Chapter 2: Theory of hedonic indices

2.1 The rationale for hedonic approaches

The categories in a classification of goods and services, even if restricted, include some widely varying sets of characteristics. Dwellings, for instance (the goods being examined here) are defined by their location, their floor space, the number of rooms, whether occupied or vacant, age, and so on. Housing rentals (services) are differentiated further by lease durations, renewal terms, and early termination clauses.

This variety is reflected in the corresponding markets, leading to differences in housing turnover or occupation rates, and widely differing prices and this in turn can pose various problems for price analysis.

The main difficulty is due to the fact that the price of a dwelling can only be observed when the transactions take place, and these are infrequent (partial observability); similarly, a rent is observable only if the service is actually used. Outside of these situations, these values – prices and rents – have no existence in economic terms.

Traditionally, the way to get round this problem is to assume *implicit values*, also called, in our context, the "estimated price" of the dwelling and the "rental value". These implicit values are known only when they coincide with transaction prices or rents, and cannot therefore be reconstructed except from models describing the composition of the prices or rents, and any changes that occur.

Hedonic approaches rely on models such as these and explain how to use them in order to construct the non-observed values and define consistent sets of price indices.

2.1.1 Hedonic model with a priori *stratification*

We shall present the hedonic approach using a simplified formulation. Let us assume that the different goods can be aggregated into strata pre-defined in such a way that price changes are approximately parallel within the same stratum. Between strata, however, prices may differ considerably.

In our application, the strata will be geographic areas, the primary areas where dwellings are located. These strata are denoted *s, s=1,…,S.* Given a dwelling *i* in stratum *s*, with characteristics $z_{i,s,t}$ (floor space, number of rooms, etc.), we use a regression model to estimate its implicit value $p^*_{i,s,t}$ on date *t*.

Within the stratum *s*, the implicit values are assumed to be such that:

$$p^*_{i,s,t} \approx c(s, z_{i,s,t})\, p^*_{s,t} \qquad (2.1)$$

- where $p^*_{s,t}$ is an implicit reference value for stratum *s* on date *t*,
- $c(s,z_{i,s,t})$ is an adjustment coefficient that takes into account the characteristics of the good and which may be stratum-dependent,
- $\approx$ means "approximately equal to".

The *adjustment* and the *reference value* are defined to within one multiplicative scalar (which poses an identification problem). It is then possible to set the reference value as corresponding to a property of pre-specified quality (reference dwelling), z_0:

$$p^*_{0,s,t} \approx p^*_{s,t} \Leftrightarrow c(s, z_0) = 1$$

let us say, for example, a two-room, used apartment on the ground floor, etc.

$\dfrac{p^*_{i,s,t}}{c(s, z_{i,s,t})}$ will be called the "reference-property equivalent price" in terms of dwelling (i,s) on date t.

The approximation (2.1) can be made to resemble an econometric model more closely by introducing appropriate error terms and specifying a parameterized form for the adjustment coefficient. Such a specification might be, for example:

$$\log p^*_{i,s,t} = \sum_{k=1}^{K} \beta_{k,s} X_k\left(z_{i,s,t}\right) + \log p^*_{s,t} + \varepsilon^*_{i,s,t} \qquad (2.2)$$

where the error terms $\varepsilon^*_{i,s,t}$ are assumed to be independent, zero-mean and of variance $\eta^2_{s,t}$ in some cases stratum- and date-dependent. The $X_k\left(z_{i,s,t}\right)$ values are explanatory variables, *K* in number, which are functions of housing characteristics or of combinations of certain characteristics.

Formula (2.2) above has the advantage of expressing a linear model in the parameters $\hat{c}_0(s,z) = \exp\left[\sum_{k=1}^{K} \hat{\beta}_{k,s} X_k(z)\right]$, which will enable us to define the adjustment coefficient, and $\log p^*_{s,t}$, which will then give the stratum reference value.

Note that the model incorporates combined effects of the stratum and other housing characteristics, $\beta_{k,s}$ coefficients being stratum specific.

2.1.2 Problematic alternative approaches

Can one reconstruct price changes without using a hedonic model? Two alternative approaches are possible: one is the repeat-sales method, but this is difficult to implement rigorously, and the other is based on observation of average prices, but produces biased results.

Although a given property *i* is seldom traded on two given successive dates *t*, *t+1* (in which case we would speak of repeat data), we can nevertheless hope to measure its change in price by comparing prices of two similar properties. If the explanatory variables $X_k(z)$ are qualitative, for example, there may be several properties in the stratum with approximately the same price levels and not only parallel changes. These prices will be used to calculate price change. However, once the quality effects reach a certain intensity and number, the market may not be liquid enough for such a comparison to be possible. Furthermore, if we base our calculations solely on repeat sales prices, we neglect a large share of the information contained in transaction data.

