11

Empirical Examples

Introduction

11.1 The purpose of this chapter is to provide additional empirical examples dealing with the construction of house price indices based on the methods that were outlined in Chapters 5-9. These are broadly defined as follows: measures of central tendency (mean or median), hedonic regression methods, repeat sales methods, and methods based on appraisal data. The following three sections of this chapter illustrate how the first three classes of methods can be implemented on very small data sets. Hopefully, working through these simple examples will enable readers to more readily follow the rather terse algebraic descriptions of the various methods that were provided in Chapters 5-9.

11.2 The following section also illustrates various methods that can be used to aggregate regional house price indices into overall house price indices. This topic was not covered in any detail in other chapters of this Handbook.

Central Tendency Methods and Stratification Methods

11.3 Central price tendency estimates, such as mean and median prices, for constructing an RPPI are among the least data intensive of all the methods currently available to compilers. The basic mean or median methods only need the selling prices of the properties in a given location to build a price index. Thus location information will be required. In addition, it is usual to stratify by the type of dwelling unit and if this is the case, then information on the type of dwelling unit will also be required.

11.4 As a first exercise, an index is constructed using the mean price. It consists in calculating the simple average of the observed prices for a sample of houses in a given period and for a given geographical area. The indicator, which can be expressed in monetary terms or in index form, is then measured simply as the change (in per cent usually) of the average price of the sampled units between two periods. ([1])

11.5 It is important that the sample of houses drawn for calculating the price indicator be representative of the target universe. Therefore some data editing may be required, the extent of which will depend on the instructions that the data provider received from the compiler and his willingness and ability to deliver the data according to the compiler's stated criteria. ([2]) For example, the sample of prices initially collected may include certain property types, such as agricultural land, commercial properties, and units found in multi-unit dwellings, which are considered outside the scope of the intended index. If this is the case, then these observations need to be excluded from the sample when measuring price trends for specific types of properties. Outliers should also be identified and removed from the sample if it is believed that they may skew or distort in any other way the outcome.

11.6 A simple numerical example using 5 and 7 price observations respectively for periods 1 and 2 ([3]) will illustrate the approach used for measuring the progression of the simple mean of house prices for a given geographical area, usually for a city or other well-defined area. ([4])

Period 1 house prices and mean

$$(350K + 352K + 378K + 366K + 402K)/5 = 370K$$

Period 2 house prices and mean

$$(360K + 350K + 382K + 395K + 380K + 400K + 450K)$$
$$/7 = 388K$$

Once the average prices for each period, e.g., a month, a quarter or a year, are obtained, it is then straightforward to calculate the period-to-period progression (typically in per cent) between \$370K and \$388K. For instance, in this specific example, average house prices have increased about 5 % over both periods.

11.7 The presence of outliers is mitigated when the median price of properties in the sample is used instead of the mean price. For instance, if one or more very expensive houses are sold in a given period, the resulting average price will likely not be typical of houses that on the market at that time. As was discussed in Chapter 4, the median approach does not however completely control for period-to-period compositional shifts in the sample of houses sold. In spite of this shortcoming the median is nevertheless a very popular residential property price indicator mainly because it is simple to compile and is not very data intensive, thus resulting in a timely indicator. Moreover, its interpretation is straightforward.

11.8 Based on the same data used for calculating the mean, the median prices from the example samples for periods 1 and 2 are found to be respectively \$366K and \$382K. Consequently, the median house price has increased 4.4 % over these two periods.

11.9 The above exercise is repeated below but with a more extensive dataset containing 5787 sampled price observations for single-family houses drawn from actual

([1]) Regardless of the form used, expressed either in terms of values or indices, the per cent change will be the same.
([2]) Of course the particular circumstances will dictate the extent of the data cleaning. If the principal user is also managing the collection of information, then the survey will be tailored to his or her needs and the extent of the cleaning will likely be less extensive.
([3]) Since the number of transactions will likely vary from period to period, the number of price observations in the sample for each period will also vary.
([4]) Note that most central tendency measures of house prices when published do not typically include indicators of statistical quality such as the coefficient of variation or standard deviation.

transactions over many years for a small municipality.([5]) Some descriptive statistics are presented in Table 11.1. Note that in this particular case, the mean price of houses sold in any year is always higher than the corresponding median. For instance, in 2002 the mean is $249 702 against 236 000 for the median; in 2008 the mean is $365 195 against $340 600 for the median. Since for any given year the sample is characterized by the sale of some higher priced units, this result is to be expected. In fact, the distribution of prices is right-skewed with a skewness coefficient ranging from 1.44 to 1.87 over the various years.([6]) Chart 11.1 illustrates the distribution of prices in 2008 for the houses that were sold that year. A similar graph constructed for the remaining years for this example yields similar price distributions.([7])

11.10 As for the annual per cent changes, they vary according to the measure of central tendency that is used here.([8]) In some years, the difference in the result between the median and mean can be quite small. For instance, in 2002 the difference is only one tenth of a percentage point (8.2 % vs. 8.1 %) with mean recording a slightly higher increase. In other years, such as in 2008, the difference is more pronounced such as in 2008 when the annual change measured using the median price increased by 6.8 % compared to an increase in the mean price of 5.2 %.

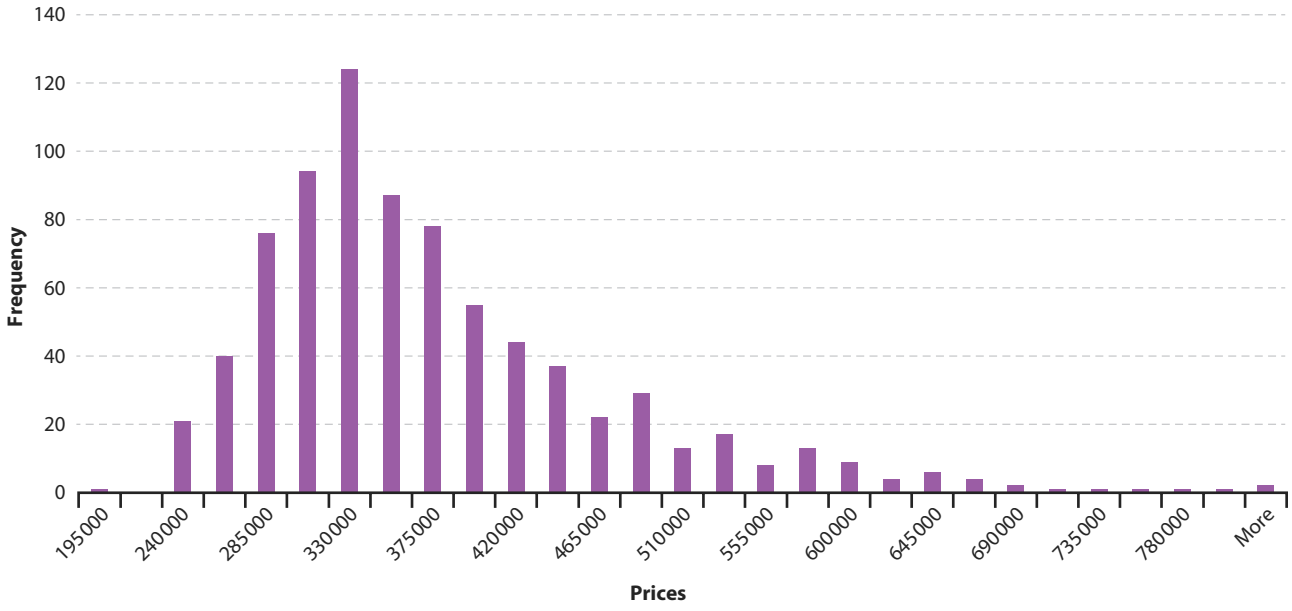
([5]) Note that the required data is obtained for calculating either the median or mean prices; the steps involved are quite simple. Most statistical software packages can do the entire exercise quite rapidly with little intervention from the compiler.
([6]) Skewness is a measure of the asymmetry of a distribution. When the degree of skewness is zero this means that the distribution is symmetric around its mean. A positive skew means that a relatively high number of observations from the sample is concentrated on the left of the centre point and vice versa.
([7]) With these particular data, the mean was always greater than the corresponding median. This result need not always hold, particularly with very small samples.
([8]) Typically, the mean price will be higher than the corresponding median price. However, when mean and median indices are formed, there is no presumption that the mean index will increase more rapidly than the median index.

