

# Philadelphia House Price Indices: Technical FAQ and Documentation

7/22/05

Kevin C. Gillen, PhD  
[gillen@econsult.com](mailto:gillen@econsult.com)

Disclaimers: The Philadelphia House Price Indices are in the public domain, and provided free of charge. Persons are free to share or otherwise use the indices as they see fit, provided that they cite the source and do not modify the original slides. However, the author requests neither the credit nor the blame for any investment or policy decisions undertaken based upon the information contained herein. Finally, although the author is affiliated with the Econsult Corporation, the indices reflect the views and opinion of the author, and not necessarily those of Econsult. © 2005, Kevin C. Gillen, All Rights Reserved.

Q: What are the Philadelphia House Price Indices?

A: The Philadelphia house price indices (hereafter, HPIs) are a set of indices characterizing the average rate of appreciation in Philadelphia house values over time. They are analogous to a Dow Jones Index, but for house prices rather than stock prices: although the actual level of the indices does not mean much, the change in the index from one time period to another does. In particular, the indices are estimated in such a way so that the percent change in the index over any time period should be representative of the average percent change in Philadelphia house values during that same period.

Q: How are the HPIs computed?

A: They are computed by estimating a hybrid hedonic regression model that uses data on house sales. The advantage of using regression estimation is that it effectively “controls” for a host of other factors which affect a home’s value, so that the underlying true appreciation rate of house values over time is effectively isolated and accurately measured. Along with a set of control variables in the regression specification, there is a set of time variables denoting the year and quarter that each property sold. It is from the estimated coefficients on these time variables that the indices are obtained.

Q: What is a “hybrid hedonic regression model”?

A: Simply put, a regression model is a statistical technique that fits an equation to data points. It systematically tries to measure how variation in a set of variables (called “independent variables”) explains the variation in the subject variable of interest (called the “dependent variable”). The objective of a regression is to explicitly parameterize the mathematical relationship between the dependent variable and the independent variables.

A regression is estimated by first collecting paired data on the variables, and then essentially fitting a line through the data points. The intercept and slope of this line describes the relationship between the variables. More specifically, the estimated equation for the line tells how then mean or average response of the dependent variable (e.g. house price) is functionally related to the level of the independent variable(s) (e.g. building square footage or time).

The particular mechanics of how this line is fitted to the data is called “Least

Squares". Ordinary Least Squares (or just "OLS" for short) is the standard default procedure used in the estimation of linear regressions. OLS fits a line through the data in such a manner that the sum of the squared distances between the line and the data points is minimized<sup>1</sup>. In this way, the overall level of differences between actual and predicted values is also minimized. Thus, the estimated regression equation describes the relationship between the dependent and independent variables in the most accurate way possible, given the dataset that the researcher is constrained to work with.

By extension, a hybrid hedonic regression is just a very specific type of regression that is applied to the pricing of composite goods. A hedonic regression equation postulates that the price of a good is the sum of the marginal prices of its individual attributes or components. The term "hedonic pricing" was originally coined by Court (1939), who estimated the pleasure (i.e. "hedonism") that people receive from the individual features of automobiles, such as size, speed, comfort, safety, etc.

Housing is one of the most salient examples of a composite good in existence. The purchase of a house entails the purchase of a location (and its proximity to various amenities or disamenities), the purchase of a structure (and its various components and features) and the purchase of a set of public services (and their various levels of quality and quantity). It is therefore reasonable to believe that the total market price of a home will vary systematically with things like: proximity to good restaurants and parks, the local crime and traffic congestion levels, the number and type of bedrooms and bathrooms, the design and age of the home, and the cost and quality of trash collection, fire and police protection and public schools, just to name a few factors.

The first hedonic regression of house prices was estimated by Rosen (1974). Using data on house sales, the author estimated a regression of house price on its structural and locational characteristics. The estimated coefficients on these characteristics gave the marginal price of a unit increase in their level. For example, what an extra bedroom, bathroom, fireplace or floor contributed to a property's overall value in dollar terms. The estimation of hedonic price regressions has since become ubiquitous in the housing economics profession.

Another type of pricing regression commonly applied to housing is a "weighted repeat-sales" regression. This regression uses data on the prices of the same physical dwellings at different points in time in order to estimate a price index. The price index is recovered from the estimated coefficients on those variables

<sup>1</sup> More formally, OLS estimates the intercept and slope of the line describing the relationship between two variables in such a manner that the squared sum of the differences between observed and predicted values are minimized. See Basic Econometrics (1995) by Damodar N. Gujarati for a good exposition.

measuring the amount of time since a home last transacted, because the percent difference in the price of the last sale and the price of a current sale is a home's rate of price appreciation. First proposed by Bailey, Muth and Nourse (1963), the weighted repeat-sales method was later extended by Case and Shiller (1987, 1989). The input to a weighted repeat-sales regression is a dataset containing multiple sales observations on the same set of properties: hence the name "repeat sales regression".

