House Price Index Methodologies

N. Edward Coulson
Department of Economics
Penn State University
University Park, PA 16802
fyj@psu.edu; 814-863-0625

Keywords: price index, regression, omitted variables, repeat sales

Prepared for the *International Encyclopedia of Housing and Home*, to be published by Elsevier, Limited..
Comments welcome.

House Price Index Methodologies

In 2005, both forbes.com (http://www.forbes.com/2005/04/26/cx_sc_0426home.html) and cnnmoney.com (http://money.cnn.com/pf/features/lists/hpci_data/index.html) published online articles on the topic of the US's most expensive housing markets. Each of the two articles constructed rankings of housing markets based on the price of a "typical" house, but the method of defining typical was quite different in the two surveys. The ranking by forbes.com used the US zip code as the area of analysis; that is, each zip code was defined as a distinct housing market. For each zip code, the survey constructed a *median sales price*, the price which is at the halfway point in a list of most expensive to least expensive units in that zip code. As forbes.com itself noted, the unit that has the median price can differ substantially from place to place, and comparing prices across these locations will be like comparing apples to oranges. Atherton, California, a community located in the bay area of Northern California, had the highest median sales price (of $2.4 million), but one of the reasons for this was because the median house in Atherton is a large house, on a large plot of land with many expensive amenities. According to the 2000 Census, the median number of rooms in Atherton's zip code (94027) is 8.2, which is certainly larger than the national median of 6.2. And so the comparison is of apples to oranges.

The comparison at cnn.com takes a different approach. They asked a leading real estate brokerage service to find, for each housing market, the price of "a 2,200-square-foot house with 4 bedrooms, 2 1/2 bathrooms, a family room and a two-car garage" located in a neighborhood "typical for corporate middle-management transferees." They are trying to compare apples to apples. The cnn.com survey seems to use market areas larger than zip codes, so it is a little bit difficult to compare its findings to that of the forbes.com report. Nevertheless it is quite instructive to note that the top place on the cnn.com list was La Jolla, California, a beach area in San Diego County. On the forbes,com list the zip code corresponding to La Jolla ranked only 73rd . It is tempting, to infer 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. Indeed, the median number of rooms for owner occupied units in the 92037 zip code was at the national median of 6.2.

The heterogeneous nature of housing precludes our ability to make price comparisons over space and time simply by taking averages in different locations or in different years. We must compare like to like using price indices. In the case of "ordinary" price indices like the Consumer Price Index, a basket of n commodities of quantities $x=x_1...x_n$ is posited. The prices of those commodities in market 1 are given as $p_1=p_{11}...p_{1n}$ and the resultant expenditure is $\Sigma p_{1i}q_i$ where summations throughout are over the index i=1...n. Prices from a spatially or temporally distinct market 2 are also gathered and the expenditure *on the same basket* is also computed. Comparison is usually taken in the form of a ratio, where market 1 (the numeraire market) expenditure is in the denominator and the comparison is in the numerator and for clarity the result is multiplied by 100:

$$P = \frac{\sum p_{2i} q_i}{\sum p_{1i} q_i} x100$$

In the case of comparing housing prices, the situation is somewhat more intricate. The quantities, X_i, are not items on the grocer's shelves but attributes of a "typical" house. In the case of the cnn.com survey these were quantities of square feet, bedrooms, baths, garage spaces, family rooms, and neighborhood quality. But what "prices" were assigned to these "commodities"? The usual answer lies within regression estimation, in this context sometimes called mass appraisal. A sample of sales within a particular market *i*, including the sale price and characteristics 1 through k of the associated property, are gathered, and a regression of the form

$$P_i = \beta_{0i} + \beta_{1i} X_1 + \ldots + \beta_{ki} X_k + e \qquad (1)$$

is run. The β's, i.e. the regression weights, or parameters, for each characteristic, are estimated for each city, and interpretable in this linear case as the *implicit prices* of the associated characteristics are estimated. For any given house, the set of attributes, X_1 through X_k can be fed into the equation and a valuation, an *appraisal*, of that house can be calculated. Since the weights are specific to each market, a price index similar to (1) can be calculated. One market is chosen as the numeraire market and cross-sectional price indexes can be calculated. The existence of the intercept term, β_0, absent from the calculation of traditional prices indexes, should be noted. Informally speaking, the value of this term will derive from all of the characteristics of the housing market that are constant across the units in that market. Representing, as it does, things like sunshine and proximity to ocean, it can be a major contributor to housing price differentials across cities.

