

Hedonic Methods and Housing Markets

Chapter 4: Price Indexes Across Space and Over Time

Edward Coulson
Department of Economics
Penn State University
fyj@psu.edu

4.1 Price indexes

In 2005, two prominent business news websites ran articles on the cost of housing across the US. According to Forbes.com, the most expensive housing in the United States was to be found in Northern California town of Atherton. The ranking for their competition was based on the median sales price of single family homes by zip code, and Atherton topped the list at a robust \$2.4 million. The median price--that is, the dollar figure that 50% of the local sales were more expensive than-- was thought to be an appropriate measure of the typical house in that zip code.

However, according to cnn.com, the US's the most expensive housing was in La Jolla, California. The cnn.com rankings were done by a real estate brokerage service, which estimated, for a large number of locations, the price of a housing units with a certain set of housing characteristics: "Coldwell Banker looked at a 2,200-square-foot house with 4 bedrooms, 2 1/2 bathrooms, a family room and a two-car garage. The neighborhood - a more subjective measure - is one 'typical for corporate middle-management transferees'".

That the two surveys had different results-- throughout the list, and not just at the top--should come as no surprise, because they are measuring different things. In the Forbes.com survey, the median sales price is measuring the price of an average housing unit *for that area*, and the average housing unit can differ substantially across different areas. As pointed out in Coulson (2009) just one indication of this is that the 2000 census shows that the median number of rooms in the Atherton zip code (94027) was 8.2, whereas the national average was 6.2. So the average house in Atherton is much bigger than the average house in the US. La Jolla, the leader in the cnn.com survey, ranked only 73rd in the Forbes.com list. The 2000 Census gives the median number of rooms in the La Jolla zip code (92037) as 6.2, the same as in the national

housing market. One can reasonably conclude that “homes in La Jolla are smaller than the mansion communities seen at the top of the Forbes list, but that on a size-and-quality adjusted basis, La Jolla is the pricier location.” (Coulson, 2009). The cnn.com list, by specifying the features of the housing unit to be priced, is comparing apples to apples, and the Forbes list is not (as was recognized by the author of the latter).

When we ask the questions: Can the author of this monograph afford a house in Atherton? the answer is of course, no. But there are two reasons. One is that housing in Atherton is expensive, and the second is that houses in Atherton are big. When we ask the same question of La Jolla, only the first answer applies. (The answer is still no.) The hedonic view of housing markets is that the first answer is the one that is relevant, and in asking the question “Which location is more expensive” or (if temporal comparisons are desired) “Is housing more expensive now than in the past?” it is necessary to compare apples to apples. While the forbes.com methodology has its uses, it has distinct disadvantages that the hedonic method is meant to correct

The cnn.com survey explicitly recognized this and defined housing in a hedonic- that is, as a set of physical and neighborhood characteristics. The analysts that compiled the prices implicitly made the assumption of an underlying hedonic equation that governed the price of housing in each location. That equation was the hedonic function:

$$P_i = \beta_{0i} + \beta_{1i} X_1 + \beta_{2i} X_2 + \beta_{3i} X_3 + \beta_{4i} X_4 + \beta_{5i} X_5 + \beta_{6i} X_6 \quad (4.1)$$

where the i subscript is an index not of houses, but of housing markets. The i index is attached to the β coefficients because those coefficients are allowed to, indeed must, be different across those different markets. But the X variables do not have any such subscript because, in the cnn.com survey, they do not change. For each market one sets X_1 , the number of square feet, equal to 2,200;

X_2 , the number of bedrooms, to 4; X_3 the number of bathrooms, to 2.5; and so on. Then to create the ranking, the brokerage firm that gathered the data, presumably (the details are not given in the article) using a comparable sales approach to find the P_i for each market by finding units with the given characteristics and noting their sale prices (which is to say, that they did not estimate the β 's). The forbes.com ranking implicitly let the X 's be different in each market, according to each market's housing stock.

To put this in a slightly different way, according to equation (4.1) there are two reasons why housing prices differ across areas: the X 's are different and the β 's are different. The cnn.com version forced the X 's to be the same and thus concentrated our attention on the β 's. But the comparison across areas remains complicated because as previous chapters have noted, each of the β 's is, in itself a price (or a closely related function thereof). Deciding which location has the most expensive housing could be decided by which has the most expensive bedrooms, or the most expensive bathrooms. But in fact all of those attribute prices have to be combined into a single number, a *price index* for housing.

The function of a price index is to provide a *single number* that can facilitate the comparison of a large number of market prices (or changes in that market). The most familiar example of this is the Consumer Price Index, which compares the price of a market basket of goods at two different points in time. That market basket consists of quantities $q_i = q_1, q_2, q_3, \dots, q_n$, one tube of toothpaste, three loaves of bread, one twelve-speed bike, and so on. These goods have associated prices, p_i . At each point in time the price of the whole market basket of goods (all evaluated at some time period t) is calculated as $(p_1q_1 + p_2q_2 + \dots + p_kq_k)_t = P_t$. Now, price out the same market basket of goods at some new time period and call that P_s . Now if you ask the question "Is time period t more expensive than time period s ?" you can answer yes if $P_t > P_s$. The hedonic version of a price index merely uses housing characteristics rather than the more usual kinds of commodities like toothpaste.