Finally, the term "hybrid hedonic pricing regression" is used to refer to a regression that combines both the hedonic and repeat sales methodologies. The regression that estimates the Philadelphia HPIs includes hedonic variables, such as floor area and number of stories, as well as repeat sales variables denoting the amount of time that has passed since each home last sold. The advantage of combining the two techniques is that it facilitates the fullest exploitation of the data possible, in order to obtain the most accurate possible measure of the underlying rate of house price appreciation.

For an excellent survey of these and other statistical techniques used to estimate house price indices, see "On Choosing Among House Price Index Methodologies" by Case, Pollakowski and Wachter (1991).

Q: Can you give an example of a house price regression?

A: Sure, no problem. Suppose that you are interested in how the value of a house varies with its overall size. In particular, the relationship you are hypothesizing is of this form:

$$P_i = \alpha + \beta \times S_i \quad (1)$$

Where:  $P_i$  = the price of the  $i$ th house  $S_i$  = the total square footage of the  $i$ th house  $i$  = an index of numbers denoting each unique house sale

The objective is to obtain numeric estimates for the parameters  $\alpha$  and  $\beta$ . With these estimates, you could just plug in any number for a dwelling's total square footage,  $S$ , into the equation in (1) and generate a prediction of the dwelling's value,  $P$ .

You could begin by collecting data on home sales during a particular time period (e.g. one quarter of the year), where you would record the sales price of the home and its total square footage. You could then put your data in table format like

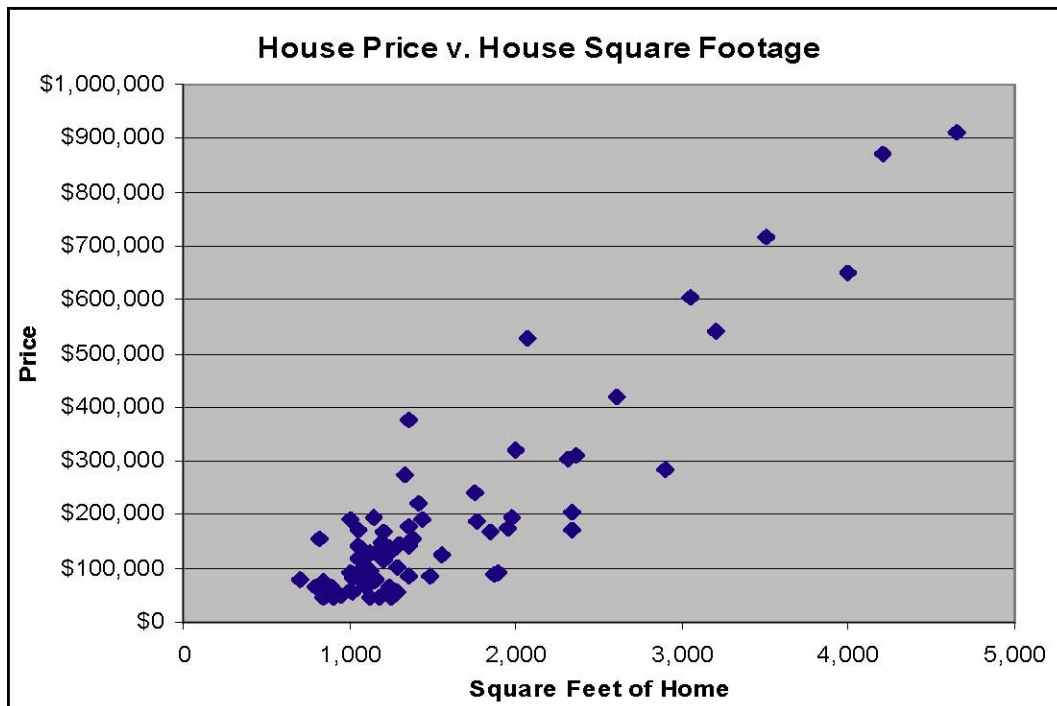
this:

$i$	$P_i$	$S_i$
1	\$85,000	1,100
2	\$100,000	1,250
3	\$125,000	1,300
.	.	.
.	.	.
.	.	.
N	\$750,000	3,200

Where  $i=1,2,\dots,N$

and  $N$ =the total number of sales records in your dataset.

I collected such data for a random sample of 100 home sales in Philadelphia during the first quarter of 2005. The next figure is a scatterplot of the data points that shows price against building square footage for this sample of data:



In general, it can be observed that the relationship between these two variables is positive: larger homes sell for higher prices. The lower left corner of the scatterplot contains small, low-priced homes while the upper right corner contains larger, higher-priced homes.

The exact numeric relationship between the two variables is obtained by estimating a regression of price on building square footage. As previously explained, the intuitive definition of a regression is that it is a statistical technique which computes a mathematical expression that best describes the relationship between two or more variables in a dataset. Mathematically, a regression computes the specific numeric parameter estimates describing this relationship between the dependent variable (in this case, house price) and the independent variable or set of independent variables (in this case, a home's square footage).

