6 and 1354.9).

5.55 The house price series generated by the three log-linear time dummy regressions described in this section, P_{H1}, P_{H2} and P_{H3}, are plotted in Figure 5.1 along with the chained stratified sample mean Fisher index, P_{FCH}. These four house price series are listed in Table 5.1. All four indices capture the same trend but there can be differences of over 2 percent between them in some quarters. Notice that all of the indices move in the same direction from quarter to quarter with decreases in quarters 4, 8, 12 and 13 except that P_{H3} – the index that corresponds to the log log time dummy model – increases in quarter 12.

Figure 5.1. Log-Linear Time Dummy Price Indices and the Chained Stratified Sample Mean Fisher Price Index



Source: Authors' calculations based on data from the Dutch Land Registry

[23] This hedonic regression model turns out to be a variant of McMillen's (2003) *consumer oriented approach to hedonic housing models*. His theoretical framework draws on the earlier work of Muth (1971) and is outlined in Diewert, de Haan and Hendriks (2010). See also McDonald (1981).

Table 5.1. Log-Linear Time Dummy Price Indices and the Chained Stratified Sample Mean Fisher Price Index

Quarter	P_{H1}	P_{H2}	P_{H3}	P_{FCH}
1	1.00000	1.00000	1.00000	1.00000
2	1.04609	1.04059	1.03314	1.02396
3	1.06168	1.05888	1.05482	1.07840
4	1.04007	1.03287	1.03876	1.04081
5	1.05484	1.05032	1.03848	1.04083
6	1.08290	1.07532	1.06369	1.05754
7	1.09142	1.08502	1.07957	1.07340
8	1.06237	1.05655	1.05181	1.06706
9	1.10572	1.09799	1.09736	1.08950
10	1.10590	1.10071	1.09786	1.11476
11	1.10722	1.10244	1.09167	1.12471
12	1.10177	1.09747	1.09859	1.10483
13	1.09605	1.08568	1.09482	1.10450
14	1.10166	1.09694	1.10057	1.11189

Source: Authors' calculations based on data from the Dutch Land Registry

5.56 Although model (5.27) performs the best of the simple hedonic regression models considered thus far, it has the unsatisfactory feature that the quantities of land and of quality adjusted structures determine the price of a property in a *multiplicative manner*. It is more likely that house prices are determined by a weighted *sum* of their land and quality adjusted structures amounts. In the following section, an additive time dummy model will therefore be estimated. The expectation is that this model will fit the data better.

Time Dummy Hedonic Regression Models using Price as the Dependent Variable

The Linear Time Dummy Hedonic Regression Model

5.57 There are reasons to believe that the selling price of a property is linearly related to the plot area of the property plus the area of the structure due to the competitive nature of the house building industry.[24] If the age of the structure is treated as another characteristic that has an importance in determining the price of the property, then the following *linear time dummy hedonic regression model* might be an appropriate one:

$$p_n^t = \alpha + \beta L_n^t + \gamma S_n^t + \delta A_n^t + \tau^t + \varepsilon_n^t \qquad (5.28)$$

$$t = 1,...,14;\ n = 1,...,N(t);\ \tau^1 \equiv 0$$

5.58 The above linear regression model was run using the data for the town of "A". The R^2 for this model was .8687, much higher than those obtained in the previous regressions[25]; the log likelihood was -10790.4 (which cannot easily be compared to the previous log likelihoods since the dependent variable has changed from the logarithm of price to just price[26]).

5.59 Using the linear model defined by equations (5.28) to form an overall house price index is a bit more difficult than using the previous log-linear or log log time dummy regression models. In the previous section, holding characteristics constant and neglecting error terms, the *relative price* for the same house over any two periods turns out to be constant, leading to an unambiguous overall index. In the present situation, holding characteristics constant and neglecting error terms, the *difference in price* for the same house turns out to be constant, but the *relative prices* for different houses will not in general be constant. Therefore, an overall index will be constructed which uses the prices generated by the estimated parameters for model (5.28)

[24] See Clapp (1980), Francke and Vos (2004), Gyourko and Saiz (2004), Bostic, Longhofer and Redfearn (2007), Davis and Heathcote (2007), Francke (2008), Diewert (2009b), Koev and Santos Silva (2008), Statistics Portugal (2009), Diewert, de Haan and Hendriks (2010), Diewert (2010) and Chapter 8 below.