The key to making such comparisons is to decide how much each of the individual prices should count in the adding up process-- that is, the quantities. In the case of the Consumer Price Index the Bureau of Labor Statistics (the government agency in charge of its calculation) must determine what set of commodities should go into the basket. The idea behind the Consumer Price Index is to replicate what a “typical” consumer might purchase over the course of a month, or whatever time period is deemed appropriate. Often, when making these comparisons, a ratio is taken and then put into a form that makes calculation of percentage changes easy. Letting P be the price index which compares time period t and time period s :

$$P = \frac{\sum (p_i q_i)_t}{\sum (p_i q_i)_s} \times 100 \quad (4.2)$$

When you do this, the time period in the denominator becomes the *base year* for the price index. It is the year to which all the other years are compared. For example if the adding-up for year s revealed that the market basket for year one was \$782 and that for year t it was \$805, the consumer price index P would be $(805/782) \times 100 = 102.9$, and this tells us that prices in the year t were 2.9% higher than in year s . Note that the value of the index in the base year is always 100.¹

Obviously, the denominator of the price index P is of exactly the form as the linear hedonic price regression; the quantities of consumer products (q) here become the prices of housing attributes (x) there, and the store prices of consumer products here (p) become the estimated prices

¹This straightforward use of this Laspeyere index to assess costs of living has always been the subject of criticism, most notably in the report of the Boskin Commission (1999). For purposes here the most significant of these is that the use of the constant market basket does not take into account shifts in consumption that take place due to the changing prices themselves. This *substitution bias* causes the Laspeyere index to overstate the true change in prices. Another important bias arises from lack of accounting of productivity changes over time.

of housing attributes (β) there (as long as the regression is linear).

4.2 City Indexes

In housing markets, we are interested in all kinds of comparisons. The type of price comparison discussed at the outset of this chapter was for housing prices in different metropolitan areas (at the same time). The study by Palmquist (1984) discussed in Chapter 2 provides an opportunity to do that too, but this time in a more complete fashion than the method chosen by *cnn.com*. Palmquist not only estimated a (linear) hedonic price index for the Atlanta metropolitan area but for five other areas as well² and Table 4.1 presents the hedonic parameters (once again) for Atlanta, as well as Denver, Houston, Louisville, Oklahoma City, and Seattle. A zero entry means that that particular attribute was not included in the regression for that city, a circumstance which arose because the FHA did not collect all the information for all the cities, possibly because of a dearth of houses with said attribute (as perhaps is the case for Swimming Pools in Seattle). In one sense, zero is the appropriate entry, since that is, in fact, the weight that the attribute gets in the hedonic function; on the other hand the absence of such characteristics might suggest the opposite—that the hedonic price is very high, perhaps because installation is very expensive, and no one wishes to buy it. In any case, we set the values of the quantities of characteristics to be the same as in the appraisal discussion of the previous chapter, and for convenience this is reprinted in Table 4.1. Note that in order to circumvent the issue of missing attributes in the index construction, the X^* values are zero for any attribute that is missing from any of the cities' hedonic regressions. The row labeled “Constant Quality Price” provides the estimated price—indeed, the appraisal—of a house with X^* in each metropolitan area—that is, $\sum p_i q_i$ for each city.

There are differences. Seattle's price is the highest, at over \$60,000, followed by Denver, at

²In addition to the cities discussed here and presented in Table 1, Palmquist also estimated a hedonic regression for Miami, but this is omitted from the present analysis because it uses the semi-logarithmic functional form.

just over \$50,000. Atlanta, the original example, turns out to be one of the lower priced markets, along with Houston, at just over \$35,000. In order to transform these numbers into a price index of the form above, we need to choose one city as a *base city* (against whom all the other cities are compared). Any of them can be chosen; in this case Atlanta serves that role. The last row of Table 1 shows the results. Obviously, Atlanta's index is 100; Houston's value of 98.1 indicates that a comparable house in that area costs about 2% less, while Seattle's value of 160 indicates that it costs 60% more there than Atlanta.

We noted above that the comparison of indexes depends rather importantly on the quantities. So it is with the hedonic price indexes here. The ranking of "most expensive cities" using one set of attributes can be reversed when using another set of attributes. An example of this phenomenon is also on display in Table 4.1. In the last column (labeled X**) a second set of attribute values is displayed. This set of attributes is meant to suggest a home of somewhat lower overall quality. In particular, both the size of the home and the size of the lot have been reduced. The last line of the Table provides the index number for each city. Changes in the ordering from most expensive to least expensive have taken place. In particular Denver is now the most expensive city, taking over from Seattle, whose price has taken a substantial drop in this new index. Thus, care must be taken to insure that the attribute sizes chosen are truly "representative" dwellings, although what constitutes representative will depend on the purpose of constructing the index.

Note that if the hedonic price regression is not linear, the construction of an index is still possible. As noted in chapter 2 the semilogarithmic functional form is often used in hedonic equations, and this was used in the excellent and comprehensive work by Malpezzi, Chun, and Green (1998) wherein the semilog form (see Chapter 2) is used. They estimate the regression

$$\log P_i = \beta_{0i} + \beta_{1i} X_{1i} + \dots + \beta_{ki} X_{ki} \quad (4.2)$$

for each of 373 metropolitan areas, separately for owner and renter property. The X's that were chosen represented "average" values³, and the authors show the effect of making the adjustment to the exponentiated fitted values discussed in Chapter 2. The price index becomes

$$P_i = \frac{\exp(\beta_{0i} + \sum \beta_{1j} X_j + .5s_i^2)}{\exp(\beta_{01} + \sum \beta_{1j} X_j + .5s_1^2)} \times 100 \quad (4.3)$$

These authors used a very large set of home observations, derived from the 1990 Census. They could not make their geographical units as narrowly defined as the two studies which introduced this chapter, but they used 15 different X variables in 272 metropolitan areas, and analyzed renter and owner markets separately. The most expensive owner-occupied housing was found in Honolulu, while the top price for rental housing was Stamford, Connecticut. The lowest-priced owner-occupied housing could be found in Joplin, Missouri and the lowest rents in Johnstown PA.⁴

What is a housing market? And why are there price differences across markets?