Table 11.1. Means, Medians, Percent Changes, Standard Deviations, and Skewness

	2002	2003	2004	2005	2006	2007	2008
Observations	777	804	894	808	834	874	796
Standard deviation	64 130	62 042	73 405	76 432	84 587	96 559	96 152
Skewness	1.63	1.51	1.71	1.87	1.58	1.46	1.44
Mean ($)	249 702	270 174	290 686	299 087	315 099	347 009	365 195
Per cent change		**8.2%**	**7.6%**	**2.9%**	**5.4%**	**10.1%**	**5.2%**
Median ($)	236 000	255 000	273 000	280 000	292 000	319 000	340 600
Per cent change		**8.1%**	**7.1%**	**2.6%**	**4.3%**	**9.2%**	**6.8%**

Source: Authors' calculations based on MLS® data for a Canadian city

Chart 11.1: Distribution of House Prices in 2008



Source: Authors' calculations based on MLS® data for a Canadian city

11.11 As is well known, location plays an important role in the determination of not only the level of house prices but also in their behaviour over time. Therefore, to improve the reliability of the indicator, a stratified or mix-adjustment approach is routinely recommended, provided of course that the information for segmenting the market (or sample of transactions) is readily available. Geographical stratification has the advantage of reducing the effects of period-to-period compositional shifts in the housing units that characterize the simple mean and median methods. A popular approach to segmenting the housing market is to group houses according to geographical area, thus ensuring a certain degree of homogeneity of the units found within the strata; other locational effects on house prices are also minimized by this method. Stratification can also benefit users by providing them with additional house price indicators for various sub-markets, such as by neighbourhood or type of house. Goodman and Thibodeau (2003) add that there is also a practical reason for grouping house by location in that geographic variables are almost always included in databases on housing transactions. This information should, when available, be leveraged since stratification makes efficient use of these data.

11.12 Some countries, such as Australia (Branson 2006), have taken advantage of the traditionally strong relationship between price and location that typifies residential real estate by stratifying the sample of properties according to geographical area or other submarket structures. This can be a viable, albeit imperfect, alternative (or compromise solution) for measuring constant quality price change in the absence of the resources and the data needed to apply some of the more sophisticated methods for constructing an RPPI such as hedonic regressions. In fact, Prasad and Richards (2008) construct a measure of median house prices for six Australian capital cities where the markets are stratified according to long-term price movements. Using a database of over 3 million observations, the authors find that their approach to measuring changes in house prices, (i.e., using the median approach but stratified by zone as defined by long term price trends), will generate results that are comparable to those using more sophisticated and data intensive methods such as hedonics or repeat sales.

11.13 Stratifying by geography thus likely ensures that the cluster of observations within each group (or stratum) is more homogeneous than observations from the entire population. Stratification can be extended to include, in addition to geography, other price determining factors such as house type and/or number of bedrooms. Grouping of houses by geography and other criteria will result in a sample of even more homogeneous properties, which is a desirable outcome for mitigating fluctuations in the index that are caused by compositional shifts in the sample that occur over time. One potential drawback however with this approach is that the compiler must be aware that a too finely defined stratum can sometimes generate a thin sample of transactions in any given period, thus resulting in some sampling bias. The objective is therefore to design the individual strata in such a way that the homogeneity of price determining characteristics is balanced against a sample size that is sufficiently robust to yield a reliable and representative measure of changes in house prices.

11.14 As previously mentioned, the construction of sub-market (or stratum) price indices that are then aggregated to the level of the market of interest will often use median prices in practice. Constructing a mixed-adjusted price index consists in first defining the stratum. The second step is to calculate the median price for houses transacted within the stratum for the period in question. Thirdly, the median prices for all sub-markets must be weighted together into an aggregate price measure for the market under study, which likely will be a city or even the country as a whole.

11.15 The following provides a simple example of the procedure and steps involved with calculating a mixed-adjustment price index for residential properties.([9])

- Step 1: Define the stratum. For the purpose of this exercise, the stratum is a geographical subdivision of a city such as the west-zone or centre town. There is no strict rule for delineating the stratum in question but geography appears to be a popular and obvious choice which can, if data permitting, be combined with other housing features such as by house type or according to number of bedrooms in order to narrow the stratum.([10])
- Step 2: Calculate the median price for a stratum such as a neighbourhood for the relevant period (month or quarter). It is assumed that the median will be the representative price of all sales in that stratum. However, the mean price could alternatively be used. Repeat this step for future periods.
- Step 3: Estimate the "average" price of houses sold for a given period by calculating a sales weighted median of the neighbourhood or stratum prices.([11])

11.16 Suppose that data on house sales for two periods (0 and 1) and three geographical regions or neighbourhoods (A, B and C) have been collected. Suppose prices are measured in thousands of dollars and that for region A in period 0, there were 4 sales with prices 290, 450, 250 and 310. Thus, the mean price for this period was 325, the median price was 300 (the arithmetic average of the two middle prices 290 and 310) and the total expenditure was 1300. For period 1, region A had 5 sales of 300, 500, 250, 400 and 275. Thus, the mean and median price for this period was 345 and 300 respectively and the total period 1 expenditure in region A was 1725. For region B, there was only one sale in each period:

([9]) This example is loosely based on an example in McDonald and Smith (2009).
([10]) This example uses the neighbourhood as the sub-stratum but in reality it can be any geographical area for which the compiler is confident that a sufficiently large enough sample of transactions is available today and in the future to generate a reliable representative price.
([11]) This is assuming that the compiler is using sales as the basis for the weighting.

500 in period 0 and 400 in period 1. Thus, the mean and median price in period 0 for region B was 500, which was also equal to expenditure in this period. The mean and median price in period 1 for region B was 400, which was also equal to expenditure in this period. For region C, there were 3 sales in each period. For period 0, the sales were equal to 200, 300 and 175 and so the median price was 200, the mean price was 225 and expenditure was 675. For period 1, the sales in region C were equal to 250, 350 and 225 and so the median price was 250, the mean price was 275 and expenditure was 825. These are the basic data for the example.

11.17 Suppose that the *median price* in each region corresponds to houses of comparable quality over the two periods being compared. Since it is desirable to have price times volume equal to expenditure in each period for each region, once a constant quality price concept has been chosen, the corresponding volume should equal expenditures divided by price. Using the median price in each region as a constant quality price for each time period leads to the data on expenditures (the v^t), prices (the p^t) and volumes or implied quantities $q^t = v^t / p^t$ that are listed in Table 11.2 below.