The study by Palmquist (1984) is convenient for exemplary purposes. Palmquist, using FHA mortgage insurance filings, estimated appraisal equations for the Atlanta, Denver, Houston, Louisville, Oklahoma City, and Seattle metropolitan areas. Table 1 provides a list of characteristics used in the regression, and the β-weights for each city. 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 any case, we set the values of the quantities of characteristics to the values in the column labeled X*. These are zero for any attribute that is missing from any of the cities' hedonic regressions. The row labeled "Constant Quality Price" provides the appraisal of a house with X* in each metropolitan area-- that is, $\beta_0 + \Sigma\beta_i X_i$ for each city.

There are differences. Seattle's price is the highest, at over $60,000 (recall this is 1984), followed by Denver, at just over $50,000. Atlanta and Houston are the lowest priced markets 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. We can then construct the price index as

$$P = \frac{\beta_{i0} + \Sigma \beta_{ij} X_j}{\beta_{A0} + \Sigma \beta_{Ai} X_j} x100 \qquad (2)$$

where the A subscripts identifies parameters from Atlanta. The last row of Table 1 shows the results of this calculation. 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 comparing indexes depends rather importantly on the quantities.

So it is with the hedonic price indexes here. It can be the case that 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 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, where it can be seen that 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.

There are three important issues that the previous simple example has ignored:

(1) Omitted variables: The vector of variables will not necessarily contain all of the important determinants of housing prices. This is understood in regression analysis, and the role of the error term is to encompass all of the unobservable influences on housing prices within the database. The problem arises when there are, roughly speaking, systematic differences in those unobservables across markets as this can cause bias in the estimated prices and therefore in the price indexes..

(2) Nonlinearity: Because housing attributes are tied bundles of the attributes, there is no particular reason why the relation between housing attributes and building price should be linear. Many alternatives to the linear model have been hypothesized, and one popular alternative used below is the semilog function

$$\ln P_i = \beta_{0i} + \beta_{1i} X_1 + \ldots + \beta_{ki} X_k + e \qquad (3)$$

in which case the parameters are not prices, per se, but semi-elasticities. The Laspeyre price index (2) is replaced by

$$P_i = \frac{\exp(\beta_{0i} + \Sigma \beta_{ij} X_j + .5 s_i^2)}{\exp(\beta_{01} + \Sigma \beta_{1j} X_j + .5 s_1^2)} x100 \qquad (4)$$

where s^2 is the estimate of the variance term of the error in (3), and is included to provide an estimate of the mean value of price, rather than the median (Malpezzi, Chun and Green, 1998). Other nonlinear forms are of course possible.

(3) Parsimony: Estimating equations (2) or (3) separately for each city market may be too much too ask of the available data. It can be plausible to assume that the intercept term is the major difference in the index parameters, since the difference in capital prices (i.e. the price of structural characteristics) may be arbitraged across locations. Thus the following regression can be employed, using the data from all markets, and assuming semilog form:

$$\log P = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \gamma_1 M_1 + \gamma_2 M_2 + \ldots + \gamma_h M_h + e \qquad (5)$$

where the γ parameters are coefficients of the h indicator variables M_j representing h of the h+1 cities in the database. Note that the semilog specification implies that these are estimates of the percentage difference between the constant quality price in the indicated city and the base city

represented by β_0, and so it itself an index number. That is to say, the price index (4) reduces to

$$P_i = \frac{\exp(\beta_0 + \Sigma\beta_j X_j + \gamma_i + .5s^2)}{\exp(\beta_0 + \Sigma\beta_j X_j + .5s^2)} x100 = \exp(\gamma_i) x100 \qquad (6)$$

For small γ, we can use the approximation $\exp(\gamma)=1+\gamma$ and interpret it directly as the percentage difference between the base market price and the *i*th market price.