Because it reflects on the estimation of the hedonic price index, it is useful to discuss why housing prices differ across housing markets. In fact, it is helpful to take a step back and discuss the definition of a housing market. Two concepts take hold. One is the law of one price— that within a market the same commodity ought to sell for the same price. But as this volume emphasizes, the commodity called "housing" barely exists. The commodities which we must be talking about are the hedonic characteristics themselves. Second, while the physical manifestation of a market is, in most

³The X's were set at the (unweighted) mean of the local means

⁴ In the cnn.com survey, Honolulu would not have been in the top ten. One obvious explanation is that Chun, Malpezzi and Green's survey comes from the 1990 census and the website data is more recent.

contexts, pretty vague, in housing markets we can fortunately be a bit more concrete by concentrating on its spatial aspects. So: for purposes of this volume, a housing market is a contiguous geographic area for which the same hedonic parameters define housing prices.

This leaves open the exact definition of geographic area, and the literature has been flexible about this. Kim (1995) was among the first to discuss the possibility that different neighborhoods within the same city (in his case, Milwaukee) had different hedonic price models and so were, in effect, different housing markets. The law of one price presumably operates within each neighborhood. However, this level of detail has not always been available (though it is becoming more so) and in any case most researchers are, like Palmquist, content to treat metropolitan areas as single housing markets (though see Goodman and Thibodeau (2006), Coulson and McMillen (2007)). Thus hedonic functions are often estimated with databases containing observations from a single metropolitan area.

But now consider the variation in prices that occurs *across* metropolitan areas, such as that just demonstrated for the Palmquist data. The issue at hand is why some metropolitan areas have higher hedonic prices than others. Shifts in either the supply or demand curve for attributes may be the cause. The reason that Seattle had a higher value of the price index than Atlanta is because the demand for residence in the former is greater than that of the latter, and this is due to the putatively higher quality of life. While one can certainly “pay for” location in desirable areas in the form of lower wages (as in the classic model of Roback (1982)), much of the “compensating differential” is observed in differences in housing prices. What concerns us here is how that higher price will be reflected in the hedonic coefficients, under the notion that demand shifts are what makes the difference.

Again, the law of one price is important. In Table 1, consider the hedonic price of dishwashers in the various cities. The coefficients, though not identical (that would be asking a lot),

are at least within a few hundred dollars of each other, and in this context that is pretty close. And well they should be, because a dishwasher is, as housing attributes go, pretty transportable. If the hedonic price of dishwashers were higher in Seattle than in Atlanta we would surely see dishwashers shipped from the latter to the former. Developers in Seattle would note that including dishwashers in their units would pass the cost-revenue test. This would be true (in the long run at least) even if the reason for the price difference was that people in Seattle liked dishwashers more than people in Atlanta. And as Seattle's supply of dishwashers increased the price of dishwashers in the two cities would converge. The law of one price takes hold for dishwashers, and a higher hedonic price of dishwashers is probably not what explains higher prices in Seattle.

So if demand explains the fact that Seattle housing prices are higher than Atlanta housing prices this *must be reflected only in the prices of hedonic characteristics that are not easily transported*. Only in this case will the law of one (hedonic) price not take hold. One obvious instance of this is land, which by its very nature cannot be transported from one location to another, and is synonymous with location in a particular area⁵. And indeed we find that the hedonic price of land in Seattle is about six times that of Atlanta. In fact, looking over all six of the cities in Palmquist's analysis, the price indices at the bottom of the Table almost perfectly coincide with the hedonic price of land at the top. That the difference in the hedonic price of land is the primary contributor to housing price differences both across US cities (Davis and Palumbo, 2005) and within a city (Bostic, Longhofer and Redfearn 2006, for the case of Wichita) has been empirically confirmed.

So even if Seattle and Atlanta were topographically identical, the difference in the demand for land between the two cities would explain the differences in their hedonic prices for land. But neither are they topographically identical. Supply differences can be the source of price differences

⁵See also chapter 5 section 2 below on the curious hedonic interaction between land and location.

as well. Without going into a detailed analysis of that topography, it seems reasonable to assume that the supply of available land around Seattle's center is rather more constrained than in Atlanta. The combination of coastal location, as well as freshwater lakes and mountainous areas all combine to limit the amount of land available for housing construction in Seattle. Rose (1989) was an early examination of this possibility, and Saiz (2008) has a systematic exploration of topography and housing supply elasticity. Supply restrictions can be manmade as well. Zoning and other regulations may serve to constrain housing quantity and raise housing prices. It may therefore be the case that demand for Atlanta and Seattle land is the same, but that the supply differences create the difference in hedonic prices (see Katz and Rosen, 1987; McMillen and McDonald, 1993; Glaeser and Gyourko, 2003; Glaeser, Gyourko and Saks, 2005; though not all of them are hedonic-based studies.)

With dishwashers at one extreme, and land at the other, there is clearly a wide variety of ways in which supply and demand interact. A thorough investigation of hedonic differences is still in the offing but, as noted in the previous chapter, the evidence contained in MacPherson and Sirmans (2005) is congruent with the above discussion— for example, they find that the coefficient on land area is more variable than that of interior square feet.