Table 11.2. Regional Expenditures, Prices and Volumes (Implicit Quantities) Using Median Prices as the Regional Prices

Period	v_A^t	v_B^t	v_C^t	p_A^t	p_B^t	p_C^t	q_A^t	q_B^t	q_C^t
0	1300	500	675	300	500	200	4.333	1.000	3.375
1	1725	400	825	300	400	250	5.750	1.000	3.300

Source: Authors' calculations based on MLS® data for a Canadian city

Note that the regional price indices for period 1 are equal to $p_A^1 / p_A^0 = 1.0$, $p_B^1 / p_B^0 = 0.80$, and $p_C^1 / p_C^0 = 1.25$ for regions A, B and C respectively. Thus there are widely differing house price inflation rates in the three regions.

11.18 At this point, we can apply normal index number theory to the problem of aggregating up the regional price movements into an overall house price inflation rate. For example, *Laspeyres* and *Paasche* overall price indices, P_L and P_P, for period 1 can be constructed. The formulae for these indices are as follows:

$$P_L \equiv [p_A^1 q_A^0 + p_B^1 q_B^0 + p_C^1 q_C^0] / [p_A^0 q_A^0 + p_B^0 q_B^0 + p_C^0 q_C^0] \quad (11.1)$$

$$P_P \equiv [p_A^1 q_A^1 + p_B^1 q_B^1 + p_C^1 q_C^1] / [p_A^0 q_A^1 + p_B^0 q_B^1 + p_C^0 q_C^1] \quad (11.2)$$

11.19 The CPI Manual (2004) recommends the construction of *superlative indices* if price and quantity data are available for the periods under consideration, as they are in the present situation. Two such superlative indices are the *Fisher ideal index* P_F and the *Törnqvist-Theil index* P_T, defined as follows for the period 1 overall indices:

$$P_F \equiv [P_L P_P]^{1/2} \quad (11.3)$$

$$P_T \equiv \exp[0.5(s_A^0 + s_A^1) \ln(p_A^1 / p_A^0)$$

$$P_T \equiv \exp[0.5(s_A^0 + s_A^1) \ln(p_A^1 / p_A^0) + 0.5(s_B^0 + s_B^1)$$

$$\ln(p_B^1 / p_B^0) + 0.5(s_C^0 + s_C^1) \ln(p_C^1 / p_C^0)] \quad (11.4)$$

where *the period t shares of sales* in regions A, B and C are given by $s_A^t \equiv v_A^t / (v_A^t + v_B^t + v_C^t)$, $s_B^t \equiv v_B^t / (v_A^t + v_B^t + v_C^t)$ and $s_C^t \equiv v_C^t / (v_A^t + v_B^t + v_C^t)$, respectively. Note that the Fisher (1922) index P_F is equal to the geometric average of the Laspeyres and Paasche indices, P_L and P_P and that the Törnqvist-Theil index P_T is equal to a share weighted geometric average of the regional price indices, p_A^1 / p_A^0, p_B^1 / p_B^0 and p_C^1 / p_C^0, where the weights are the arithmetic averages of the period 0 expenditure shares, s_A^0, s_B^0 and s_C^0, and the period 1 expenditure shares, s_A^1, s_B^1 and s_C^1.

11.20 The results for the four indices defined by (11.1)-(11.4) are listed in Table 11.3 below. It should be noted that the two superlative indices, P_F and P_T, are fairly close to each other while the Laspeyres index P_L lies above these superlative indices and the Paasche index P_P lies below them. This is a typical empirical result.

11.21 Organizations that compile residential property price indices tend to use somewhat different formulas when aggregating over regions. A common form of aggregation is to use a *weighted* average of the regional price indices to form an overall index, using the sales weights of period 0 (or some average of sales weights that pertain to periods prior to period 0). Denote the share weighted index that uses the sales weights of period 0 by P_0 and the share weighted index that uses the sales weights of period 1 by P_1. The period 1 values [12] for the indices P_0, P_1 and the arithmetic average of P_0 and P_1, denoted by P_A, are defined as follows:

$$P_0 \equiv s_A^0 (p_A^1 / p_A^0) + s_B^0 (p_B^1 / p_B^0) + s_C^0 (p_C^1 / p_C^0) \quad (11.5)$$

$$P_1 \equiv s_A^1 (p_A^1 / p_A^0) + s_B^1 (p_B^1 / p_B^0) + s_C^1 (p_C^1 / p_C^0) \quad (11.6)$$

$$P_A \equiv 0.5 P_0 + 0.5 P_1 \quad (11.7)$$

[12] The period 0 values for all of the indices defined in this section are set equal to 1.

The above three indices are also listed in Table 11.3.([13]) It can be seen that P_0 is equal to P_L and is about 0.26 percentage points above the Fisher index P_F in period 1, while P_1 is about 1.77 percentage points above P_F. This result is not unexpected; the indices P_0 and P_1 do not generally closely approximate superlative indices and so their use is not recommended.

([13]) Fisher (1922; 466) showed that P_0 defined by (11.5) is equal to the Laspeyres index P_L defined by (11.1). Fisher also attributed the index P_1 defined by (11.6) to Palgrave.

Table 11.3. Overall House Price Indices using Median Prices and Alternative Formulae to Aggregate over Regions A, B and C

Period	P_F	P_T	P_L	P_P	P_0	P_1	P_A	P_{GL}	P_{GP}
0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
1	1.02515	1.02425	1.02778	1.02253	1.02778	1.04280	1.03529	1.01590	1.03267

Source: Authors' calculations based on MLS® data for a Canadian city

11.22 Two additional indices are listed in Table 11.3: the *geometric Laspeyres and Paasche price indices*, P_{GL} and P_{GP}. The period 1 values for these indices are defined as follows:

$$P_{GL} \equiv \exp[s_A^0 \ln(p_A^1 / p_A^0) + s_B^0 \ln(p_B^1 / p_B^0) + s_C^0 \ln(p_C^1 / p_C^0)] \quad (11.8)$$

$$P_{GP} \equiv \exp[s_A^1 \ln(p_A^1 / p_A^0) + s_B^1 \ln(p_B^1 / p_B^0) + s_C^1 \ln(p_C^1 / p_C^0)] \quad (11.9)$$

Thus, the period 1 values for each of these two indices are equal to share weighted geometric averages of the regional price indices, p_A^1 / p_A^0, p_B^1 / p_B^0 and p_C^1 / p_C^0, where P_{GL} uses the regional share weights pertaining to period 0, s_A^0, s_B^0 and s_C^0, and P_{GP} uses the regional share weights pertaining to period 1, s_A^1, s_B^1 and s_C^1. From Table 11.3 it can be seen that the geometric Laspeyres index P_{GL} is approximately 1 percentage point below the superlative indices P_F and P_T while the geometric Paasche index P_{GP} is approximately 1 percentage point above the superlative indices.([14])

([14]) It can be verified that the geometric mean of P_{GL} and P_{GP} is exactly equal to P_T. Thus if P_{GL} is below P_T, then P_{GP} will necessarily be above P_T.

Hence, the use of the geometric Laspeyres or Paasche formulae cannot be recommended when constructing aggregates of regional price indices; these formulae are unlikely to closely approximate a superlative index, which can readily be constructed using regional data on house price sales.

