# House Price Indexes: Methodology and Revisions[1]

**Joseph M. Silverstein**
Federal Reserve Bank of Philadelphia
June 2014

Accurate measurements of house prices are important for a number of reasons. Housing is usually the most important investment a household ever makes, and home equity is typically the largest component of household wealth. In addition, housing is an important source of collateral for household borrowing, and, as we have seen, its value can be subject to considerable fluctuation. Also, research has shown that having a mortgage that is larger than the value of the underlying house is associated with an elevated risk of default.

Measuring aggregate house prices, estimated using house price indexes (HPIs), tends to be surprisingly difficult. Furthermore, the methodologies that are most accurate in the long run are subject to considerable revision in the short run. For example, monthly revisions to a commonly used HPI for the United States released by CoreLogic tend to be large up to about the third revision. There is also some evidence that these revisions tend to be downward, although there is currently no clear theoretical explanation for this phenomenon. Therefore, HPI users who are interested in the most recent data should interpret them with caution.

In this report, I review the literature on HPI methodologies, revisions, and sources of bias. I first describe the methodology common to all "repeat-sales" indexes, some improvements on that methodology, and some competing approaches. I then discuss some sources and characteristics of revisions to repeat-sales indexes.

---

[1] The views expressed here are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. When he wrote this report, Joe Silverstein was a research analyst in the Research Department of the Philadelphia Fed.

**Methods of Measuring House Prices**

At first glance, aggregate house prices appear easy to measure: Just take the average of all the house prices in the region. But houses are bought and sold only occasionally. At times other than the time of sale, the value of a house is determined through an appraisal. But this is only an educated guess of the sale price of that house if it were sold at the time of the appraisal. So an HPI is usually constructed using only observed sale prices.

Suppose one takes the average of all the sales data available to construct the index, as the Bureau of Labor Statistics (BLS) does when constructing the consumer price index (CPI). On first sight, this method appears to accurately measure house prices, but a number of problems arise. Because of the highly skewed wealth distribution in the United States, a small percentage of houses are extremely expensive, skewing the average upward. Therefore, a median price would be more representative of prices for the vast majority of houses. This methodology is termed the *median sales price method*. The National Association of Realtors (NAR) computes such an index.

The median sales price methodology, however, is not ideal for many scenarios. Suppose that in September 2013, in Zip code 54321, more than half of the houses were mobile homes that sold at low prices, so that the median house price was only $60,000. But in October 2013, a smaller proportion of low-price mobile homes sells in 54321, and now the median house price is $150,000. For 54321, a median price index would indicate a 150 percent rise in house prices from September to October 2013. But this enormous rise does not accurately reflect the condition of the housing market in 54321; it reflects only a change in the quality of the housing sold. The BLS controls for such fluctuations in quality in the CPI and producer price index (PPI) by constructing these indexes using primarily homogeneous goods that are very similar across time and markets. For example, the BLS might always use bags of U.S. extra fancy grade golden delicious apples weighing 4.4 pounds to represent the *apples* category in the CPI.[2] Since there are many such apples for sale, the BLS is able to construct a constant-quality index of homogeneous goods that are highly comparable across time, using only a small sample of all apples sold. Unfortunately, this method is not possible with housing, since no two houses are the same. Even if some houses were the same, not enough data are available to identify groups of highly similar houses.

The *repeat-sales methodology* is one common solution to this problem. In this methodology, housing quality is controlled for by comparing sales of the same house across time. This entails comparing only paired sale prices for the same house. An index constructed in this way is called a *repeat-sales* index.

One drawback of this approach is that sale prices are not entered into the index until they

---

[2] See http://stats.bls.gov/cpi/cpifaq.htm#Question_7.

are paired with a subsequent sale.[3]  A repeat sale does not occur very often, so a lot of data are not used. Also, the houses with paired sales may not be representative of the housing market as a whole. For example, cheaper "starter" homes tend to sell more often. Furthermore, it is possible that the difference in composition for these new sales pairs will be systematic across time, potentially leading to a bias in the index. This will be true, for example, if the composition of houses sold depends on the business cycle.