[25] However, recall that the levels adjusted measure of fit for the log log model described by (5.27) was .8880, which is higher than .8687.

[26] Marc Francke has pointed out that it is possible to compare log likelihoods across two models where the dependent variable has been transformed by a known function in the second model; see Davidson and McKinnon (1993; 491) where a Jacobian adjustment makes it possible to compare log likelihoods across the two models.

and evaluated at the sample average amounts of L, S and the sample average age of a house A.[27] The resulting quarterly prices for this "average" house were converted into an index, P_{H4}, which is listed in Table 5.2 and charted in Figure 5.2.

5.60 The hedonic regression model defined by (5.28) is perhaps the simplest possible one but it is a bit too simple since it neglects the fact that the interaction of age with the selling price of the property takes place via a multiplicative interaction with the structures variable and not via a general additive factor. In what follows, model (5.28) is re-estimated using quality adjusted structures as an explanatory variable rather than just entering age A as a separate stand alone characteristic.

The Linear Time Dummy Model with Quality Adjusted Structures

5.61 The *linear time dummy hedonic regression model with quality adjusted structures* is described by

$$p_n^t = \alpha + \beta L_n^t + \gamma(1-\delta A_n^t)S_n^t + \tau^t + \varepsilon_n^t \qquad (5.29)$$

$$t = 1,...,14;\ n = 1,...,N(t);\ \tau^1 \equiv 0$$

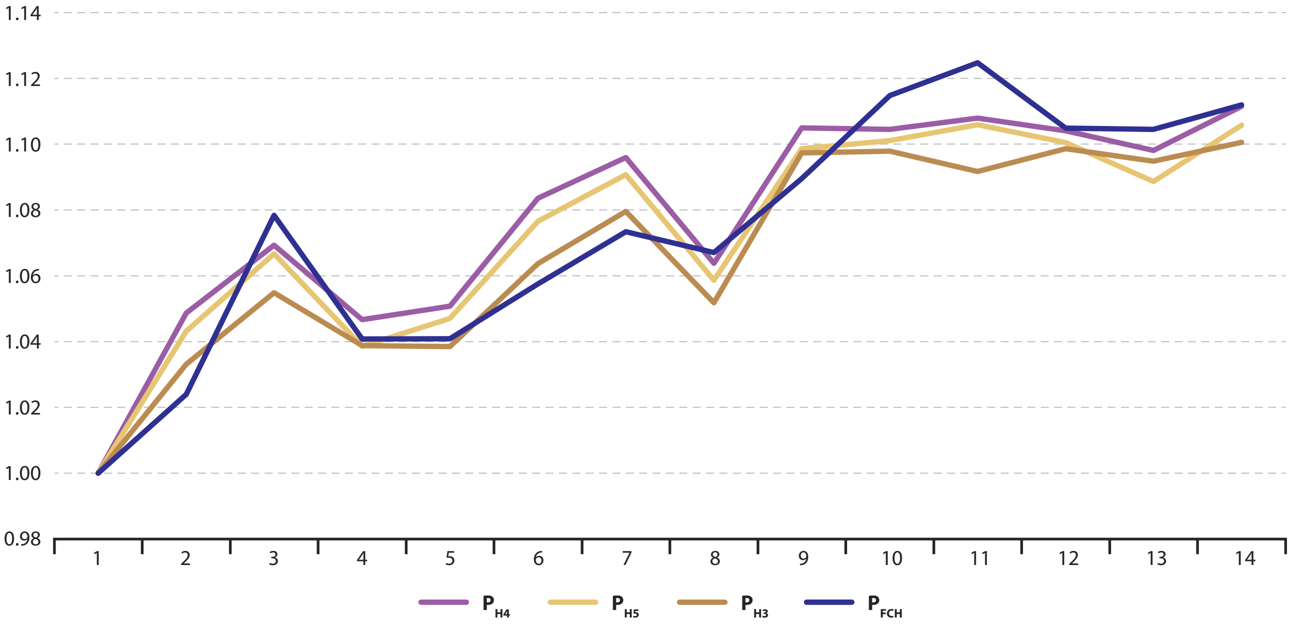
This is the most plausible hedonic regression model so far. It works with quality adjusted (for age) structures S^* equal to $(1-\delta A)S$ instead of having A and S as completely independent variables that enter into the regression in a linear fashion.