There is a statistical approach to this issue. To facilitate the discussion, simplify the problem by considering only two cities: Atlanta and Denver. We estimate the hedonic price index for each (just as Palmquist did). If these two cities are the same market (unlikely, certainly, but just suppose) then the regression coefficients ought to be very close— though assuredly not exactly identical because of the randomness in the sample selection. The popular Chow test can be used to determine whether the hedonic parameters of the different cities can be considered statistically distinct or not⁶. If the Chow test suggests that the coefficients of the housing and

⁶The Chow test separately estimates a regression model for each of two samples and then combines them into a single sample and estimates the model once again. It then compares the sum of squared residuals from the combined model with the sum of the two SSR's when they

neighborhood variables are the same then the two city-samples can be combined into one regression. Then the price indexes would be the same. On the other hand we may still wish to entertain the oft-considered possibility that while the coefficients are in general seen to be statistically the same, the intercept terms might still be different. We could allow for that possibility in a regression model that combined the samples for both cities, and make no distinction between the two except for a dummy variable:

$$P = a_0 + a_1X_1 + a_2X_2 + \dots + a_kX_k + bD \quad (4.4)$$

where D is a binary variable equal to one if the observation is from the Denver housing market and zero if from Atlanta. The Denver coefficient then represents the *additive* difference between otherwise identical Denver and Atlanta houses. Since price indexes are *ratios* of one price to another, its calculation is a bit problematic, because the value of the Denver index (given Atlanta's value of 100) is not invariant to the underlying value of the Atlanta house. The base value for Atlanta must be decided first. Call this A; the value of Denver's index is $100 \cdot (A+b)/A$. More convenient for the purpose of calculating price indices that take ratio form is the semi-log hedonic function:

$$\log P = a_0 + a_1X_1 + a_2X_2 + \dots + a_kX_k + bD \quad (4.5)$$

in which case the price index becomes

$$P_i = \frac{\exp(\beta_0 + \sum \beta_j X_j + \gamma_i + 5s^2)}{\exp(\beta_0 + \sum \beta_j X_j + 5s^2)} \times 100 = \exp(\gamma_i) \times 100 \quad (4.6)$$

were considered individually. If the coefficients for Atlanta and Denver are "close enough" then the SSRs will be as well. If not, then the difference will be statistically significant and we can reject the hypothesis that the two cities have the same hedonic parameters.

and if γ is “small” we can interpret it as the percentage difference in price between the two cities.

In point of fact the Chow test is almost surely going to reject the hypothesis that the Denver and Atlanta are the same housing markets⁷. Thus the stratification of hedonic databases by metropolitan areas, as in the Palmquist study, is appropriate (and is the course usually followed by others, including Malpezzi, Ozanne and Thibodeau (1980), Ozanne and Thibodeau (1983), Follain and Blackley (1986), and Malpezzi, Chun and Green (2002); but see also Butler (1980).

4.2 Time indexes

We turn then to the problem of estimating housing price indexes for a given location/housing market over time. At its heart this presents no new issues. In the first instance one can gather data where the sales or appraisals occur at different points in time, and treat those points in time as if they were different markets. The (say) Atlanta market in 1990 is a distinct market from Atlanta in 2000 or any other year, and one can estimate separate hedonic regressions for each of the two markets, and proceed as above. Or one can simply estimate a single regression with dummy variables indicating the different time periods in the data. These can be over any level of time aggregation that the data will support: years, quarters and even months. The regression would then take the form:

$$\log P = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_k X_k + b_1 T_1 + b_2 T_2 + \dots + b_h T_h \quad (4.7)$$

where T_j is a binary variable indicating that the sale, or the observation of the housing price took place in time period j . There are $h+1$ distinct time periods; and a_0 is the intercept term which represents the normalized period. Setting that period's index to 100, the subsequent period's indexes are derived from the b_j in a manner similar to that described above.

⁷But as indicated above, not all hedonic attributes will have different prices across cities, if they are easily transportable, like dishwashers.

In research applications this latter method seems to be preferred over the use of separate hedonic regressions for each time period. at least in a relative sense. Researchers seem to be more willing to assume constant hedonic coefficients for the same location over time, than for different locations at the same time (see Follain and Blackley (1986); Meese and Wallace (1991), etc.)

Following on from some previous arguments, this makes some intuitive sense. It was suggested above that spatially distinct markets will have differences in supply and/or demand that would contribute to creating statistically significant differences in hedonic parameters. This is less likely to be an issue when the same market is examined at different points in time, although there are no guarantees that this would be the case. It is always helpful to use Chow tests to verify its validity.

Whether the regressions are estimated separately or not, it is often the case that smoothness is an issue. Our intuitive belief is that housing prices do proceed somewhat smoothly across time, and a common problem with the estimation of the type described above is that the indexes behave erratically. Researchers have often followed a few different strategies to impose more smoothness on the time index. All of them, obviously, involve restricting the variability of the dummy variable estimates. A common method (e.g. Blackley and Follain) is to replace the set of dummy variables with a time trend. A time trend models the variation of the time component with a single variable:

$$\log P = a_0 + a_1X_1 + a_2X_2 + \dots + a_kX_k + cW \quad (4.8)$$

This variable W takes on the value 1 for the chronologically first time period in the data base, 2 for the second, and so on. It thus assumes that the time index increases at a constant percentage (if the model is logarithmic) or amount (if it is in levels) each period. It can be seen that this *time trend* model is a specialized case of the previous approach where the coefficients in (4.7) behave according to the pattern $c^*_j = b_j$. Thus the time trend can be tested using usual hypothesis testing procedures. Note also that c can be negative, if the data indicate falling housing prices.

There are obvious problems using time trends, since the restriction on the b_j 's may not be true, and certainly will not if housing prices exhibit both increases and decreases over the sample period. One then has to balance the flexibility of the form to allow increases and decreases over time, with the desirability of "smoothness". One might alleviate this problem with the use of higher exponents of W in the regression. Including W^2 in the model would allow either a fall and subsequent rise, or the opposite. Additional cubic or even higher terms would capture any possible pattern in prices. Or one may invoke a spline or other nonparametric estimator of the time parameters, similar to those discussed in Chapter 2.