Another feature of the repeat-sales methodology is that any time two sales can be paired, the values of the house between the two sales dates are imputed using the two sales prices, which often means that every HPI value all the way back to the earlier purchase date is revised slightly to reflect the addition of a new sales pair. Since many new sales pairs with many different sales dates are added to the index with each new index release, this in practice means that every value of the HPI will be revised (often substantially) with each new release, with the most recent releases usually getting revised the most. The revisions are especially significant at lower levels of aggregation, such as Zip codes, where they can be as much as 5 percent in one month. These large revisions mean that the most recent data in a repeat-sales index must be interpreted with caution. Although the additional data are likely to increase the accuracy of the index, the most recent releases contain a mixture of information about both current and past changes in house prices, meaning that some degree of noise is added to the most recent index values.  As a result, it is possible for the additional data to reduce the accuracy of particular index values, although this is not typical.  However, controlling for the quality of houses sold is a serious issue and must be dealt with in order to have an accurate index, which is why repeat-sales is nonetheless the primary methodology used today. Some examples of repeat-sales indexes are the S&P/Case-Shiller index, the CoreLogic HPI, the Conventional Mortgage HPI and Freddie Mac HPI (both published by Freddie Mac), and the Federal Housing Finance Agency (FHFA) HPI.

The *hedonic methodology*, an alternative approach to house price index construction, estimates the typical effects of a house's attributes on price using regression. This approach allows the index to accurately measure the changes in the value of a home over time based on a single sale through inference using the typical value associated with the changes in house attributes over time. As such, the hedonic methodology can be thought of as a generalization of the repeat-sales methodology to allow changes in the value of particular attributes over time (i.e., the constant-quality assumption between paired sales is relaxed). For example, a simple linear regression (i.e., a best-fit line) might indicate that each additional bathroom adds $10,000 to the value of a house in a

---

[3] If a house has sold more than twice, only nonoverlapping sales pairs will be used.  For example, a house that sold three times will have the first and second sales and the second and third sales as its two sales pairs.  The pair formed by the first and third sales will not be included.

particular region, on average. Then a constant-quality index can be constructed based on attributes without the need to use only paired sales in the index and without assuming the underlying homes do not change in quality between two sales. This methodology is viewed favorably by many economists, but it is difficult to obtain enough data to accurately estimate the effects of the attributes, because in the United States these attributes are recorded only at the county level and are never aggregated into a national data set. For this reason, few hedonic indexes are available. The FNC HPI is an example of a hedonic index, but it began relatively recently, in October 2011, and has yet to catch on.

A number of researchers have also proposed hybrid repeat-sales/hedonic methodologies. Nagaraja, Brown, and Wachter (2012) compare such methodologies with a number of other common pure repeat-sales methodologies, and they conclude that these hybrid methodologies are better. However, this issue is beyond the scope of this report, and no indexes based on these methodologies are widely available.

**The Repeat-Sales Methodology: Further Details**

Bailey, Muth, and Nourse (1963) developed the repeat-sales methodology. The basic idea is to estimate the index values via a regression of the observed sale prices on the times of sale. Then the index values can be obtained from the estimated regression coefficients.[4] In the appendix, I present a simple numerical example of repeat-sales index construction to better elucidate the methodology.

In general, all index values are subject to revision with each new release because the regression best-fit line will change in slope as new data are added. For example, suppose we have a monthly HPI with 450 months of data. Then adding a new sales pair in month 451 for a house that sells in month 0 and month 451 and another pair for a house that sells in month 200 and month 451 could end up revising every month of HPI data by changing the regression coefficients on the times of sale.[5] The HPI presumably becomes more accurate with revisions as more data flow in. This creates a tension between the freshness of the data and its accuracy, which is one reason why it is important to understand the properties of revisions.

There is a body of literature on HPI revisions. Clapp and Giaccotto (1999) show that

---

[4] In practice, since an HPI is supposed to measure relative sale prices rather than absolute differences in sale prices, the dependent variable needs to be the ratio of sale prices at subsequent sales dates, and the independent variable needs to be the ratio of index values on those dates. This means that the equations have to be log-linearized beforehand in order to run a linear regression and then exponentiated after the regression in order to obtain the index values.