A second approach often suggested is to compare the average price of properties in stratum *s* traded at *t+1*, with the average price of properties in the same stratum traded at *t*, in the hope of proxying change in the reference value. This approach is biased, however, as the quality structure of traded properties is not stable over time. To illustrate this difficulty, let us consider a case in which only a single property is traded on each date in the stratum, with the property noted as being of quality z_t at date *t*. The observed price ratio would be approximately:

$$\frac{c(s,z_{t+1})}{c(s,z_t)} \frac{p^*_{s,t+1}}{p^*_{s,t}}$$

and would differ from the change in reference value because of the change in the adjustment coefficient.

2.2 Using the hedonic model

The model is essentially used to compensate for the partial inobservability of the data and to adjust for quality effects. The process consists of several steps, as outlined below.

- Step 1: estimation of adjustment coefficients using a predefined set of transaction data, hereafter called *estimation stock*. Choice of a set of dwellings, called the *reference stock*, whose prices will be tracked.
- Step 2: at each date *t*, use of actual transaction data and adjustment coefficients estimated in step 1 to reconstruct the values of reference stock dwellings.
- Step 3: use of reference values and adjustment coefficients to construct an array of price indices and an expert valuation system.

2.2.1 Estimating adjustment coefficients

The model is mainly used to consider an estimation stock consisting of transactions occurring in a predefined period $t=1,\ldots,T_0$ (called *estimation period*). The transaction data provide price-quality pairs (price $p_{j,s,t}$, quality $z_{j,s,t}$), $j=1,\ldots, J_{s,t}$, $s=1,\ldots,S$, $t=1,\ldots,T_0$, where $J_{s,t}$ denotes the number of transactions in stratum s in period t.

We then estimate parameters $\beta_{k,s}, k=1,...,K$ in each stratum, using ordinary least squares, from equation (2.2). From this we deduce the estimated adjustment coefficient of the stratum:

$$\hat{c}_0(s,z) = \exp\left[\sum_{k=1}^{K} \hat{\beta}_{k,s} X_k(z)\right]$$

and also, if the precision of the estimated coefficients allows, a value range for each of these terms, :

$$\left[\hat{c}_1(s,z), \hat{c}_2(s,z)\right]$$

These adjustments will remain unchanged throughout a future period, whose length has to be specified.

2.2.2 Estimating reference value for date t

Let us now consider a date t, subsequent to the estimation period and transaction data: $p_{j,s,t}$, $z_{j,s,t}$, for each of the properties j in stratum s on date t. We have:

$$\log p_{j,s,t} = \log c\left(s, z_{j,s,t}\right) + \log p^*_{s,t} + \varepsilon_{j,s,t}$$

$$\cong \log \hat{c}_0\left(s, z_{j,s,t}\right) + \log p^*_{s,t} + \varepsilon_{j,s,t}$$

after replacing the adjustments by their proxies obtained in the first step.

We obtain $J_{s,t}$ estimations of the reference property price in stratum s on date t, $p^*_{s,t}$, and we deduce the approximation of $p^*_{s,t}$ using ordinary least squares:

$$\log \hat{p}^*_{s,t} = \frac{1}{J_{s,t}} \sum_{j=1}^{J_{s,t}} \left[\log p_{j,s,t} - \log \hat{c}_0\left(s, z_{j,s,t}\right)\right]$$

$$\Leftrightarrow \hat{p}^*_{s,t} = \prod_{j=1}^{J_{s,t}} \left[\frac{p_{j,s,t}}{\hat{c}_0\left(s, z_{j,s,t}\right)}\right]^{\frac{1}{J_{s,t}}}$$

$p^*_{s,t}$ is thus estimated from the geometric mean of the reference-property equivalent prices. By taking standard deviations into account, we can also propose a range for the reference value. More specifically, note that:

$\hat{\eta}^2_{s,t}$ is the empirical variance of the values $\log\left(p_{j,s,t} / \hat{c}_0\left(s, z_{j,s,t}\right)\right), j=1,...,J_{s,t}$.

This empirical variance proxies variance $\eta^2_{s,t}$ of the error term. The reference value admits a logarithm lying between $\log \hat{p}^*_{s,t} - 2\hat{\eta}_{s,t}$ and $\log \hat{p}^*_{s,t} + 2\hat{\eta}_{s,t}$, and the range is:

$$\left(\hat{p}^*_{1,s,t} = \exp\left(-2\hat{\eta}_{s,t}\right)\hat{p}^*_{s,t}\ , \hat{p}^*_{2,s,t} = \exp\left(2\hat{\eta}_{s,t}\right)\hat{p}^*_{s,t}\right)$$

2.2.3 Constructing a valuation system

We can now estimate the implicit values of any property at date *t*, of quality z, which is not necessarily traded.