5.62 The results for this model were a clear improvement over the results of model (5.28). The log likelihood increased by 92 to -10697.8 and the R^2 increased to .8789 from the previous .8687. The estimated decade depreciation rate was $\hat{\delta} = 0.1119\ \ (0.00418)$, which is reasonable as usual. This linear regression model has the same property as the model (5.28): house price *differences* are constant over time for all constant characteristic models but house price *ratios* are not constant. So again an overall index will be constructed which uses the prices generated by the estimated parameters in (5.29) and evaluated at the sample average amounts of L, S and the average age of a house A. The resulting quarterly house prices for this "average" model were converted into an index, P_{H5}, which is listed in Table 5.2 and charted in Figure 5.2. For comparison purposes, P_{H3} (the time dummy Log Log model index) and P_{FCH} (the chained stratified sample mean Fisher index) will be charted along with P_{H4} and P_{H5}. The preferred indices thus far are P_{FCH} and P_{H5}.

5.63 It can be seen that again, all four indices capture the same trend but there can be differences of over 2 percent between the various indices for some quarters. Note that all of the indices move in the same direction from quarter to quarter with decreases in quarters 4, 8, 12 and 13, except that P_{H3} increases in quarter 12.

[27] The sample average amounts of L and S were 257.6 m² and 127.2 m² respectively and the average age of the detached dwellings sold over the sample period was 1.85 decades.

Figure 5.2. Linear Time Dummy Price Indices, the Log Log Time Dummy Price Index and the Chained Stratified Sample Mean Fisher Price Index



Source: Authors' calculations based on data from the Dutch Land Registry

Table 5.2. Linear Time Dummy Price Indices, the Log Log Time Dummy Price Index and the Chained Stratified Sample Mean Fisher Price Index

Quarter	P_{H4}	P_{H5}	P_{H3}	P_{FCH}
1	1.00000	1.00000	1.00000	1.00000
2	1.04864	1.04313	1.03314	1.02396
3	1.06929	1.06667	1.05482	1.07840
4	1.04664	1.03855	1.03876	1.04081
5	1.05077	1.04706	1.03848	1.04083
6	1.08360	1.07661	1.06369	1.05754
7	1.09593	1.09068	1.07957	1.07340
8	1.06379	1.05864	1.05181	1.06706
9	1.10496	1.09861	1.09736	1.08950
10	1.10450	1.10107	1.09786	1.11476
11	1.10788	1.10588	1.09167	1.12471
12	1.10403	1.10044	1.09859	1.10483
13	1.09805	1.08864	1.09482	1.10450
14	1.11150	1.10572	1.10057	1.11189

Source: Authors' calculations based on data from the Dutch Land Registry

5.64 A problem with the hedonic time dummy regression models considered thus far is that the prices of land and quality adjusted structures are not allowed to change in an unrestricted manner from period to period. The class of hedonic regression models to be considered in the following section does not suffer from this problem.

Hedonic Imputation Regression Models

5.65 The theory of *hedonic imputation indices* explained earlier is applied to the present situation as follows. For each period, run a linear regression of the following form:

$$p_n^t = \alpha^t + \beta^t L_n^t + \gamma^t(1-\delta^t A_n^t)S_n^t + \varepsilon_n^t \qquad (5.30)$$

$t = 1,\ldots,14;\ n = 1,\ldots,N(t)$

Using the data for the town of "A", there are only four parameters to be estimated for each quarter: α^t, β^t, γ^t and δ^t for $t = 1,\ldots,14$. Note that (5.30) is similar in form to the model defined by equations (5.29), but with some significant differences:

- Only one depreciation parameter is estimated in the model defined by (5.29) whereas in the present model, there are 14 depreciation parameters; one for each quarter.
- Similarly, in model (5.29), there was only one α, β and γ parameter whereas in (5.30), there are 14 α^t, 14 β^t and 14 γ^t parameters to be estimated. On the other hand, model (5.29) had an additional 13 time shifting parameters (the τ^t) that required estimation.