11.23 The above methods for aggregating over regional price indices assumed that median prices in each region correspond to houses of comparable quality over the two periods being compared. Now suppose that instead of using median prices in each region to represent constant quality house prices, it was decided to use mean prices in each region. Again, since it is desirable to have price times volume equal to expenditure in each period for each region, once it is decided to use mean prices as the constant quality a price concept, the corresponding volume should equal expenditures divided by price. Thus using the mean price in each region as a constant quality price for each time period leads to the data on regional expenditures (the v^t), prices (the p^t) and volumes (or implied quantities $q^t = v^t / p^t$) that are listed in Table 11.4 below.

Table 11.4. Regional Expenditures, Prices and Volumes (Implicit Quantities) Using Mean Prices as the Regional Prices

Period	v_A^t	v_B^t	v_C^t	p_A^t	p_B^t	p_C^t	q_A^t	q_B^t	q_C^t
0	1300	500	675	325	500	225	4	1	3
1	1725	400	825	345	400	275	5	1	3

Source: Authors' calculations based on MLS® data for a Canadian city

11.24 Using means instead of medians as the constant quality price in each region changes the regional price indices. The mean-based period 1 regional price indices are equal to $p_A^1 / p_A^0 = 345/325 = 1.06154$, $p_B^1 / p_B^0 = 400/500 = 0.80$, and $p_C^1 / p_C^0 = 275/225 = 1.2$ for regions A, B and C respectively. Again, there are widely differing house price inflation rates in the three regions when mean prices are used in place of median prices.

11.25 Using means instead of medians, the various overall price indices defined by formulae (11.1) to (11.9) can be calculated. The following counterpart to Table 11.3 is obtained using these formulae applied to the data in Table 11.4.

Table 11.5. Overall House Price Indices using Mean Prices and Alternative Formulae to Aggregate over Regions A, B and C

Period	P_F	P_T	P_L	P_P	P_0	P_1	P_A	P_{GL}	P_{GP}
0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
1	1.05305	1.05222	1.05253	1.05357	1.05253	1.07101	1.06177	1.04187	1.06267

Source: Authors' calculations based on MLS® data for a Canadian city

It can be seen that the use of mean prices instead of median prices for each region has led to very different indices; the superlative indices P_F and P_T are now about 3 percentage points higher in period 1. However, the use of mean prices has led to Laspeyres and Paasche indices, P_L and P_P, that are fairly close to their superlative counterparts. Since the base period share weighted index P_0 is numerically equal to P_L, P_0 is also fairly close to P_F and P_T. However, the other two shared weighted indices, P_1 and P_A, are well above the superlative indices. Finally, the Geometric Laspeyres index, P_{GL}, is well below P_T and the Geometric Paasche index, P_{GP}, is well above P_T. In any case, the use of mean prices in the housing context is not recommended since the mean price of a house in a region is unlikely to hold the quality of the houses constant over time.

Hedonic Regression Methods

11.26 Chapter 5 discusses the use of hedonic techniques for calculating house price indices. There are various ways of applying this technique when calculating price indices in general and residential property price indices in particular. The handbook presents three variants of the hedonic approach. These are: the time dummy variable method, the characteristics prices (or imputation) method, and the stratified hedonic method. Compared to the other approaches, all these hedonic methods are typically more data intensive, often requiring more information compared to the other approaches for constructing constant quality house price indices. This is because, in addition to data on prices, some pertinent characteristics (both structural and environmental) for each observation that is used in the regression are needed with hedonic methods. In principle, the more detailed the set of characteristics is and the larger the sample of housing units, the more reliable and accurate will be the resulting price index.([15])

11.27 A hedonic model expresses the price of a good as a function of its price-determining characteristics (or attributes). Chapter 5 covered two frequently used functional forms, which are the linear model and the logarithmic-linear (or semi-log) model, although other options (e.g., the Box-Cox technique) are often also treated in the literature, they are not covered here. The semi-log form is convenient because the interpretation of the regression coefficients is straightforward: once multiplied by 100, the coefficients can be interpreted as the percent change in the price of the house that results from a unit change in the explanatory variable.

11.28 To illustrate as plainly as possible how the various hedonic house price indices are constructed, the extensive version of the dataset used for calculating the mean and median prices above will also be consulted for the following examples. To simplify the presentation, the number of price-determining characteristics will be limited to four (continuous) variables. These are: lot size (land), number of bedrooms (rooms), number of bathrooms (bath), and age (age). The initial results for a regression using OLS with a semi-log functional form for a single year (2008) are summarised in Table 11.6.

([15]) Although most hedonic regressions on house prices in the literature will often use many more explanatory variables, some studies and the examples in Chapter 5 show that reliable hedonic price indices can be obtained with as few as four independent variables.

Table 11.6. Log-linear Regression Results for a Simple Example

Source |	SS	df	MS			Number of obs =	796
						F(4, 791) =	156.02
Model |	20.0634692	4	5.0158673			Prob > F =	0.0000
Residual |	25.4293063	791	.032148301			R-squared =	0.4410
						Adj R-squared =	0.4382
Total |	45.4927755	795	.057223617			Root MSE =	.1793
lprice |	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]	
rooms |	.1156791	.0098159	11.78	0.000	.0964108	.1349473	
bath |	.0999522	.0095996	10.41	0.000	.0811086	.1187958	
age |	-.002561	.0004173	-6.14	0.000	-.0033801	-.001742	
land |	9.39e-06	1.28e-06	7.31	0.000	6.87e-06	.0000119	
_cons |	12.0647	.0383342	314.72	0.000	11.98945	12.13995	

Source: Authors' calculations based on MLS® data for a Canadian city

11.29 From the regression on a sample of 796 price observations it is found that all four explanatory variables have the expected sign and are significantly different from 0 (using a t-test). The adjusted R-squared (or coefficient of determination) is 44 %, i.e., variations in lot size, the number of bedrooms, bathrooms, and age account for 44 % of house price variability. By adding more explanatory variables to the regression, the R-squared would increase. In fact, by adding three independent variables (the presence of a fireplace, the presence of a garage, and the age squared to account for the non-linearity associated with this variable) improved the adjusted R-squared to 54 %.

11.30 The regression results can be interpreted as follows:

- An extra square foot of lot size will increase the price of the house by 0.000939%, *ceteris paribus*.
- Each additional bedroom adds 11.6% to the price of a house, *ceteris paribus*.
- A house with an extra bathroom cost almost 10% more than a house without the extra bathroom, *ceteris paribus*.
- By adding one year to the house, its price declines (or the housing unit depreciates) by 0.2%, *ceteris paribus*.

The Latin locution *ceteris paribus* means "all variables other than the ones being studied are assumed to be constant". Turning to the variable "number of bedrooms" as an example, it cannot be concluded that houses with more bedrooms will always cost more; other factors are at play that can affect the price of the house such as its location and age, and overall quality of its construction. What is meant by qualifying the statement by *ceteris paribus* is that when houses vary only in terms of the number of bedrooms for instance (i.e., they are comparable in all other respects) then those with more bedrooms will cost more.

11.31 What follows are simplified examples of the various methods, as discussed in Chapter 5, for calculating hedonic price indices. The time dummy variable method is presented first. All examples use OLS regressions.