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 in the spirit of comment (3) above, one can simply estimate a single regression with indicator variables for 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:

$$\log P = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \gamma_1 T_1 + \ldots + \gamma_h T_h \qquad (7)$$

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 β_0 is the intercept term which represents the normalized period. This is exactly analogous to equation (5) and the construction of the index is as in (6)

In research applications this latter method seems to be preferred over the use of separate hedonic regressions for each time period. 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. This makes some intuitive sense. Spatially distinct markets will have large differences in supply and/or demand that would contribute to creating statistically significant differences in the regression 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.

A parsimonious version of (7) would replace the time variables with a time trend

$$\log P\beta = {}_0 + \beta_1 X_1 + \ldots + \beta_k X_k + cW$$

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. This model is a specialized case of the previous approach where the coefficients in (7) behave according to the pattern $c*j= \gamma_j$. 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.

A temporal price index can alternatively be constructed using the increasingly popular method of *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:

$$\log P_1 = \beta_0 + \beta_1 X_{11} + \ldots + \beta_k X_{k1} + \gamma_1 T_1 + e_1 \qquad (8)$$

and

$$\log P_2 = \beta_0 + \beta_1 X_{11} + \ldots + \beta_k X_{k1} + \gamma_2 T_2 + e_2 \qquad (9)$$

because of course the values of all the other time indicators are equal to zero. Now subtract (8) from (9). If none of the attribute sizes changed between the two sales (i.e. the X values are constant) the price difference would be:

$$\log P_2 - \log P_1 = \gamma_2 T_2 - \gamma_1 T_1 + e_2 - e_1 \qquad (10)$$

Now, the difference between two error terms is just another error term (although perhaps one with different properties, K. Case and Shiller (1989)). Therefore we can write the above down in the following way:

$$\Delta \log P = \gamma_2 T_2 - \gamma_1 T_1 + v \qquad (11)$$

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 (4.11)) , and write down a regression model of the form:

$$\Delta \log P = \gamma_1 T_1 + \gamma_2 T_2 + \ldots \gamma_k T_k + v \qquad (12)$$

where the T_j's are no longer indicator variables in the strictest sense. Instead, they take on the value of -1 for those observations which had "first sales" in that time period and +1 for those observations which had their second sale during the time period (and zero otherwise). For any given observation the "fitted value" will be something like equation (11) with 1 and 2 being replaced by the appropriate first and second sale time periods.

Running the above regression provides estimates of the b_j's and, with one further modification, the sequence of γ_j's form a *mulitiplicative 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 is to let the first time period be the normalization. In the logarithmic model above, this amounts to setting the initial price equal to zero. The γ-terms are then values relative to the first time period. The index construction proceeds as in equation (6). This model basically underlies the well-known price indexes provided by the US

Office of Federal Housing Enterprise Oversight (http://www.ofheo.gov/HPI.asp) and by Freddie Mac (http://www.freddiemac.com/finance/cmhpi).

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. This is especially significant since both *observed and unobserved attributes are eliminated* and thus comment (1) above no longer has as much force. The differencing operation which takes place in the repeat sales model removes all attribute levels and so that source of 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 time indexes of property, then the repeat sales model has considerable appeal. The research of K. Case and Shiller (1989) and the increased availability of databases with repeat sales in them has caused the use of this model has exploded.

The repeat-sales model is not, however, without its own faults (Meese and Wallace, 1997). One, the database is restricted to properties with multiple sales, which may be a small portion of the overall database, and moreover, such properties may be a nonrandom sample. Properties that sell multiple times within a given time frame may be systematically different. Also the 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.