The estimate is:

$$\hat{p}^{*}_{s,t}\hat{c}_0(s,z)$$

the range can be taken as equal to:

$$\left(\hat{p}^{*}_{1,s,t}\hat{c}_1(s,z),\ \hat{p}^{*}_{2,s,t}\hat{c}_2(s,z)\right)$$

2.2.4 Constructing an array of indices

The hedonic method has enabled us to define primary indices stratum by stratum:

$$I_{s,t} = \hat{p}^{*}_{s,t}, s = 1,...,S,\ t = 1,...,T$$

which can serve as a base for constructing composite indices. This construction uses traditional methods, such as a Laspeyres-index approach. For this, we define a basket of properties (housing stock), called the reference stock, from the initial period, specifying the qualities *z* and strata *s* concerned. *Z* denotes the different qualities introduced in the basket (reference stock) and $N_{z,s}$ the number of properties (dwellings) with characteristics *z* in stratum *s*.

the composite index is defined from the value of this basket:

$$I_t = \sum_{s=1}^{S}\sum_{z\in Z} N_{z,s}I_{s,t}$$

This basket can also serve to construct composite indices disaggregated in a coherent manner. For example, we can introduce a "two-room" index [French *deux pièces*] and consider the sub-stock composed only of two-room dwellings:

$$I_t(deux\,pièces) = \sum_{s=1}^{S}\ \sum_{z\in Z\ (\,deux\,pièces\,)}\ N_{z,s}\ I_{s,t}$$

or an index for a given region, calculated on a sub-set of strata:

$$I_t(région) = \sum_{s\in région}\ \sum_{z\in Z} N_{z,s}I_{s,t}$$

For a standardised approach, the convention is to take a base year, say *t = 0*. The standardisation should then be carried out index by index, giving indices with base *100* at *t = 0* written as:

$$I_{t/0} = 100\ I_t / I_0$$
$$I_{t/0}(deux\,pièces) = 100\ I_t(deux\,pièces) / I_0(deux\,pièces)$$

2.3 Making the hedonic approach more robust

The hedonic approach, which relies on estimates, can be sensitive to their precision or to parameter instability. It may be useful to aggregate some quality variables or strata in order to make the results more significant.

2.3.1 The search for underlying scores

Let us consider the adjustment coefficients. For each stratum, we have estimated a set of parameters $\hat{\beta}_{1,s},...\hat{\beta}_{K,s}$ defining the stratum *score* $s : \sum_{k=1}^{K} \hat{\beta}_{k,s} X_k$.

We can examine whether these *S* scores depend on a smaller number of underlying scores. This approach is as follows:

1. We define the matrix $\hat{B}$, of size *(K,S)*, whose columns are the vectors $\left(\hat{\beta}_{1,s},..., \hat{\beta}_{K,s}\right)$;
2. We determine the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 ... \geq \hat{\lambda}_S$ and the associated eigenvectors $\hat{\alpha}_1,...,\hat{\alpha}_S$ of matrix $\hat{B}'\hat{B}$, where $\hat{B}'$ denotes the transposition of $\hat{B}$.
3. The number of eigenvalues S_0 significantly different from zero yields the number of independent underlying scores which are given by:

$$Z_l = \hat{\gamma}_l' X = \hat{\alpha}_l' \hat{B}' X, l = 1,...,S_0$$

4. To make the model more robust, we constrain the adjustment coefficient to take the form:

$$c(s,z) = \exp\left[\sum_{l=1}^{S_0} \lambda_{l,s} Z_l\right]$$

As S_0 is less than both *K* and *S*, and often fairly small, there are far fewer parameters to estimate in this constrained form of adjustment coefficient.

Even if it requires a partial disaggregation of the scores Z_l deduced from the approach above, it is customary to select sub-scores that contain only variables with interpretations of the same type. For example, one sub-score will adjust for the physical characteristics of the dwelling, a second for its amenities, a third for the quality of the environment, a fourth for location, and so on. We thus obtain a hierarchical structure of the effects of the variables, making it easier to set up, interpret and update expert valuation systems.

2.3.2 Strata aggregation

Similarly, we can examine the strata via changes in the corresponding indices $I_{s,t}$. By analysing empirical correlations between these time series, we may be able to identify strata whose reference values are moving in parallel. If the result is interpretable, we can then aggregate those strata.

2.4 Monitoring the specification

Parameter values may change over time and adjustment coefficients determined in the estimation period may deteriorate. It is important to set up instruments to monitor the quality of the model so as to be able to identify the point at which we need to re-estimate it, and also to develop ideas about any changes that should be made. We can then go on to a suitable examination of estimation residuals.