[5] Note that at least two houses must be added. Otherwise, the additional house price data would all be absorbed into the 451st month due to the additional degree of freedom.

revised repeat-sales index releases are more efficient than first releases in the sense that revised releases have lower standard errors (which the authors demonstrate theoretically as well). They also observe that a larger sample size is not necessarily associated with significantly smaller revisions. Using Los Angeles housing sales data from 1988 to 1994 and Fairfax County, VA, data from 1981 to 1991, they observe that revisions tend to be downward. They contend that removing or down-weighting "flips" — houses that were bought and sold within one to two years — causes the observed downward revisions to disappear. This is happening, they argue, because flipped houses do not appreciate at the same rate as other properties since they might be improved between sales, leading to upward bias. As more revisions are released, the proportion of sales pairs that are flips will decrease as the typical time between sales pairs increases, leading to downward revisions. They observe much smaller revisions when re-estimating their repeat-sales index without flips, supporting their hypothesis that flips are the primary source of revisions in repeat-sales indexes.

Bourassa, Cantoni, and Hoesli (2013) argue that revisions are largely due to both flips and forced sales such as foreclosures. Since, they argue, these transactions do not reflect underlying market conditions, they propose several new index methodologies that use robust regression techniques to down-weight such outliers.[6] They apply these robust regression techniques to data for Louisville, KY, from the first quarter of 2003 through the second quarter of 2005 and find that mean absolute values of quarterly revisions drop from 0.23 percent (using the conventional methodology) to 0.08 percent (using the robust methodology).

Clapham et al. (2006) study some other aspects of repeat-sales index revisions. They compare them with revisions to hedonic indexes and study the effects of revisions on house price derivatives, which are financial contracts that derive their value from house prices. By constructing both repeat-sales and hedonic HPIs using Swedish data from 1981 to 1999, they are able to directly compare the size of HPI revisions across the two index types.[7] They find the hedonic methodology to be less subject to revision than repeat-sales. They suggest that the difference in revision size occurs in large part because repeat-sales data points include information about past prices, whereas single-sale hedonic data points do not. They observe systematic downward revisions in repeat-sales indexes, but they leave the explanation of this phenomenon to future research. Clapham et al. (2006) also compare the size of cumulative HPI revisions with the size of cumulative CPI revisions over several years to benchmark the size of the revisions. CPI revisions are on the order of 0.2 percent and are comparable in size to the revisions to the hedonic HPI they constructed in their paper. By contrast, repeat-sales revisions can be of an entirely different order of magnitude at 2

---

[6] More specifically, they construct three indexes using the Huber estimator, the bisquare M-estimator, and the $L_1$ estimator.

[7] Sweden has a national property tax system, so data are available to accurately estimate a hedonic index.

percent to 4 percent. They also point out that HPI revisions can have a large effect on housing derivatives such as equity insurance products and housing futures contracts because trades are settled on the initial release of housing prices. Since revision stability is very important for housing derivative contract settlement, they recommend using a hedonic index instead of repeat-sales for this application.

Following up on Clapham et al. (2006), Deng and Quigley (2008) conduct an analysis of the magnitude and bias of revisions to the Office of Federal Housing Enterprise Oversight (OFHEO) HPI, which is now called the FHFA HPI. They consider a six-year period between two releases (the first quarter of 2001 and the first quarter of 2007), so revisions are likely to be permanent and not just aberrations. Usually, the six-year revisions are small, but sometimes they can be quite large. They find that over the six years, 25 percent of the average absolute revisions to the MSA-level index changed it by more than 1.5 percent, and 15 percent of the average absolute revisions exceeded 2 percent.[8] Long-term revisions were almost always downward during the period. They perform a panel regression across metropolitan statistical areas (MSAs) of percentage revisions on lagged HPI values along with quarter dummies and MSA dummies for every quarter and MSA. The regression results reveal that the coefficients on some of the lags are significant, indicating that lagged HPI values could potentially be used to predict revisions. They also regress percentage revisions on lagged prices and quarter dummies separately for each MSA and find that the average $R^2$ is 28 percent. This leads them to conclude that the recent HPI history in an MSA, known at the time of the initial release, possibly explains around 28 percent of the index revisions over the next six years. However, it must be emphasized that this is only one study, so there is still no conclusive evidence that index revisions are predictable.