Thus the hedonic imputation model involves the estimation of 56 parameters, the time dummy model only 17, so it is likely that the hedonic imputation model will fit the data much better.

5.66 In the housing context, precisely matched models across periods do not exist; there are always depreciation and renovation activities that make a house in the exact same location not quite comparable over time. This lack of matching, say between quarters *t* and *t*+1, can be overcome in the following way: take the parameters estimated using the quarter *t*+1 hedonic regression and price out all of the housing models (i.e., sales) that appeared in quarter *t*. This generates *predicted quarter t+1 prices for the quarter t models*, $\hat{p}_n^{t+1}(t)$, as follows:

$$\hat{p}_n^{t+1}(t) \equiv \hat{\alpha}^{t+1} + \hat{\beta}^{t+1}L_n^t + \hat{\gamma}^{t+1}(1-\hat{\delta}^{t+1}A_n^t)S_n^t \qquad (5.31)$$

$t = 1,\ldots,13;\ n = 1,\ldots,N(t)$

where $\hat{\alpha}^t$, $\hat{\beta}^t$, $\hat{\gamma}^t$ and $\hat{\delta}^t$ are the parameter estimates for model (5.30) for $t = 1,\ldots,14$. Now we have a set of pseudo matched quarter *t*+1 prices for the models that appeared in quarter *t* and the following *Laspeyres type hedonic imputation (or pseudo matched model) index*, going from quarter *t* to *t*+1, can be formed: [28]

$$P_{HIL}^{t,t+1} \equiv \frac{\sum_{n=1}^{N(t)} 1\hat{p}_n^{t+1}(t)}{\sum_{n=1}^{N(t)} 1p_n^t} \qquad (5.32)$$

$t = 1,\ldots,13$

[28] Due to the fact that the regressions defined by (5.30) have a constant term and are essentially linear in the explanatory variables, the sample residuals in each of the regressions will sum to zero. Hence the sum of the predicted prices will equal the sum of the actual prices for each period. Thus the sum of the actual prices in the denominator of (5.32) will equal the sum of the corresponding predicted prices and similarly, the sum of the actual prices in the numerator of (5.34) will equal the corresponding sum of the predicted prices.

As mentioned earlier, the quantity that is associated with each price is 1 as each housing unit is basically unique and can only be matched through the use of a model.

5.67 The same method can be applied going backwards from the housing sales that took place in quarter t+1; take the parameters for the quarter t hedonic regression and price out all of the housing models that appeared in quarter t+1 and generate predicted prices, $\hat{p}_n^t(t+1)$, for these t+1 models:

$$\hat{p}_n^t(t+1) \equiv \hat{\alpha}^t + \hat{\beta}^t L_n^{t+1} + \hat{\gamma}^t(1-\hat{\delta}^t A_n^{t+1})S_n^{t+1} \qquad (5.33)$$

t = 1,...,13; n = 1,...,N(t+1)

Now we have a set of "matched" quarter t prices for the models that appeared in period t+1 and we can form the following *Paasche type hedonic imputation (or pseudo matched model) index*, going from quarter t to t+1:

$$P_{HIP}^{t,t+1} \equiv \frac{\sum_{n=1}^{N(t+1)} 1 p_n^{t+1}}{\sum_{n=1}^{N(t+1)} 1 \hat{p}_n^t(t+1)} \qquad (5.34)$$