The Time Dummy Variable Method

11.32 The time dummy variable method is based on the estimation of a logarithmic-linear hedonic regression model where the data are pooled across all periods. The model is given by equation (6.5) and is repeated here for convenience:

$$\ln p_n^t = \beta_0 + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t \qquad (11.10)$$

where D_n^τ is dummy variable which is equal to one if the observation comes from period τ $(\tau = 1,\ldots,T)$ and is zero otherwise. The time dummy variable for the base period 0 – i.e., the start period from which the subsequent price changes will be compared – is left out to avoid perfect collinearity of all dummies with the intercept term β_0, known as the 'dummy trap'. With the time dummy variable approach the base period and the subsequent comparison periods, $t = 1,\ldots,T$, are the same units of time, i.e., a month, a quarter, or a year, depending on the particular circumstances such as the needs of the users or data availability.

11.33 The exponential or anti-logarithm of the estimated regression coefficient $\hat{\delta}^\tau$ measures the percent change in 'constant quality' property prices between the base period and period *t*. To understand why $\exp(\hat{\delta}^\tau)$ is a measure of quality adjusted, pure price change, the following steps have been worked out. The predicted logarithm of price in period 0 for property *i*, given its base period characteristics, z_{nk}^0 $(k = 1,\ldots,K)$, is

$$\ln \hat{p}_n^0 = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k z_{nk}^0 \qquad (11.11)$$

In period 1, the predicted logarithm of price must be evaluated at the property's *base period characteristics*, because quality should be held constant, hence

$$\ln \hat{p}_n^{1*} = \hat{\beta}_0 + \hat{\delta}^1 + \sum_{k=1}^{K} \hat{\beta}_k z_{nk}^0 \qquad (11.12)$$

Taking the differences between the estimates for both periods yields

$$\ln \hat{p}_n^{1*} - \ln \hat{p}_n^0 = \ln(\hat{p}_n^{1*} / \hat{p}_n^0) = \hat{\delta}^1 \qquad (11.13)$$

Expression (11.13) does not depend on *n*. That is, the result holds for all houses in the sample. As pointed out in Berndt (1991), the estimate of δ^t can be interpreted as the change in the logarithm of price due to the passage of time, holding all other variables constant. Taking the anti-log of $\hat{\delta}^1$ gives the estimated price index for period 1:

$$P_{TD}^{01} = \exp(\hat{\delta}^1) \qquad (11.14)$$

A similar exercise can be done for all other periods. The time dummy price index going from the base period to a comparison period t $(0 < t \leq T)$ therefore is

$$P_{TD}^{0t} = \exp(\hat{\delta}^t) \qquad (11.15)$$

Obviously, the time dummy hedonic index for the base period is equal to 1.

11.34 The following example illustrates the procedure for calculating a time dummy price index. Suppose that detailed information about the houses that were transacted over two years ($t = 2006$ to $t = 2007$) is available. Using the same information as in the basic data set above, the data for all periods are combined into the following pooled regression equation:

$$\ln p_n^t = \beta_0 + \beta_1 Lotsize_n + \beta_2 Bedroom_n + \beta_3 Bathroom_n + \beta_4 Age_n + \delta^1 D_n^1 + \varepsilon_n^t \qquad (11.16)$$

The left-hand side of equation (11.16) has the logarithm of the price of house *i* in year *t* (2006 or 2007) as the dependent variable. The right-hand side has the same explanatory variables (except for the time dummy variables) that one would find in a one period hedonic regression. In this particular case the explanatory variables are: lot size, number of bedrooms, number of bathrooms, and age; the respective parameters range from β_1 to β_4. Since this is a pooled regression, the estimated parameters (or regression coefficients) will be constrained over the years for which data are used in the regression. The error term ε_n^t indicates if an observed value is above or below the regression line. Also on right-hand side of the equation is the intercept term, β_0.

11.35 The regression results using the basic data set are listed in Table 11.7. The coefficient of interest is the one associated with year 2007, $\hat{\delta}^{07}$. Its value is 0.0781548. This coefficient is then transformed to arrive at an estimate of the price index (or the per cent change in prices) for houses between years 2006 and 2007. This transformation consists in taking the anti-logarithm of coefficient $\hat{\delta}^{07}$: $P_{TD}^{07/06} = \exp(0.0781548) = 1.08129$. Thus, the per cent change in house prices between years 2006 and 2007, holding constant all the characteristics of the house, is 8.1 %. Note that the mean and the median yielded increases of 10.1 % and 9.2 %, respectively, for this same period.

11.36 If a third period (year 2008) is added, then the hedonic regression equation becomes:

$$\ln p_n^t = \beta_0 + \beta_1 Lotsize_n + \beta_2 Bedroom_n + \beta_3 Bathroom_n + \beta_4 Age_n + \delta^1 D_n^1 + \delta^2 D_n^2 + \varepsilon_n^t \qquad (11.17)$$

Table 11.8 contains the regression output. The value of the time dummy coefficient for year 2008 is 0.1332734. Taking its anti-logarithm generates a value of $e^{0.1332734} = 1.14$, showing an increase in the constant quality house price index of 14 % between the base year, 2006 and the most recent year, 2008. By contrast, the price progression over the same period generated by the mean and median was respectively 16 % and 17 %.

Table 11.7. Results from a Pooled Regression for Years 2006 and 2007

Source |	SS	df	MS			Number of obs =	1708
						F(5, 1702) =	286.64
Model |	48.4501865	5	9.6900373			Prob > F =	0.0000
Residual |	57.5372376	1702	.033805663			R-squared =	0.4571
						Adj R-squared =	0.4555
Total |	105.987424	1707	.062089879			Root MSE =	.18386
lprice |	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]	
rooms |	.0840483	.0069071	12.17	0.000	.0705009	.0975957	
bath |	.121815	.0071529	17.03	0.000	.1077855	.1358444	
age |	-.0029137	.0003183	-9.15	0.000	-.0035381	-.0022894	
land |	.0000137	9.24e-07	14.78	0.000	.0000119	.0000155	
d2007 |	.0781548	.0089128	8.77	0.000	.0606736	.095636	
cons |	11.96531	.0273032	438.24	0.000	11.91176	12.01886	

Source: Authors' calculations based on MLS® data for a Canadian city

Table 11.8. Results from a Pooled Regression for Years 2006 to 2008

Source |	SS	df	MS			Number of obs =	2504
						F(6, 2497) =	366.64
Model |	73.4886776	6	12.2481129			Prob > F =	0.0000
Residual |	83.4154327	2497	.033406261			R-squared =	0.4684
						Adj R-squared =	0.4671
Total |	156.90411	2503	.06268642			Root MSE =	.18277
lprice |	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]	
rooms |	.0942001	.0056566	16.65	0.000	.083108	.1052923	
bath |	.1139931	.0057443	19.84	0.000	.102729	.1252572	
age |	-.0028112	.0002538	-11.08	0.000	-.0033089	-.0023135	
land |	.0000122	7.51e-07	16.28	0.000	.0000108	.0000137	
d2007 |	.0781257	.008856	8.82	0.000	.0607598	.0954916	
d2008 |	.1332734	.0090681	14.70	0.000	.1154916	.1510552	
_cons |	11.95724	.0225891	529.34	0.000	11.91295	12.00154	

Source: Authors' calculations based on MLS® data for a Canadian city

11.37 This technique can be extended to more than three periods as more periods become available. This consists in pooling more periods of data and adding additional time dummy variables. However, multi-period pooled regressions are not necessarily ideal for constructing a time series since adding new periods of data will likely modify the results from the previous periods. For instance, in the above example, when year 2008 is added to the previously pooled regression, the coefficient for year 2007 becomes 0.0781257, which in this specific case is only slightly different compared to the estimate obtained with the regression of Table 11.7, where the corresponding coefficient was 0.0781548. Moreover, the stability of the coefficients in a pooled regression can become an issue as the number of periods expands.