**Properties of Revisions**

I will give a more detailed illustration of the properties of revisions using an HPI developed by CoreLogic, presenting some information about how this HPI behaves.[9]

CoreLogic releases HPI indexes at the national, state, and Zip code level. There is a large degree of variation by geographic level in the magnitude of revisions and in the rate at which those revisions decrease in size over time and converge to their final values. Figure 1 shows the average

---

[8] The first quarter of 2001 release and the first quarter of 2007 release both report HPI values for each quarter from the first quarter of 1975 through the first quarter of 2001, and they calculate the six-year revision in each of these quarters. Then they take the average of the absolute revisions.

[9] For further details on the properties and economic significance of revisions to the CoreLogic index, see the paper by Elul, Silverstein, and Stark.

absolute values of the percentage monthly revisions to the CoreLogic index.[10]  For example, the chart shows that the average absolute percentage change from the sixth release to the seventh release (revision 7) of the CoreLogic U.S. HPI is 0.162 percent.  Figure 2 does the same thing for the *cumulative* revisions.[11]  For example, Figure 2 shows that the average absolute percentage change from the first release to the seventh release (cumulative revision 7) of the CoreLogic U.S. HPI is 1.005 percent.

**Figure 1**
**Average Absolute Value of Percentage Monthly Revisions**
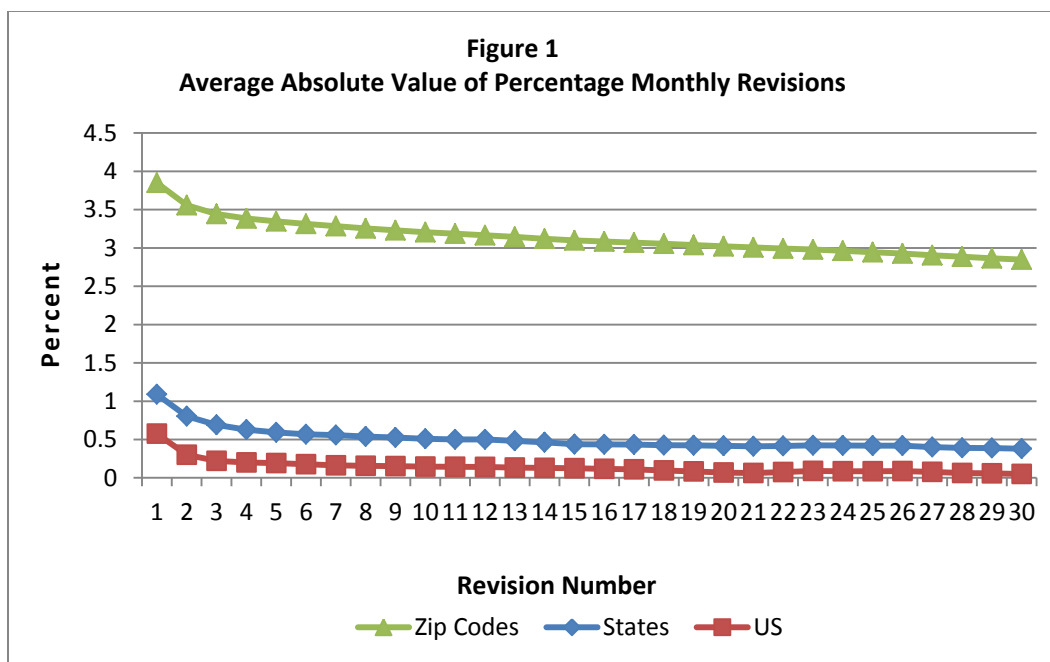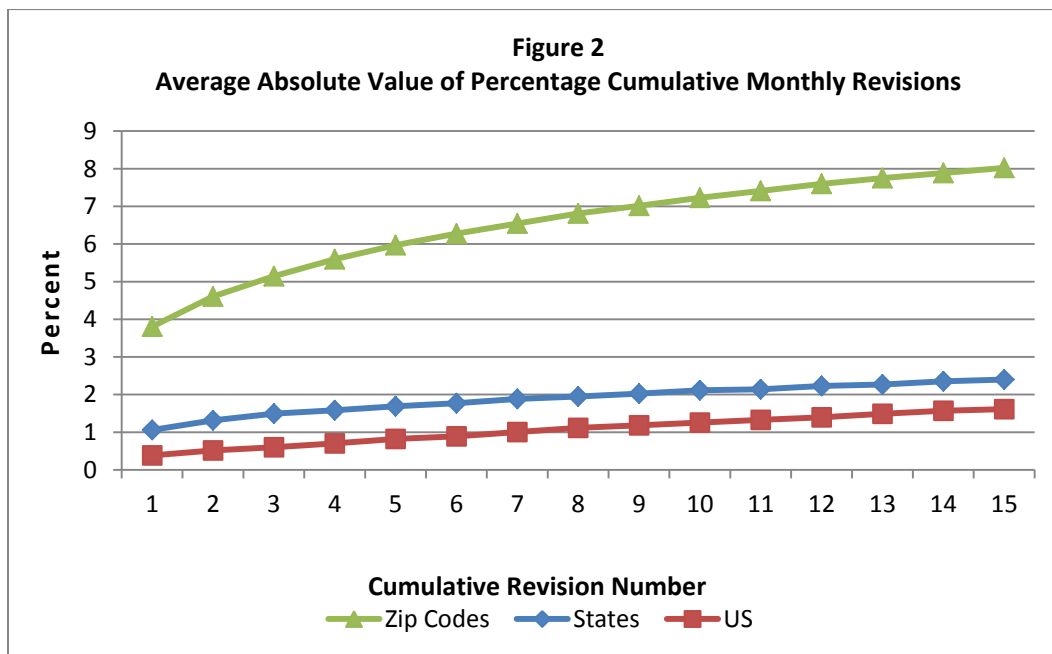