4.3 Repeat Sales

Suppose you had a database of actual sales, and moreover that it included the same house twice: a *repeat sale*. For convenience, label these observations 1 and 2 and equally conveniently assume that the sales took place in time periods one and two. The individual appraisals for these two observations are:

$$\log P_1 = \alpha_0 + \alpha_1 X_{11} + \alpha_2 X_{21} + \dots + \alpha_k X_{k1} + b_1 T_1 + e_1 \quad (4.9)$$

and

$$\log P_2 = \alpha_0 + \alpha_1 X_{12} + \alpha_2 X_{22} + \dots + \alpha_k X_{k2} + b_2 T_2 + e_2 \quad (4.10)$$

because of course the values of all the other time dummies are equal to zero. Now subtract (4.9) from (4.10). If none of the attribute sizes changed between the two sales (i.e. the X values are constant) the price difference would be:

$$\log P_2 - \log P_1 = b_2 T_2 - b_1 T_1 + e_2 - e_1 \quad (4.11)$$

Now the difference between two error terms is just another error term (although perhaps one with different properties, see below). Therefore we can write the above down in the following way:

$$\Delta \log P = b_2 T_2 - b_1 T_1 + v \quad (4.12)$$

so that the percentage change in price is just the difference in index values b_2 and b_1 plus an error term specific to that observation.

Now imagine an entire database that has such pairs of observations of home sales. That is, each house has two observations, a first sale, and a second sale. Combine each pair into a single *repeat-sales observation* (as in (4.12)), and write down a regression model (without observation indexes) of the form:

$$\Delta \log P = b_1 T_1 + b_2 T_2 + \dots + b_r T_r + v \quad (4.13)$$

where the T_j 's are no longer dummy variables in the strictest sense. Instead, they take on the value of -1 for those observations which had their first sale in that time period, and +1 for those observations which had their second sale during the time period. For any given observation the "fitted value" will be something like equation (1) with 1 and 2 being replaced by the appropriate first and second sale time periods.

Estimating the model (4.13) provides estimates of the b_j 's and, with one further modification, the sequence of b_j 's form a *multiplicative repeat sales index* as presented originally by Bailey, Muth and Nourse (1963). The modification is the usual one, that an index needs a normalization. The usual procedure, following these authors (and similar to the methods described in the previous sections) is to let the first time period be the normalization. In the logarithmic model above, this amounts to setting b_1 equal to zero. The first term in (1) then drops out, and the

rest of the b-terms are then indexes relative to the first time period.

This model, with some modifications, underlies the well-known price indexes provided by the US Office of Federal Housing Finance Administration (<http://www.fhfa.gov>) and by Freddie Mac (<http://www.freddiemac.com/finance/cmhpi>).

Bailey, Muth and Nourse compare price indexes achieved in this way to more naive price comparisons such as chain indexes. A naive chain index would work in the following way: take all of the observational pairs that were first sold in period one and then sold in period 2. The average percentage difference in this limited set of repeat sales would provide an estimate of b_2 and that estimate would, like all sample means, be unbiased (recall that b_1 is already known to be zero). You could get a similar estimate of b_3 by looking only at housing sales from time period 1 which were then repeated in time period 3, and so on. An unbiased price index could be achieved in this way. But this estimate of the price index obviously ignores information about pairs of sales that took place in periods 2 and 3 (or other time-pairs). This information can be used to provide another estimate of b_3 : you could take a weighted average of the “2-1 average” and the “3-2” average to get a second “3-1” average. This is, according to Bailey, Muth and Nourse, a less naive chain index, but one which still does not take advantage of all of the information available in the sample.

In a research program that revived interest in the repeat sales paradigm, Case and Shiller (1989) argue that the estimation of the repeat sales model above needs a further modification. They state that the longer the time gap that separates the repeat sales, the wider the range of price changes that will be observed. They propose, that is, that the prices follow a *random walk*. For any given house:

$$\log P_t = \log P_{t-1} + v_t \quad (4.14)$$

where v_t has a zero mean and a variance σ^2 . The random walk is a standard model for asset prices

because it implies that future stock movements are in effect unpredictable, and that therefore it is a natural outcome of perfect arbitrage in asset markets. While we wish to avoid the topic of whether housing prices behave efficiently in this way (but see Case and Shiller (1989)) the possibility that prices do follow a random walk has an impact on the estimation of the repeat sales model. For note that

$$\log P_t = \log P_{t-r} + v_t + v_{t-1} + \dots + v_{t-r+1} \quad (4.15)$$

The log difference taken over k periods is the sum of k shocks to the price of the housing unit. Each time period brings its own unique circumstances to bear on the price of the unit and each period will bump the price one way or the another. So while these shocks on average will have no effect on the percentage change in the house price, they have a substantial effect on the variance. We assume that the shocks are independent of each other, and so the variance of the sum is equal to the sum of the variances: that is

$$\text{Var}(\log P_t - \log P_{t-r}) = r\sigma^2 \quad (4.16)$$

in which case we see that the regression in (4.13) is heteroskedastic (as described in Chapter 2). The wider the time gap, the bigger the variance⁸.

For purposes of estimation, in addition to this source of variance Case and Shiller assume that there is a time-invariant source of variance that arises from the idiosyncracies of the particular property. Thus the total variance is $\sigma_1^2 + k\sigma_2^2$, where the first term is the “idiosyncratic variance) and the second is the variance related to the time between sales. Case and Shiller propose the following estimation procedure. They estimate the basic repeat sales model using ordinary least squares, and then regress the squared residual from that regression on a constant term and the measure of time between sales. The coefficient of the latter is taken to be the estimate of the

⁸Dreiman and Pennington-Cross (2004) suggest that r might depend on a variety of factors that vary from submarket to submarket.

variance term σ_2^2 while the intercept term is an estimate of σ_1^2 . They then deflate each observation by the square root of the sum above, and reestimate the whole model⁹. The FHFA indexes use this correction.