11.38 An alternative approach mentioned in Chapter 5 is to use the adjacent-period time dummy variable technique. If the hedonic regression is based on two consecutive periods τ and $\tau+1$, the hedonic relationship becomes:

$$\ln p_n^t = \beta_0 + \delta^{\tau+1} D_n^{\tau+1} + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t \qquad (11.18)$$

In the context of the three periods of data used in the above examples, a hedonic regression is first run for periods 0 and 1, and then a second regression is run for periods 1 and 2 using the four characteristics. The regression output for the first adjacent period regression is obviously the same as in Table 11.7, and the resulting period-to-period price index yields an estimate of 108.1. Table 11.9 shows the regression output for adjacent years 2007 and 2008.

Table 11.9. Results from a Pooled Regression for Years 2007 and 2008

Source |	SS	df	MS			Number of obs =	1670
						F(5, 1664) =	271.91
Model |	45.441478	5	9.0882956			Prob > F =	0.0000
Residual |	55.6172267	1664	.033423814			R-squared =	0.4497
						Adj R-squared =	0.4480
Total |	101.058705	1669	.060550452			Root MSE =	.18282
lprice |	Coef.	Std. Err.	t	P>|t|	[95 % Conf.	Interval]	
rooms |	.1041401	.0068861	15.12	0.000	.0906337	.1176465	
bath |	.1070142	.0068881	15.54	0.000	.093504	.1205244	
age |	-.0026926	.0003045	-8.84	0.000	-.0032899	-.0020953	
land |	.0000117	9.42e-07	12.42	0.000	9.85e-06	.0000135	
d2008 |	.0555370	.0089625	6.20	0.000	.073116	.037958	
_cons |	12.07482	.026871	449.36	0.000	12.02212	12.12753	

Source: Authors' calculations based on MLS® data for a Canadian city

11.39 The constant quality price index is calculated as the antilogarithm of the coefficient for year 2008 (0.0555370), so that the index becomes $\exp(0.0555370) = 1.057$. Recall that this is the price change from period 2007, not from the base period 2006. From these results, a time series can be constructed by *chaining* the two period-to-period indices (starting with the value 1 for the base period): $P_{TD}^{07/06} = 1.081$; $P_{TD,chain}^{08/06} = 1.081 \times 1.057 = 1.143$. This result differs only slightly from the full-period pooled regression (see Table 11.8) where we estimated a price change of 14.0 % over the entire period. Now, with chaining adjacent period time dummy indices, the estimated price change is 14.3 %.

Characteristics Prices or Imputation Method

11.40 The next hedonic regression approach presented in Chapter 5 is the characteristics prices or hedonic imputation method, henceforth simply the characteristics method. Applying this method to the same data as previously used, a quality-adjusted price index is estimated. For ease of presentation and interpretation, a *linear* model will be regressed to generate the results.([16])

11.41 The characteristics prices approach uses the implicit prices of the characteristics of the model (the regression coefficients) as the basis for constructing the price index, in a similar way as in a typical price index formula, but where the regression coefficients assume the role of the prices and the quantities are the quantities are the number of units of characteristics. Thus, the hedonic equation is estimated for each time period separately. The linear hedonic models for the base period 0 (2006) and for period 1 (2007) are

$$p_n^0 = \beta_0^0 + \beta_1^0 Lotsize_n + \beta_2^0 Bedroom_n + \beta_3^0 Bathroom_n + \beta_4^0 Age_n + \varepsilon_n^0 \quad (11.19)$$

$$p_n^1 = \beta_0^1 + \beta_1^1 Lotsize_n + \beta_2^1 Bedroom_n + \beta_3^1 Bathroom_n + \beta_4^1 Age_n + \varepsilon_n^1 \quad (11.20)$$

11.42 Estimating these equations on the sample data from 2006 and 2007, respectively, using OLS regression, generates the results shown in Tables 11.10 and 11.11. In this example, the implicit price of an extra bedroom in 2006 is $24329 while each additional bathroom will add $43190 to the price of the house. The results for 2007 in this highly simplified example are understandably different from those for 2006: an additional bedroom now seems to increase the price by $35147, while the price of an extra bathroom is now estimated to be $43463.([17])

([16]) There is nothing to prevent however the use of a semi-log or log functional form. Both can be used with this hedonic approach.

([17]) Note that the coefficients for the number of bedrooms are somewhat volatile between both years. This is to be expected because hedonic regressions are often characterized by the presence of multicollinearity between these two predictor variables. It should be stressed however that multicollinearity does not in itself affect the accuracy of the overall index. This phenomenon is only an issue if an accurate monetary value is needed for the value of an additional bedroom and/or for an additional bathroom, such as would be the case with a property assessment exercise. It should also be added that for the purpose of this simplified exercise, the sample size is relatively small. This can also explain why sometimes the results are not quite as robust as is often the case with larger samples.

Table 11.10. Results from a Regression for 2006

Source |	SS	df	MS			Number of obs =	834
						F(4, 829) =	141.49
Model |	2.4182e+12	4	6.0454e+11			Prob > F =	0.0000
Residual |	3.5420e+12	829	4.2726e+09			R-squared =	0.4057
						Adj R-squared =	0.4029
Total |	5.9601e+12	833	7.1550e+09			Root MSE =	65365
price |	Coef.	Std. Err.	t	P>|t|	[95 % Conf. Interval]		
rooms |	24329.78	3557.79	6.84	0.000	17346.45	31313.12	
bath |	43190.01	3734.288	11.57	0.000	35860.24	50519.79	
age |	-1083.309	164.5957	-6.58	0.000	-1406.382	-760.2357	
land |	5.168582	.4474175	11.55	0.000	4.290378	6.046787	
_cons |	98333.45	14450.86	6.80	0.000	69968.88	126698	

Source: Authors' calculations based on MLS® data for a Canadian city

Table 11.11. Results from a Regression for 2007

Source |	SS	df	MS			Number of obs =	874
						F(4, 869) =	169.68
Model |	3.5694e+12	4	8.9236e+11			Prob > F =	0.0000
Residual |	4.5702e+12	869	5.2592e+09			R-squared =	0.4385
						Adj R-squared =	0.4359
Total |	8.1397e+12	873	9.3238e+09			Root MSE =	72520
price |	Coef.	Std. Err.	t	P>|t|	[95 % Conf. Interval]		
rooms |	35147.31	3777.91	9.30	0.000	27732.41	42562.2	
bath |	43463.76	3858.683	11.26	0.000	35890.33	51037.19	
age |	-1059.767	173.0922	-6.12	0.000	-1399.495	-720.0394	
land |	5.829323	.5388036	10.82	0.000	4.771814	6.886831	
_cons |	79248.85	14337.87	5.53	0.000	51107.95	107389.7	