Source: CoreLogic.

As can be seen in Figure 1, the United States as a whole experiences an average absolute first revision of about 0.576 percent in our data, but the average absolute revision size is already down to about 0.221 percent by the third revision and gradually decreases after that. The average

---

[10] More precisely, Figure 1 shows the means of the means (over geographic region) of the absolute values of the percentage monthly revisions to the CoreLogic index using all of the data available to me as of November 5, 2013. For the U.S., I have monthly releases from September 2009 through August 2013, except for March 2010 which is not available. This gives me 45 data points for each revision of national data.  For the states and Zip codes, I have monthly releases from March 2011 through August 2013, which gives me 29 data points for each revision. As an example of my calculations for Zip code revision 25, the absolute values of the 29 data points are averaged for each Zip code, and then the average of those averages is taken to arrive at the value of 2.944 percent.

[11] Figure 2 shows the means of the means (over the geographic region) of the absolute values of the percentage monthly *cumulative* revisions to the CoreLogic index.  Because of data limitations, I am unable to provide as many cumulative revisions as monthly revisions because the amount of data available to me decreases with the cumulative revision number. For this reason, I averaged a fixed number of values across each geographic region equal to the number of values available as of the 15th cumulative revision, restricting to the most recent cumulative revision values for the lower-numbered cumulative revisions. For the national data, I use the 26 most recent values of the cumulative change. For the states and Zip codes, fewer data are available, and I can average only the 15 most recent values.

absolute first revision of the state data is 1.09 percent, but it goes down to 0.691 percent by the third revision. For the Zip codes, the revisions decrease from 3.85 percent in the first revision to 3.44 percent in the third revision. This shows that at the state level, revisions do not diminish as quickly as at the national level, and at the Zip code level they barely diminish at all. So Zip code data converge much more slowly than state or national data to the final value of the HPI. Across all levels of aggregation, revision size decreases substantially from the first to the third release, indicating that HPI users should be careful when interpreting index values that have been revised less than twice.