At each date *t* the residuals are:

$$\hat{\varepsilon}_{j,s,t} = \log p_{j,s,t} - \log \hat{c}_0\left(s, z_{j,s,t}\right) - \log p^*_{s,t}, j = 1,..., J_{s,t}$$

They must be aggregated in order to eliminate quality effects and steer the monitoring process towards the parameters $\beta_{k,s}$ which are liable to be affected. To this end, we can consider various marginal characteristics,

for example: two rooms (d.p. [for French *deux pièces*]), used dwelling (a. [for French *ancien*]), etc., and compute the mean residuals for each.[18]

$\hat{\varepsilon}_{s,t}(d.p.) =$ mean of $\hat{\varepsilon}_{j,s,t}$ values for two-room dwellings of stratum *s* at date *t*,

$\hat{\varepsilon}_{s,t}(a.) =$ comparable mean for used dwellings, etc.

If the model is properly specified, such means should vary around zero. We shall therefore search for recurring divergences. For example, if we find $\hat{\varepsilon}_{s,t}(a.)$ values that are too often positive for stratum s_0, from a particular date t_0, we may need to adjust the value of a parameter for an explanatory variable X_{k_0} as a function of dwelling age. If a trend emerges in the series $t \to \hat{\varepsilon}_{S_0,t}(a.)$ this may mean that the proportionality hypothesis for price changes within stratum *s* is no longer fulfilled and that this stratum needs to be decomposed.

2.5 Extending the basic model

The hedonic model used by way of illustration until now has a disadvantage in that it pre-defines homogeneous strata for price changes. We can generalise this model as follows.

Let us begin by writing formula (2.2) in a condensed form:

$$\log\ p^*_{i,s,t} = \sum_{s_0=1}^{S}\sum_{k=1}^{K} \xi_{s_0}(i,s,t)\beta_{k,s_0} X_k(z_{i,s,t}) + \sum_{s_0=1}^{S} \log\ p^*_{s_0,t}\ \xi_{s_0}(i,s,t) + \varepsilon^*_{i,s,t}$$

where ξ_{s_0} denotes the dummy variable of stratum s_0, i.e. the variable that equals 1 if the observation is in stratum s_0, otherwise 0. In this form, the model applies to all data in all strata, and remains linear in the various parameters:

$$\beta_{k,s_0}, k = 1,...,K,\ s_0 = 1,...,S,\ \log p^*_{s_0,t},\ s_0 = 1,...,S,\ t = 1,...,T$$

Its limitations clearly appear in the expressions of the explanatory variables: the "constant term" part includes highly specific stratum × quality cross-effects, whereas the parts giving the dynamics of the reference values $p^*_{s_0,t}$ do not incorporate quality effects other than the stratum dummy.

An enlarged model may thus be written:

$$\log p^*_{i,t} = c_0(z_{i,t};\theta_0) + \sum_{l=1}^{L} c_l(z_{i,t};\theta_l) f_{l,t} + \varepsilon^*_{i,t} \qquad (2.3)$$

where $f_{1,t},...,f_{L,t}$ are dynamic factors to be determined, $c_0(z_{i,t};\theta_0),...,c_L(z_{i,t};\theta_L)$ are adjustment coefficients giving, for example, sensitivities to factors, with parameters $\theta_0,...,\theta_L$ to be estimated. There is no longer any need to distinguish stratum *s* which is reincorporated among the other characteristics of the dwelling.

The hedonic approach can be applied from such a model. For example, in the current situation at date *t*, factor values will be estimated using ordinary least squares on the proxied model:

$$\log p_{j,t} \cong \hat{c}_0(z_{j,t}) + \sum_{l=1}^{L} \hat{c}_l(z_{j,t}) f_{l,t}$$

where the $\hat{c}_l(z_{j,t})$ values are determined for the estimation period. We shall not go any further into a discussion on the corresponding estimation procedures. Let us simply note that model (2.3) enables us to determine the best factor forecasts, assuming all implicit prices are observable. These forecasts appear as linear combinations of price logarithms, with coefficients whose sum can be standardised to unity:

[18] If the number of transactions proves insufficient to compute such means, the period over which the mean is computed can always be lengthened by considering two or three consecutive dates.

$$f_{l,t}^{*} \cong \sum_{i=1}^{N} \pi_l(z_{it}) \log p_{i,t}^{*}$$

$$\Leftrightarrow \exp f_{l,t}^{*} = \prod_{i=1}^{N} \left[p_{i,t}^{*}\right]^{\prod_{l(z_{i,t})}}$$

Each factor is thus implicitly linked to a time-varying basket (stock) of composition $(\pi_l(z_{i,t}))$, with *i* varying), such that the change in f_l is close to that of the basket (stock) value. This is referred to as the *factor-mimicking basket (stock)* (Huberman, Kandel, Stambaugh, 1987). In other words, we can make a suitable choice of disaggregated indices whose changes will reproduce those of the underlying factors.