t = 1,...,13

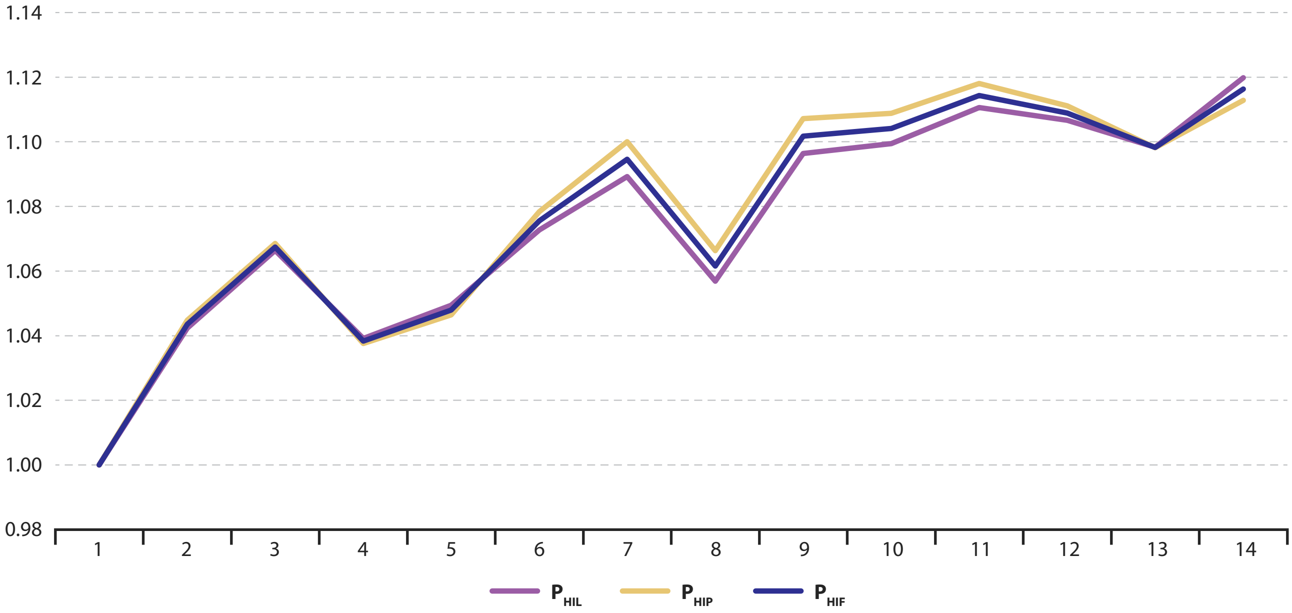
5.68 Once the above Laspeyres and Paasche imputation price indices have been calculated, the corresponding *Fisher type hedonic imputation index* going from period t to t+1 can be formed by taking the geometric average of the two indices defined by (5.32) and (5.34):

$$P_{HIF}^{t,t+1} \equiv \left[P_{HIL}^{t,t+1} P_{HIP}^{t,t+1}\right]^{1/2} \qquad (5.35)$$

t = 1,...,13

5.69 The resulting chained Laspeyres, Paasche and Fisher imputation price indices, P_{HIL}, P_{HIP} and P_{HIF}, based on the data for the town of "A", are plotted below in Figure 5.3 and are listed in Table 5.3. The three imputation indices are amazingly close. The Fisher imputation index is our preferred hedonic price index thus far; it is better than the time dummy indices because imputation allows the price of land and of quality adjusted structures to change independently over time, whereas the time dummy indices shift the hedonic surface in a parallel fashion. The empirical results indicate that, at least for the present data set for the town of "A", the Laspeyres imputation index provides a close approximation to the preferred Fisher imputation index.

Figure 5.3. Chained Laspeyres, Paasche and Fisher Hedonic Imputation Price Indices