Examination of the repeat sales regression reveals the obvious advantage of this method, which is that the *actual attributes of the property are not among the regressors*, and so it is unnecessary to estimate attribute prices. (In that sense the repeat sales regression is not even a hedonic model.) This is especially significant since both observed and unobserved attributes are eliminated from consideration. Recall from Chapter 2 that one of the difficulties faced by hedonic researchers is that of unobserved attributes, the omission of which from the hedonic regression can cause the included parameter estimates to be biased. The differencing operation which takes place in the repeat sales model removes all attribute levels and so the bias is eliminated from the parameter estimates. If the goal of the investigator is not to estimate attribute prices but merely to derive constant quality price indexes, then the repeat sales model has a great deal of merit. Since the publication of Case and Shiller (1989) and the availability of sufficiently rich databases, use of this model has exploded.

The repeat-sales model is not, however, without its own faults. The major premise of the model is that the attributes do not change between sales, and that the coefficients of those attributes do not change either. This pair of assumptions is what allows the cancellation to take place. But if these assumptions are not true then some modifications of the model are required.

There is one attribute that is clearly not constant over the inter-sales period, and that is the age of the dwelling. The problem that this causes can be most starkly represented when the coefficient of age is allowed to change over time in the most flexible way possible (see also the

⁹This is an example of the Weighted, or Generalized, Least Squares estimator, found in any econometrics textbook.

discussion in Chapter 3). In this case, each age level would have its own binary variable. Let $A_j = A_1 \dots A_p$ represent binary variables that equal one if the age of the dwelling is j . Removing all of the other variables from the hedonic, the repeat sales model gives

$$\Delta \log P = \alpha_2 A_2 - \alpha_1 A_1 + b_2 T_2 - b_1 T_1 + \nu \quad (4.17)$$

for any given repeat sale, and the full regression model is

$$\Delta \log P = \alpha_1 A_1 + \dots + \alpha_p A_p + b_1 T_1 + b_2 T_2 + \dots + b_k T_k + \nu \quad (4.18)$$

where the value of the A variables is specified much like that of the T variables: $A_j = -1$ at the time of the first sale and $A_j = 1$ at the time of the second sale. The problem here is clear from the discussion of age, time and vintage in the previous section. In that model, the level of price could not be a function of both age and vintage if the observations were all at the same point in time, and could not be a function of age, vintage and time even if observations come at different time periods. Here, observations come at different time periods but are differenced, hence vintage (which is *per force* an unchanging variable) necessarily vanishes from the model, and the age variables and time variables are perfectly correlated.

As in the previous case, the resolution of the problem is to impose some kind of restrictive functional form on either the time dummy variables or the age dummy variables. Since the purpose of the repeat sales model is to create time indexes with as little restriction as possible, it is the age variable that must be constrained. It would seem that the standard thing to do is let age enter the hedonic linearly, in which case the repeat sales regression would have the form

$$\Delta \log P = \alpha_1 \Delta Age + b_1 T_1 + b_2 T_2 + \dots + b_k T_k + \nu \quad (4.19)$$

and while the variable Δage is of course the length of time between sales, it is not perfectly correlated with the set of T dummy variables in the regression.

What is always true of age might also be true with almost any other structural or neighborhood characteristic. Remodeling of housing occurs, and neighborhood characteristics can certainly change over time. Case and Quigley (1991) construct a model which attempts to deal with both the possibility that the characteristic size might change but also the possibility that its coefficient might move over time as well. The latter problem is dealt with first. The model they have in mind is (again suppressing the observation index i):

$$\log P_t = \beta_0 + (\alpha_1 + \beta_1 t)x_{1t} + \dots + (\alpha_k + \beta_k t)x_{kt} + e_t \quad (4.20)$$

thus the parameter associated with each attribute is decomposed into two parts: a base value α , and a part that grows at rate β .

Note that from a practical standpoint, this model requires that each attribute is entered twice into the regression: once on its own (and having coefficient α) and once multiplied by the time variable t (and having coefficient β). Because of this, repeat sales are not required, only that the observable sales take place at different points in time (that there be different t 's). For the observations that *are* repeat sales, the counterpart to (4.11) is now

$$\log P_t - \log P_{t-r} = \beta_1 r x_{1t} + \dots + \beta_k r x_{kt} + e_t - e_r \quad (4.21)$$

The variables in the repeat sales version of (4.20) would then have to be constructed as the product of the time span between sales (r) and the attribute level (x). The regression coefficients would then be (log) time trends for each of the attribute prices.

While this model has its advantages, it essentially reverts to describing house prices as linear trends. Note that this trend, the per-time-unit change in the price, can be found by setting $r=1$

$$E(\log P_t - \log P_{t-1}) = \beta_1 x_{1t} + \dots + \beta_k x_{kt} \quad (4.22)$$

so that each housing unit has its own trend which depends on the observed attributes. One has by this time lost the ability to use the repeat-sales regression (like 4.13) to construct a temporal price index. This is because (4.12) assumes that the changes in housing prices over time are the same for all houses (that is, that the hedonic parameters are constant), and (4.20) denies this assumption. If one is to use (4.22) to construct an index, it is necessary to create and use a representative list of X's as we did when creating the city price indexes above.