At a more local level, there is immense variation. For example, using these data, the average absolute value of first revisions in Oregon is 0.30818 percent, but the average absolute value of first revisions in West Virginia is 3.962 percent, an order of magnitude higher. At the Zip code level, the difference is even larger: 3.37 percent of Zip codes covered by CoreLogic have no new sales pairs as of the first revision and therefore have first revisions of 0 percent, but the same percentage of Zip codes experience first revisions in excess of 7.75 percent. In the CoreLogic data, the average revisions usually converge over time to zero. However, if a geographic region has very large absolute first revisions, then it also tends to have large absolute second and third revisions and so on, even though they still decrease over time.



**Figure 2**
**Average Absolute Value of Percentage Cumulative Monthly Revisions**

Source: CoreLogic.

In Figure 2, we see that cumulative revisions increase over time but tend to eventually plateau, albeit very slowly, meaning that years elapse before HPI values fully stabilize. The smaller the geographic area, the greater the cumulative revisions, which makes sense by comparison with Figure 1.

**Concluding Remarks**

The primary takeaway messages of this report are that housing prices are very difficult to measure and that repeat-sales indexes are subject to substantial revision, particularly in the short run and when measured at lower levels of aggregation. Initial releases of repeat-sales HPIs may also be subject to upward bias, resulting in downward revisions. Users of HPIs should be aware that a given index value may undergo large revisions initially but will then quickly stabilize. Thus, caution is warranted when relying on the first two HPI releases for a given period.

**References**

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse. "A Regression Method for Real Estate Price Index Construction," *Journal of the American Statistical Association*, 58:304 (1963), pp. 933-942.

Bourassa, Steven C., Eva Cantoni, and Martin Hoesli. "Robust Repeat Sales Indexes," *Real Estate Economics*, 41 (2013), pp. 517-541.

Butler, J.S., Yan Chang, and Amy Crews Cutts. "Revision Bias in Repeat-Sales Home Price Indices," Freddie Mac Working Paper 05-03 (2005).

Case, Karl E., and Robert J. Shiller. "Prices of Single Family Homes Since 1970: New Indexes for Four Cities," *New England Economic Review*, (September/October 1987), pp. 45-56.

Clapham, Eric, Peter Englund, John M. Quigley, and Christian L. Redfearn. "Revisiting the Past and Settling the Score: Index Revision for House Price Derivatives," *Real Estate Economics*, 34 (2006), pp. 275-302.

Clapp, John M., and Carmelo Giaccotto. "Revisions in Repeat-Sales Price Indexes: Here Today, Gone Tomorrow?" *Real Estate Economics*, 27 (1999), pp. 79-104.

Deng, Yongheng, and John M. Quigley. "Index Revision, House Price Risk, and the Market for House Price Derivatives," *Journal of Real Estate Finance and Economics*, 37 (2008), pp. 191-209.

Elul, Ronel, Joseph M. Silverstein, and Tom C. Stark. "Understanding House Price Index Revisions," mimeo, May 2014, www.elul.org/papers/hpi.pdf.

Nagaraja, Chaitra H., Lawrence D. Brown, and Susan S. Wachter. "Repeat Sales House Price Index Methodology," *Journal of Real Estate Literature* (forthcoming), available at Social Science Research Network Accepted Paper Series 2186661 (2012).

Nakamura, Leonard. "How Much Is That Home Really Worth? Appraisal Bias and House-Price Uncertainty," Federal Reserve Bank of Philadelphia *Business Review* (First Quarter 2010).

Shiller, Robert J. "Arithmetic Repeat Sales Price Estimators," *Journal of Housing Economics*, 1:1 (1991), pp. 110-126.

Wang, Ferdinand T., and Peter M. Zorn. "Estimating House Price Growth With Repeat Sales Data: What's the Aim of the Game? *Journal of Housing Economics*, 6 (1997), pp. 93-118.

**Appendix**

In this appendix, I explain the details of the Bailey, Muth, and Nourse (BMN) repeat-sales HPI construction methodology using a simple numerical example.[12]

Suppose there are 258 homes in Springfield, HI. Fifteen of these houses are sold in 1990, 18 in 1991, and 14 in 1992. Only three sales can be matched with other sales of those same houses from 1990 through 1992. Below, I chart the sales pairs that the county property tax assessor was able to put together for the three-year period. I denote the first (base) year as year 0, the second as year 1, and the third as year 2.