Source: Authors' calculations based on MLS® data for a Canadian city

11.43 The next step is to compute a hedonic price index from the regression results. A price index for 2007 compared to period 2006 can, for example, be expressed as

$$P^{01} = \frac{\hat{\beta}_0^1 + \beta_1^1\bar{z}_1^0 + \hat{\beta}_2^1\bar{z}_2^0 + \hat{\beta}_3^1\bar{z}_3^0 + \hat{\beta}_4^1\bar{z}_4^0}{\hat{\beta}_0^0 + \hat{\beta}_1^0\bar{z}_1^0 + \hat{\beta}_2^0\bar{z}_2^0 + \hat{\beta}_3^1\bar{z}_3^0 + \hat{\beta}_4^1\bar{z}_4^0} = \frac{\sum_{k=0}^{K}\hat{\beta}_k^1\bar{z}_k^0}{\sum_{k=0}^{K}\hat{\beta}_k^0\bar{z}_k^0} \quad (11.21)$$

where $\bar{z}_k^0$ is the sample mean value of the k-th characteristic in the base period; $\bar{z}_0^0 = 1$. Price index compilers will recognize that the index described by (11.21) is a Laspeyres-type price index: the estimated characteristics prices in period 0 (2006) and period 1 (2007), $\hat{\beta}_k^0$ and $\hat{\beta}_k^1$, are weighted by the average base period quantities of the characteristics. Put differently, the average base period quantities for all characteristics are valued at their implicit prices in the base period and in the current period. Table 11.12 lists the average sample values for the characteristics in this example. Using these values and the coefficients from Tables 11.10 and 11.11, the Laspeyres-type hedonic index between the base year (2006) and 2007 is computed as

$$P^{07/06} = \frac{79248 + (35147 \times 3.63) + (43463 \times 2.76) + }{98333 + (24329 \times 3.63) + (43190 \times 2.76) + }$$

$$\frac{+ (-1059 \times 23.89) + (5.829323 \times 6719)}{+ (-1083 \times 23.89) + (5.168582 \times 6719)} = 1.082$$

The 8.2 % increase in prices so obtained compares, in this particular case, quite closely with the 8.1 % obtained using the time-dummy approach from Table 11.7.

Table 11.12. Mean Values of the Characteristics for the Base Period (2006)

	Mean	Std. Err.	[95 % Conf. Interval]	
rooms	3.633094	.0244034	3.585194	3.680993
bath	2.767386	.0269044	2.714578	2.820195
age	23.88969	.5693338	22.77219	25.00719
land	6719.492	184.8605	6356.644	7082.339

Source: Authors' calculations based on MLS® data for a Canadian city

11.44 For subsequent periods, the compiler has a decision to make. He or she can use the same base year quantities to calculate the subsequent indices using the Laspeyres formula but replacing the implicit prices in the numerator with the relevant ones. Alternatively, quantities (mean characteristics) from the previous period could be used to generate period-to-period price indices. These bilateral indices would then be chained to create a continuous time series of linked indices. Other options are also available, and these are discussed in Chapter 5, but the mechanics of constructing the index remain essentially the same as presented here.

The Repeat Sales Method

11.45 The most significant problem with using (non-stratified) median or mean transaction prices to measure trends in houses prices is that the variation in the composition of the sample of properties sold from period to period is not always accurately accounted for. This issue can be partially circumvented by constructing an RPPI based on the repeat sales method, which was discussed in Chapter 6. In fact, one very popular house price index that is closely scrutinized in the U.S., the Case-Shiller house price index, is based on the repeat sales methodology.

11.46 The strategy for constructing a repeat sales house price index is quite straightforward. It consists in comparing the change in the price of identical properties that have sold at two points in time. In other words, it uses matched (or like-for-like) sampling as the basis for selecting the units that will be used in the calculation of the index. For the repeat sales approach to be tractable, one must have access to a large database of transactions covering a fairly long period. Otherwise the data needs are relatively modest: with the basic repeat sales method, only information on the dwellings address (or another location identifier) is required in order to identify which units have sold repeatedly, in addition of course to the selling price and the sale date.([18])

11.47 A simple example can illustrate the application of the repeat sales methodology.([19]) Assuming the objective is to estimate an annual index of price change between 2008 and 2010, Table 11.13 shows data for a small number of transactions. Property A sold in 2008 for $100 000 and sold again in 2009 for $120 000; property B is sold in 2008 for $175 000 and sold again in 2010 for $220 000; property C sold in 2009 for $180 000 and sold again in 2010 at the same price.

([18]) One assumption is that the quality of the house has not changed over the period between the two sales. If information about the features of the property is available to the compiler, then it is possible to exclude from the calculation those observations that have undergone significant changes over time and that are likely to affect the price and thus distort the index. Furthermore, given that high turnover is often a sign that certain undesirable features for that particular property may be at play so that these observations can also be excluded from the calculation. It should also be mentioned that repeat-sales indices are not always strictly constant quality price indices since houses are often subject to some loss in value over time as a result of depreciation. Consequently, repeat-sales price indices typically underestimate true house price inflation, unless some corrective adjustment is made to the estimates. If the purpose of the index is to act as a short- to medium-term indicator of house prices, then the issue of depreciation which the repeat-sales approach does not handle adequately can perhaps be set aside.

([19]) The example is partially drawn from the Canadian Teranet-National Bank® repeat sales price index documentation: http://www.housepriceindex.ca/Default.aspx.

Table 11.13. Repeat Sales Data

	2008	2009	2010
Property A	$100 000	$120 000	No sale
Property B	$175 000	No sale	$220 000
Property C	No sale	$180 000	$180 000
Average	*$137 500*	*$150 000*	*$200 000*

As a first step, the price change over the 2008 to 2010 period is estimated using the mean of prices approach. The annual average prices from 2008 to 2010 are respectively $137 000, $150 000 and $200 000. The corresponding year-to-year changes in average prices are 9.1 % and 33.3 % for the periods 2009/2008 and 2010/2009.

11.48 These results are now compared with those obtained if the repeat sales technique is used. Let *P* be the price relative of the house between the second and first sale for each completed transaction ([20]) from 2008 to 2010. The logarithm of *P* will serve as the *dependent variable* in a repeat sales regression. Three repeated sales are identified in Table 11.13 for the period 2008 to 2010. The first repeat sale, for property A, has a *P* value of 1.200 (i.e., the price relative between its sale prices in 2009 and 2008); the second repeat sale, which occurs for property B, has a *P* value of 1.257 (the price relative between its selling prices in 2010 and 2008); property C is the third repeat-sales transaction which has a *P* value of 1 because the price of this property did not change from 2009 and 2010.

([20]) Geltner and Pollakowski (2006) use the term "round trip".

11.49 The *independent variables* in a repeat sales regression are dummy variables, which take the value -1 during the year of the initial sale, then take the value +1 in the period of the second sale, and finally take the value 0 for all other periods. The estimated dummy variable coefficients from the regression are used to calculate the repeat sales price index. Table 11.14 summarizes the values of the dummy variables for properties A to C. For example, since property A is sold for a second time in 2009, the dummy variable D2009 takes the value of 1 but D2010 takes a value of 0 since this property A is not sold after 2009. A similar reasoning applies to the other properties and the other years. Note that to avoid perfect collinearity, the first period (2008) is disregarded from the explanatory variables and the regression. In other words, if the first sale occurs at the base year, then there is no dummy variable for that period.