Case and Quigley go further, and modify the model allow the trend to vary over time. While they do not write their model in this way, it can be shown that their model, to the extent that the database contains repeat sales, is precisely the repeat sales methodology applied to attributes themselves (see also Case, Pollakowski and Wachter (1991)). That is, the hedonic equation for a given observation is:

$$\log P_t = \beta_0 + (\alpha_1 + \beta_{1t})x_1 + \dots + (\alpha_k + \beta_{kt})x_k + e_t \quad (4.23)$$

so that, on subtracting the repeat sale at time period t-r we get, again for an individual observation:

$$\log P_t - \log P_{t-r} = (\beta_{1t} - \beta_{1r})x_{1t} + \dots + (\beta_{kt} - \beta_{kr})x_{kt} + e_t - e_r \quad (4.24)$$

For the sample as a whole, proceeding as above, we get the regression model

$$\log P_t - \log P_s = \sum_{i=1}^T \sum_{j=1}^k T_{ij}(\beta_{ij})x_j \quad (4.25)$$

where, i indexes the time period, j indexes the attribute and, in a fashion similar to the original Bailey, Muth and Nourse models:

$$\begin{aligned} T_{ij} &= \mathbf{1} \quad \text{if } i = t \\ &= -\mathbf{1} \quad \text{if } i = s \end{aligned}$$

Note that this allows the price index to have any pattern over time; it has all of the

flexibility-- indeed it nests-- all of the previous models. Its one drawback, the drawback of all highly flexible models, is that the researcher must estimate a large number of parameters. A large sample is required, obviously one that has a large number of (repeat) sales, at various points in time, with a wide variety of attribute levels at each point in time. Case and Quigley (1991) only use six attributes in their estimation, and Case, Pollakowski and Wachter (1991) are explicit in discussing the limitations that were necessary for the model's price indexes to behave coherently.

The situation is complicated even more by an issue discussed previously, that the "simple" repeat sales model assumes that there are no changes in the attribute level, x , between the sales. If this is not true, and we fail to take such changes into account, then we may overstate price changes (if x 's with positive value rise). This is not merely a question of changes in the physical structure of a unit, such as an addition which increases the number of square feet. These might be relatively rare. But perhaps not so rare are changes in the neighborhood, which can get better or worse rather quickly, and affect a larger number of local transactions because of developments in the surrounding environment.

Suppose that, as above, sales take place at both time t and $t-r$. The hedonic for the first sale is

$$\log P_{t-r} = \beta_0 + (\alpha_1 + \beta_{1t-r})x_{1t-r} + \dots + (\alpha_k + \beta_{kt-r})x_{kt-r} + e_{t-r} \quad (4.28)$$

and the second

$$\log P_t = \beta_0 + (\alpha_1 + \beta_{1t})x_{1t} + \dots + (\alpha_k + \beta_{kt})x_{kt} + e_t \quad (4.29)$$

Then subtracting the one hedonic as before, but now taking into account changes in x , yields the individual hedonic

$$\begin{aligned} \log P_t - \log P_{t-\tau} &= \alpha_1(x_{1t} - x_{1t-\tau}) + \beta_{1t}x_{1t} - \beta_{1t-\tau}x_{1t-\tau} \\ &+ \dots + \alpha_k(x_{kt} - x_{kt-\tau}) + \beta_{kt}x_{kt} - \beta_{kt-\tau}x_{kt-\tau} + e_t - e_{t-\tau} \end{aligned} \quad (4.30)$$

and the hedonic regression model becomes

$$\log P_t - \log P_s = \sum_{j=1}^k \alpha_j (x_{jt} - x_{jt-\tau}) + \sum_{i=1}^T \sum_{j=1}^k T_{ij} (\beta_{ij}) x_{jt} \quad (4.31)$$

At this point one is faced with a choice of samples. Case and Quigley create three data sets: (a) those units with no repeat sales, (b) those with repeat sales and no changes in attributes, and (c) those with repeat sales and changes in attributes. They then note the important fact that the α parameters in the first and third groups ought to be the same; and that the β parameters in the second and third ought to be the same as well. (This is the case with and without the possibility that the trends are the same across attributes.) Therefore they simultaneously estimate all three and impose the parameter equality thereby suggested¹⁰. They report gains in accuracy from estimating everything together in the manner suggested by 4.31.

This is of interest because it is not entirely clear what the “correct” set of data is. Combining one-time sales with repeat sales is not necessarily a good idea. A big advantage of using repeat-sales, recall, is that by differencing the data to form a repeat sales observation, the effects of unobserved housing characteristics are purged. Not accounting for these may lead to bias in the estimates of the parameters, and so combining the two data sets may have this problem as well. On the other hand, using only repeat sales, even after the adjustments for changing parameters, and changing attributes

¹⁰B. Case and Quigley also correct for heteroskedasticity, but not, evidently in the way that K. Case and Shiller did. Repeat sales observations are allowed to have a different variance than the “ordinary” hedonic observations.

(as in Case and Quigley) still has the disadvantage of throwing away an awful lot of observations.

So one has a choice of samples, all of which involve various tradeoffs of these types. One can obviously estimate an ordinary hedonic (with or without time dummies); one can limit this hedonic to those without repeat sales or merely treat the repeat sales as ordinary observations in the database. One can estimate a repeat sales model either in the classic Bailey-Muth-Nourse fashion or in the manner of (2), and one could exclude repeat sales that exhibited changes in the attributes (or not). Or one could estimate a model just above, using only repeat sales that *did* exhibit changes in the attributes. In the latter couple of cases, one is faced with the most severe fault of the repeat sales paradigm, which is that in limiting the regression data to that associated with repeat sales, the researcher is ignoring a lot of information about the housing market.

Case, Pollakowski, and Wachter (1991) try to come to grips with these tradeoffs. They estimate fourteen different models of time index construction, from the simplest Bailey, Muth and Nourse model, to the most complex hybridization. They do not end up in favor of the hybridization favored by Case and Quigley over simpler repeat sales models--- adding one-time sales to the model may add enough bias that the increased efficiency is possibly not worth the tradeoff.