**Table 1**

|  | Year 0 | Year 1 | Year 2 |
|---|---|---|---|
| **House 1** | $200,000 | $194,000 |  |
| **House 2** |  | $420,000 | $400,000 |
| **House 3** | $110,000 |  | $130,000 |

We want to construct a repeat-sales housing index for Springfield for each year, and we will call the three index values $I_0$, $I_1$, and $I_2$. We set $I_0 = 100$ so that period 0 is the base period. The most important feature that we want our index to have is that it should measure percentage changes in the value of housing. Therefore, we need the ratios of the index values between time periods to be the same as the ratios of the house prices the index is supposed to measure. So we will want the index to satisfy the following equations:

$$\frac{I_1}{100} = \frac{194,000}{200,000}$$
$$\frac{I_2}{I_1} = \frac{400,000}{420,000}$$
$$\frac{I_2}{100} = \frac{130,000}{110,000}$$

Note that this is a system of three equations in two unknowns, so in general we cannot find numbers $I_1$ and $I_2$ that simultaneously satisfy all three equations. Instead, we can perform a linear regression to get as "close" as possible to solving all three equations simultaneously.

In order to run a linear regression, we need to have a linear system of equations. But the above system is not linear due to the presence of the $I_2/I_1$ term. We can linearize the equations by taking the log of both sides, thereby separating out the fractional terms:

---

[12] This example owes much to Wang and Zorn (1997).

$$\log\left(\frac{194}{200}\right) = \log(I_1) - \log(100)$$

$$\log\left(\frac{400}{420}\right) = \log(I_2) - \log(I_1)$$

$$\log\left(\frac{130}{110}\right) = \log(I_2) - \log(100)$$

Now define $\beta_1 = \log I_1$ and $\beta_2 = \log I_2$. So we want

$$\log\left(100 \cdot \frac{194}{200}\right) = 1 \cdot \beta_1 + 0 \cdot \beta_2$$

$$\log\left(\frac{400}{420}\right) = -1 \cdot \beta_1 + 1 \cdot \beta_2$$

$$\log\left(100 \cdot \frac{130}{110}\right) = 0 \cdot \beta_1 + 1 \cdot \beta_2$$

This system of three equations can be written as one equation using matrix notation:

$$\begin{bmatrix} \log\left(100 \cdot \frac{194}{200}\right) \\ \log\left(\frac{400}{420}\right) \\ \log\left(100 \cdot \frac{130}{110}\right) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Denote this system as $\vec{y} = \mathbf{X}\vec{\beta}$. Then, assuming certain assumptions are satisfied, ordinary least squares regression tells us that the best-fit coefficients are $\hat{\vec{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'}\vec{y}$. [13] Applying this to our example, we have

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left( \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \log\left(100 \cdot \frac{194}{200}\right) \\ \log\left(\frac{400}{420}\right) \\ \log\left(100 \cdot \frac{130}{110}\right) \end{bmatrix} \approx \begin{bmatrix} 4.6568 \\ 4.6901 \end{bmatrix}$$

Note that $\hat{\beta}_1 = \log \hat{I}_1$ and $\hat{\beta}_2 = \log \hat{I}_2$. So $\hat{I}_1 = e^{\hat{\beta}_1}$ and $\hat{I}_2 = e^{\hat{\beta}_2}$. Then we have

$$\hat{I}_0 \stackrel{\text{def}}{=} 100 \qquad and$$
$$\hat{I}_1 = e^{4.6568} = 105.30 \quad and$$
$$\hat{I}_2 = e^{4.6901} = 108.87$$

---

[13] Assuming mean-zero errors that have equal variances (homoscedasticity), the Gauss-Markov Theorem states that $\hat{\vec{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'}\vec{y}$ is the best linear unbiased estimator (BLUE) of the coefficient vector $\vec{\beta}$. In this sense, the estimate of the coefficients $\hat{\vec{\beta}}$ "comes closest" to solving $\vec{y} = \mathbf{X}\vec{\beta}$ by minimizing the sum of squared residuals $\sum_i \epsilon_i^2 = \vec{\epsilon}'\vec{\epsilon} = \left(\vec{y} - \mathbf{X}\vec{\beta}\right)'\left(\vec{y} - \mathbf{X}\vec{\beta}\right)$. This regression methodology is called ordinary least squares (OLS). See http://en.wikipedia.org/wiki/Ordinary_least_squares for further discussion of OLS and http://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Least_squares_estimator_for_.CE.B2 for its derivation.