Table 11.14. Dummy Variables for Repeat Sales

	P	D2009	D2010
Property A	1.200	1	0
Property B	1.257	0	1
Property C	1.000	-1	1

11.50 Given these repeat sales data, the regression equation – which has no intercept term – can be expressed as (see also equation (6.3):

$$\ln P_n^t = \gamma^{2009} D_n^{2009} + \gamma^{2010} D_n^{2010} + \varepsilon_n^t \qquad (11.22)$$

where ε_n^t is an error term ("white noise"). The anti-logarithm of the estimated parameters, i.e. $\exp(\hat{\gamma}^{2009})$ and $\exp(\hat{\gamma}^{2010})$, will represent the price indices of the housing unit for each period when compared to the base period 2008. Using Ordinary Least Squares (OLS) to estimate equation (11.22) on the data from Table 11.14, the resulting repeat sales price indices are 1.219 and 1.238 for 2009 and 2010, respectively. The year-to-year growth rates of 21.9 % and 23.8 % for this example are quite different from those found with the simple average approach, which were 9.1 % and 33.3 %. ([21])

11.51 The simple repeat sales model can be improved. One way of accomplishing this is by reducing the statistical noise in the index series generated. As pointed out by Geltner and Pollakowski (2006), the source of the estimation error (or noise) in property price indices is explained by the fact that the observed transaction prices are randomly distributed around the "true" but unobservable market values. The authors add that this noise is present in any house price index, regardless of how the index is constructed. To mitigate the effects of the noise the sample of repeated sales can be expanded, data availability permitting.

([21]) There are very few observations so no meaningful conclusions should be drawn from this simplified example. It should only be used for illustrative purposes.

11.52 As previously pointed out, an OLS regression can be used to obtain the set of price changes. The Bailey, Muth, and Nourse (1963) model is a classic example of the OLS repeat sales methodology using the technique outlined above. However, subsequent research has suggested that the basic OLS repeat sales method may be improved by applying a weighted least squares (WLS) technique. In a nutshell, the method consists in giving more weight in the regression to the observations that are deemed more accurate. In the context of the repeat sales method, giving less weight to properties for which a long time span has elapsed between sales and *vice versa* corrects for this inherent problem, better known as the heteroskedasticity problem.

11.53 Case and Shiller (1987) suggest the following three-stage approach:

1. Estimate model (11.22) by OLS regression and retain the vector of regression residuals.

2. Run an OLS regression of the squared residuals on a constant term and the time interval between sales.

3. Run an OLS regression of model (11.22) but where each observation is divided through by the square root of the fitted value from the second-stage regression.

The third stage is a weighted least squares regression of model (11.22) that accounts for the presumed heteroskedasticity.

Table 11.15. Unweighted Repeat Sales Regression

Source |	SS	df	MS			Number of obs =	1186
						F(6, 1180) =	379.41
Model |	32.5127473	6	5.41879122			Prob > F =	0.0000
Residual |	16.8531146	1180	.014282301			R-squared =	0.6586
						Adj R-squared =	0.6569
Total |	49.365862	1186	.04162383			Root MSE =	.11951

diflnprice |	Coef.	Std. Err.	t	P>|t|	[95 % Conf.	Interval]
dy2003 |	.0613539	.0086332	7.11	0.000	.0444157	.0782921
dy2004 |	.1198942	.0082047	14.61	0.000	.1037969	.1359915
dy2005 |	.1431862	.008343	17.16	0.000	.1268173	.159555
dy2006 |	.1845885	.0084578	21.82	0.000	.1679945	.2011826
dy2007 |	.2658241	.0083474	31.85	0.000	.2494468	.2822015
dy2008 |	.3438869	.0087587	39.26	0.000	.3267025	.3610713

Source: Authors' calculations based on MLS® data for a Canadian city

Table 11.16. Weighted Repeat Sales Regression

Source |	SS	df	MS			Number of obs =	1186
						F(6, 1180) =	348.90
Model |	2098.21619	6	349.702699			Prob > F =	0.0000
Residual |	1182.72363	1180	1.00230816			R-squared =	0.6395
						Adj R-squared =	0.6377
Total |	3280.93982	1186	2.76639108			Root MSE =	1.0012

ndifprice |	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ndy2003 |	.0635307	.0085609	7.42	0.000	.0467345	.0803269
ndy2004 |	.1211754	.0081162	14.93	0.000	.1052516	.1370992
ndy2005 |	.1437457	.0082962	17.33	0.000	.1274688	.1600226
ndy2006 |	.1864151	.0084621	22.03	0.000	.1698127	.2030175
ndy2007 |	.2689894	.0084844	31.70	0.000	.2523433	.2856356
ndy2008 |	.3491619	.0091085	38.33	0.000	.3312913	.3670325

Source: Authors' calculations based on MLS® data for a Canadian city

11.54 Moving to the larger and more realistic set of data on single-family houses that were previously used for most of the previous examples of this chapter, two versions of the repeat sales method are illustrated. The results are first computed for the unweighted repeat sales regression approach and are presented in Table 11.15. Table 11.16 presents the results for the weighted version of the repeat sales regression. Note that for this particular set of data, all the coefficients are significantly different from 0 and that no intercept is used in the regressions for the repeats sales approach. One often cited drawback of the repeat sales method is that it is wasteful of data. The current exercise confirms this. Of the 5787 observations that were in the database at the start, only 1186 (or about 20 %) are found to be units that sold more than once during the 6 or so years.

11.55 Similar to the time dummy hedonic model presented earlier, the corresponding price indices are obtained by taking the antilogarithm of the estimated coefficient as

the dependent variable is the logarithm of the price. For example, the regression for the unweighted repeat sales approach yields a coefficient of 0.2658241 for 2007; taking the antilogarithm yields $\exp(0.2658241) = 1.3045$ (or 130.5 once rounded and multiplied by 100). The indices for the entire 2002 to 2008 period are shown in Table 11.17. Note that the indices are quite similar, regardless whether the unweighted or weighted repeat sales versions are used. This is a feature of this particular dataset and may not necessarily hold true for house price indices estimated from other sources.

Table 11.17. Repeat Sales Price Indices (2002 = 100)

Year	Unweighted	Per cent change	Weighted	Per cent change
2002	100.0		100.0	
2003	106.3	6.3	106.6	6.6
2004	112.7	6.0	112.9	5.9
2005	115.4	2.4	115.5	2.3
2006	120.3	4.2	120.5	4.4
2007	130.5	8.5	130.9	8.6
2008	141.0	8.1	141.8	8.3

Source: Authors' calculations based on MLS® data for a Canadian city

11.56 Table 11.18 summaries the index results using the various methods presented here using the extended dataset for year 2007. The simple mean shows the largest increase of all the estimated indices at 10.1 % with the median being slightly lower at 9.2 %. The hedonic indices increased by 5.7 % and 5.9 % for the adjacent year pooled and characteristics prices approaches, respectively (calculation not shown above). By contract, the repeat sales weighted and unweighted indices increased by 8.5 % and 8.6 %, respectively. Although the sample size is somewhat small to make any generalisation, one important observation is noteworthy. The non-quality adjusted indicators, i.e., the mean and median, generate the highest growth rates, while the hedonic methods generate the smallest. The repeat sales approaches, although they control for many potential aspects of quality, do not control for age. Therefore, it is not so surprising that the price increases obtained with this approach are larger than those obtained with the hedonic approaches.

Table 11.18. Growth Rates in Percent for the Various House Price Indices (2007)

Mean	Median	Pooled hedonics	Characteristics hedonics	Repeat sales unweighted	Repeat sales weighted
10.1	9.2	5.7	5.9	8.5	8.6

Source: Authors' calculations based on MLS® data for a Canadian city