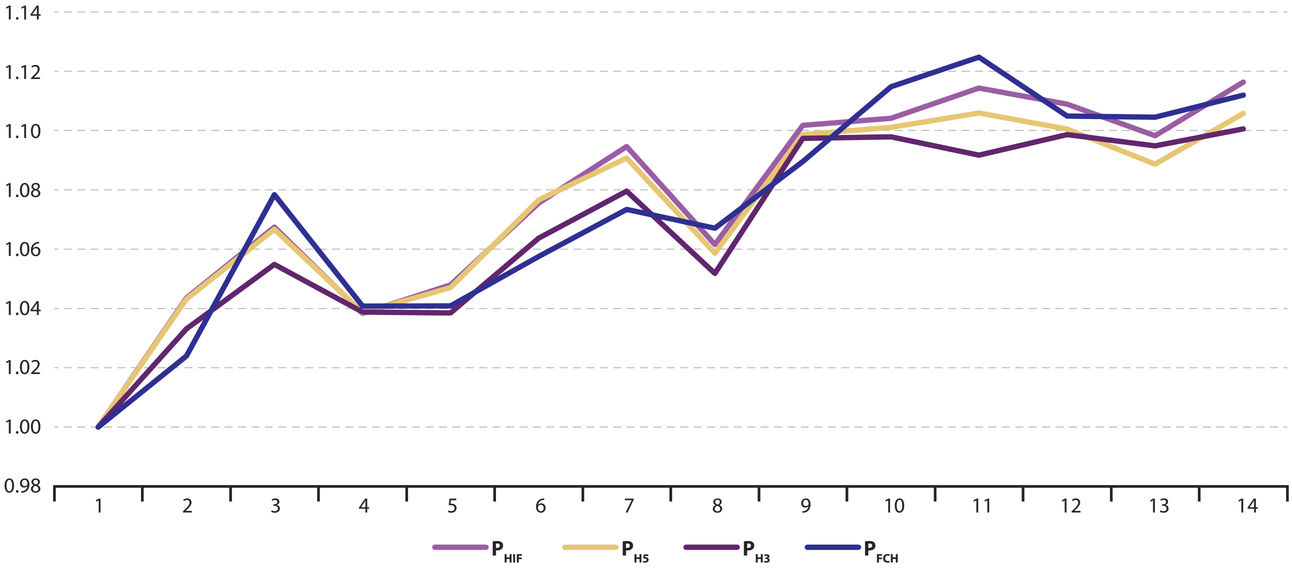
Source: Authors' calculations based on data from the Dutch Land Registry

Table 5.3. Chained Laspeyres, Paasche and Fisher Hedonic Imputation Price Indices

Quarter	P_{HIL}	P_{HIP}	P_{HIF}
1	1.00000	1.00000	1.00000
2	1.04234	1.04479	1.04356
3	1.06639	1.06853	1.06746
4	1.03912	1.03755	1.03834
5	1.04942	1.04647	1.04794
6	1.07267	1.07840	1.07553
7	1.08923	1.10001	1.09460
8	1.05689	1.06628	1.06158
9	1.09635	1.10716	1.10174
10	1.09945	1.10879	1.10411
11	1.11062	1.11801	1.11430
12	1.10665	1.11112	1.10888
13	1.09830	1.09819	1.09824
14	1.11981	1.11280	1.11630

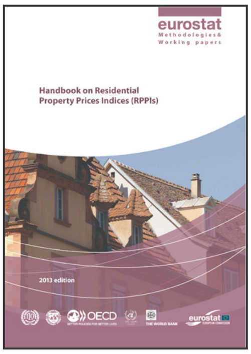
Source: Authors' calculations based on data from the Dutch Land Registry

Figure 5.4. The Fisher Imputation Price Index, the Chained Stratified Sample Mean Fisher Price Index, the Linear Time Dummy Price Index and the Log Log Time Dummy Price Index



Source: Authors' calculations based on data from the Dutch Land Registry

5.70 To conclude: our two "best" indices are the Fisher imputation index P_{HIF} and the stratified chained Fisher index P_{FCH}. Overall, the imputation index P_{HIF} should probably be preferred to P_{FCH} since the stratified sample indices will have a certain amount of unit value bias which will most likely be greater than any functional form bias in P_{HIF}. These two "best" indices are plotted in Figure 5.4 along with the log-log time dummy index P_{H3} and the linear time dummy index with quality adjusted structures P_{H5}. All of the price indices except P_{H3} show downward movements in quarters, 4, 8, 12 and 13 and upward movements in the other quarters; P_{H3} moves up in quarter 12 instead of falling like the other indices.



From:

Handbook on Residential Property Price Indices

Access the complete publication at:

https://doi.org/10.1787/9789264197183-en

Please cite this chapter as:

de Haan, Jan and Erwin Diewert (2013), “Hedonic Regression Methods”, in OECD*, et al.*, *Handbook on Residential Property Price Indices*, Eurostat, Luxembourg.

DOI: https://doi.org/10.1787/9789264197183-7-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.