This indicates that prices in Springfield rose over the first three years of the 1990s.

Most HPIs today do not use the basic BMN methodology demonstrated above, but instead build upon the same regression model while correcting for various limitations associated with it. For example, the BMN methodology calculates house prices as a geometric mean, which can be shown to always be less than or equal to the arithmetic mean. Since the arithmetic mean conforms to our intuitive understanding of an average, this means that BMN indexes are biased downward.[14] A revised methodology that corrects for this problem and others was introduced in Case and Shiller (1987) and described in more detail in Shiller (1991).[15] Shiller termed this methodology *(arithmetic) weighted repeat sales*, and it is often referred to as the Case-Shiller methodology.

**Revisions to Repeat-Sales HPIs**

Suppose the Springfield county property tax assessor is able to match two more sales when it releases its 1994 HPI. These additional houses were last sold in 1990 and 1991 but were not included in the HPI in those years because they were single sales that could not be paired with previous sales. The data that the tax assessor had on record showed that a house sold for $100,000 in 1990, and it now also has a new sale on record in 1994 for $115,000. It also now knows that a house that first sold in 1991 for $150,000 sold again in 1993 for $175,000. Let's see how the addition of these two new sales affects the index:

**Table 2**

|  | Year 0 | Year 1 | Year 2 | Year 3 |
|---|---|---|---|---|
| **House 1** | $200,000 | $194,000 |  |  |
| **House 2** |  | $420,000 | $400,000 |  |
| **House 3** | $110,000 |  | $130,000 |  |
| **House 4** | $100,000 |  |  | $115,000 |
| **House 5** |  | $150,000 |  | $175,000 |

---

[14] See http://en.wikipedia.org/wiki/Inequality_of_arithmetic_and_geometric_means for further discussion and proof.
[15] Another innovation of the Case-Shiller index is that it relaxes the assumption of homoscedasticity (that is, that the variance of house prices remains the same across time). Intuitively, sales pairs separated by a longer period will be associated with a larger error term (or an error term with higher variance).

Now, we have

$$
\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \quad and \quad y = \begin{bmatrix} \log\left(100 \cdot \dfrac{194}{200}\right) \\ \log\left(\dfrac{400}{420}\right) \\ \log\left(100 \cdot \dfrac{130}{110}\right) \\ \log\left(100 \cdot \dfrac{115}{100}\right) \\ \log\left(\dfrac{175}{150}\right) \end{bmatrix}
$$

Then,

$$
\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \left( \begin{bmatrix} 1 & -1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & -1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \log\left(100 \cdot \dfrac{194}{200}\right) \\ \log\left(\dfrac{400}{420}\right) \\ \log\left(100 \cdot \dfrac{130}{110}\right) \\ \log\left(100 \cdot \dfrac{115}{100}\right) \\ \log\left(\dfrac{175}{150}\right) \end{bmatrix} \approx \begin{bmatrix} 4.6403 \\ 4.6819 \\ 4.7697 \end{bmatrix}
$$

Therefore,

$$
\begin{aligned}
\hat{I}_0 &= 100 \\
\hat{I}_1 &= e^{4.6403} = 103.58 \\
\hat{I}_2 &= e^{4.6819} = 107.97 \\
\hat{I}_3 &= e^{4.7697} = 117.88
\end{aligned}
$$

Note that all the index values (except for the base period) have now been revised, including period 2 (the index for 1992). This occurred even though no new houses sold in period 2, because the new houses first sold prior to period 2 and therefore revised the estimated aggregate house price appreciation in period 2.

14