As an example, there is one attribute that is clearly not constant over the inter-sales period, and that is the age of the dwelling. Allowing this X variable to change over time causes the difference between (8) and (9) to become

$$\Delta \log P = \beta_1 \Delta Age + \gamma_1 T_1 + \ldots + \gamma_k T_k + v \qquad (13)$$

although in some such specifications collinearity can become problematic (Coulson and McMillen, 2008). What is always true of age might be sometimes be true of almost any other structural or neighborhood characteristic, but with any such change, as long as it is observable (i.e. involves the X characteristics), similar modifications can be made. The repeat sales price index can still be characterized by the sequence of γs assuming that all of the ΔX terms are set to zero.

Potential changes in the β coefficients themselves are somewhat more difficult to handle. The most general case is when each β coefficient has a different value in each time period. B. Case and Quigley (1991) discuss the possibility that each β has a distinct deterministic trend, but also the more general model

$$\log P_t = \beta_0 + \beta_{1t} x_{1t} + \ldots + \beta_{kt} x_{kt} + e_t$$

(also see Clapp and Giaccotto (1998)). Subtracting the repeat sale at time period t-r we get the individual change in appraisal:

$$\log P_t - \log P_{t-r} = (\beta_{1t} - \beta_{1r}) x_{1t} + \ldots + (\beta_{kt} - \beta_{kr}) x_{kt} + e_t - e_r \qquad (14)$$

and for the sample as a whole, the regression model becomes

$$\log P_t - \log P_s = \sum_{i=1}^{T} \sum_{j=1}^{k} T_{ij} (\beta_{ij}) x_j + v \qquad (15)$$

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

$$T_{it} = 1 \quad if \quad i = t$$
$$= -1 \quad if \quad i = t - r$$

Several things can be noted. First, each sequence of β parameters may be interpreted as a repeat sales index not for the unit, but for the X characteristic itself. Second, a consequence of this is that the one of the advantages of the repeat sales index is lost; a set of benchmark values X* must be selected, in the manner of Table 1, in order to construct the price index sequence for housing itself, but is a straightforward modification of the above methods. Third, if temporal parameter variation is suppressed, then this model reverts to the described in text following equation (13). Finally, as Case and Quigley (1991) and Clapp and Giaccotto (1997) note, it is straightforward to combine this repeat sales model with data on one-time sales, since the β parameters from a one time sale at time t ought to be equal to the corresponding βs in (15), although the doubts about the comparability of such samples expressed above may prevent this.

REFERENCES

Martin J. Bailey, Richard F. Muth, and Hugh 0. Nourse, (1963) "A Regression Method for Real Estate Price Index Construction," *Journal of the American Statistical Associa-tion ,*58, 933-942.

Bradford Case and John Quigley (1991) "The Dynamics of Real Estate Prices" *The Review of Economics and Statistics*,73, 50-58

Karl Case and Robert Shiller (1989) "The Efficiency of the Market for Single Family Homes," *American Economic Review,* 79, 125-137.

John Clapp and Carmelo Giaccotto (1998) "Price Indices Based on the Hedonic Repeat-Sales Method: Application to the Housing Market" *Journal of Real Estate Finance and Economics*, 16, 5-26

Edward Coulson and Daniel McMillen (2008) "Estimating time, age and vintage effects in housing prices" *Journal of Housing Economics*, 17, 138-151

Stephen Malpezzi, Gregory H. Chun, Richard K. Green (1998) "New Place-to-Place Housing Price Indexes for U.S. Metropolitan Areas, and Their DeterminantsReal Estate Economics", *Real Estate Economics*, 26, 235-274

Richard Meese and Nancy Wallace (1997) "The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression and Hybrid Approaches. *The Journal of Real Estate Finance and Economics* 14 ,51 -73

Raymond Palmquist (1984) "Estimating the Demand for the Characteristics of Housing" *The Review of Economics and Statistics*, 66, 394-404

TABLE 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		
=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
evel 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		