Moreover, Meese and Wallace discount the hybrid model (1997) suggested by Case and Quigley because of the possibility that the hedonic parameters may be different in repeat samples than in ordinary hedonic observations. That is to say, they wish to question the joint estimation of parameters from hedonic and repeat sales because of the possibility of unobservable differences in the houses that are subject to repeat sales and those that are only sold once. This is because the housing units which are found in the repeat sales database may be different than those found in the single sales database. (Clapp and Giacotto (1999) make a similar point, that houses with only a year or two between sales have a higher appreciation rates, presumably because there are unobservable improvements to the property.) Meese and Wallace use a Chow test as described above, permitting

separate parameters to be estimated for repeaters and one time sales. They do find a difference for their sample, and if that is the case, there is no justification for any joint treatment of repeat sales and ordinary hedonic observations.

To summarize, in constructing temporal price indexes, repeat sales estimators have one great advantage: they require no information about housing attributes. This advantage should not be understated. However, using *only* repeat sales information creates limitations on the available sample, and is only useful for assessing the time index of housing prices. Combining repeat sales with ordinary one-time sales alleviates those issues, but creates other ones, since there are now many options with respect to models and sample selection. The excellent comparison in Case (2004) of many of these models nevertheless reveals a broad consistency in the temporal patterns of prices¹¹.

¹¹Interestingly, Case (2004) includes in his comparison a price index based on the year to year observations of a single housing unit. This extreme reduction of the comparable sales methodology, not surprisingly, produces a absurdly highly variation in the “index” compared to any of the other methods.

TABLE 4.1

Palmquist (1983) estimates of hedonic price indexes for six US cities

ATTRIBUTE	Atlanta	Denver	Houston	Louisville	Ok. City	Seattle	Value for X*	Value of X**
Intercept	-9337.32	4398.511	-12156.8	1116.21	2901.192	-9526.05	1	1
Lot Area (square feet)	0.0813	0.1474	0.0998	0.0745	0.1423	0.6542	40000	25000
Improved Area (square feet)	15.0576	12.7203	12.7237	8.4252	8.6116	17.921	1400	1000
Improved Area ²	-0.0022	-0.0019	-0.0002	-0.0023	0.0007	-0.0032	1960000	1000000
Number of Baths	1821.32	1881.861	477.7357	3611.45	1169.399	2527.32	2	2
Year Built	134.4473	79.34	111.402	71.27	106.004	101.9034	70	70
Number of stalls in Garage	1451.094	21989.28	1838.58	1602.43	1694.6060	1319.142	2	2
Number of stalls in Carport	1198.081	601.4742	682.5717	999.5843	1097.116	483.6459	0	0
=1 if garaged is detached	-1006.91	-820.9986	-739.4174	-409.3972	-1277.08	-479.62	0	0
=1 if wiring is underground	710.0944	510.15	1239.945	2156.105	449.7995	672.2081	1	1
=1 if dishwasher	1710.118	984.5379	1153.738	2027.138	1028.8940	1006.522	1	1
=1 if garbage disposal	292.5529	473.8454	783.4335	1214.163	866.3541	696.6563	1	1
=1 if central air conditioning	1937.391	0	1998.0340	2113.566	1606.441	0	1	1
=1 if wall air conditioning	604.6657	0	984.1632	642.5249	285.40880	0	0	0
=1 if ceiling fan	344.714	570.0075	-165.0138	977.5475	560.0915	300.8057	0	0
=1 if sold in 1976	-1114.5	-2432.459	-1758.73	-1179.69	-1616.08	-2207.6	0	0
=1 if "excellent condition"	1007.502	1434.456	759.2975	384.2958	1084.787	1243.15	1	1
=1 if "fair condition"	-2227.37	-2095.13	-1352.85	-2538.34	-1042.07	-1626.36	0	0
=1 if "poor condition"	0	-4316.13	0	-3390.74	-8880.12	153.1606	0	0
=1 if brick or stone exterior	622.333	979.6494	1568.272	2390.842	1241.053	3981.132	0	0
=1 if full basement	1852.194	2229.443	0	3219.733	0	3712.41	1	1
=1 if partial basement	1108.292	2218.805	0	2201.489	0	2748.483	0	0

=1 if fireplace	1114.569	2118.643	2418.986	1604.151	2416.365	1334.085	1	1
=1 if swimming pool	3274.725	0	0	0	3426.925	0	1	1
level of air pollution	-45.47	-26.0403	-11.8616	-15987	-0.2232	-8.865	0	0
median age in census tract	-58.1812	49.0941	119.2348	47.7201	2.8702	-108.976	36	36
median family income in census tract	0.0788	0.1655	-0.0044	0.0249	-0.0976	0.3854	25,000	25,000
% of workers in tract with blue-collar jobs	-52.1812	-15.0316	27.0144	-44.1273	-51.5020	-76.8352	45	45
% of houses in tract with new occupants (< 5 yrs)	-32.4515	-61.5007	-30.7873	5.9894	-2.7746	-30.8822	14	14
% of tract population that is non-white	-1516.5	-4465.67	-2455.11	-5561.16	-3412.13	-6155.91	0	0
% of tract population over 24 that is HS graduate	1.2341	0.3271	0.4575	0.9166	0.2563	-0.3471	68	68
% of structures with >1 person per room	35.9097	64.3941	80.8062	-16.2552	-39.3587	199.9203	0	0
number of work destinations per square mile in tract	16.6915	13.9396	26.8967	3.7791	-7.0332	8.7971	0	0
Constant Quality Price	\$35,695.53	50235.08	35038.77	40020.18	41373.61	60748.87		
Index (X*) (Atlanta=100)	100.0	140.32	98.1	112.1	115.9	170.2		
Index (X**) (Atlanta =100)	100.00	145.86	93.27	123.86	114.02	144.31		