Continuing with the example shown in the scatterplot, the regression of house price on square footage was estimated using a specialized statistical analysis software known as SAS, and it produced the following results:

$$P_i = -137,340 + 208.59 \times S_i$$

(2) (-7.22) (18.95)  
N=100, R<sup>2</sup>=0.8211

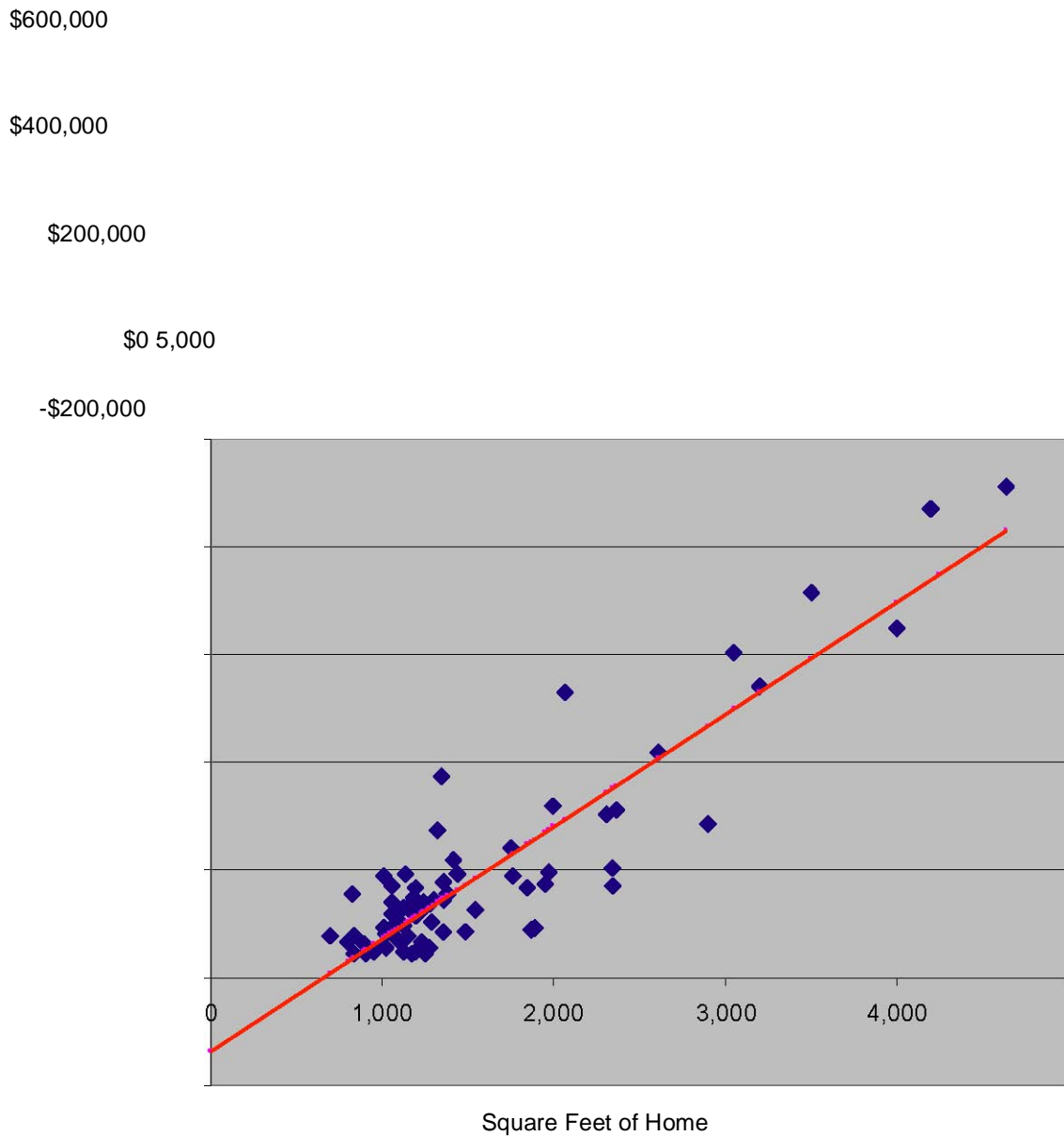
In words, this equation states that a house with zero square feet would have a negative value of -\$137,340, but the value of the home would increase by \$208.59 for every square foot that was added to the house. For a typical Philadelphia house that has an average of 1,260 square feet, this equation predicts that home would be worth \$125,483. Furthermore, the large values of the t-scores (the numbers in parentheses) indicate that the square footage of a home is a statistically significant factor in explaining house prices. The R<sup>2</sup> value of 0.8211 implies that 82.11% of the variation in house prices in this sample are explained by variation in the total size of the dwelling.

Here is a plot of the estimated linear pricing equation against the data points used to estimate the equation:

#### House Price as a Function of Square Footage

Price  
\$1,000,000

\$800,000



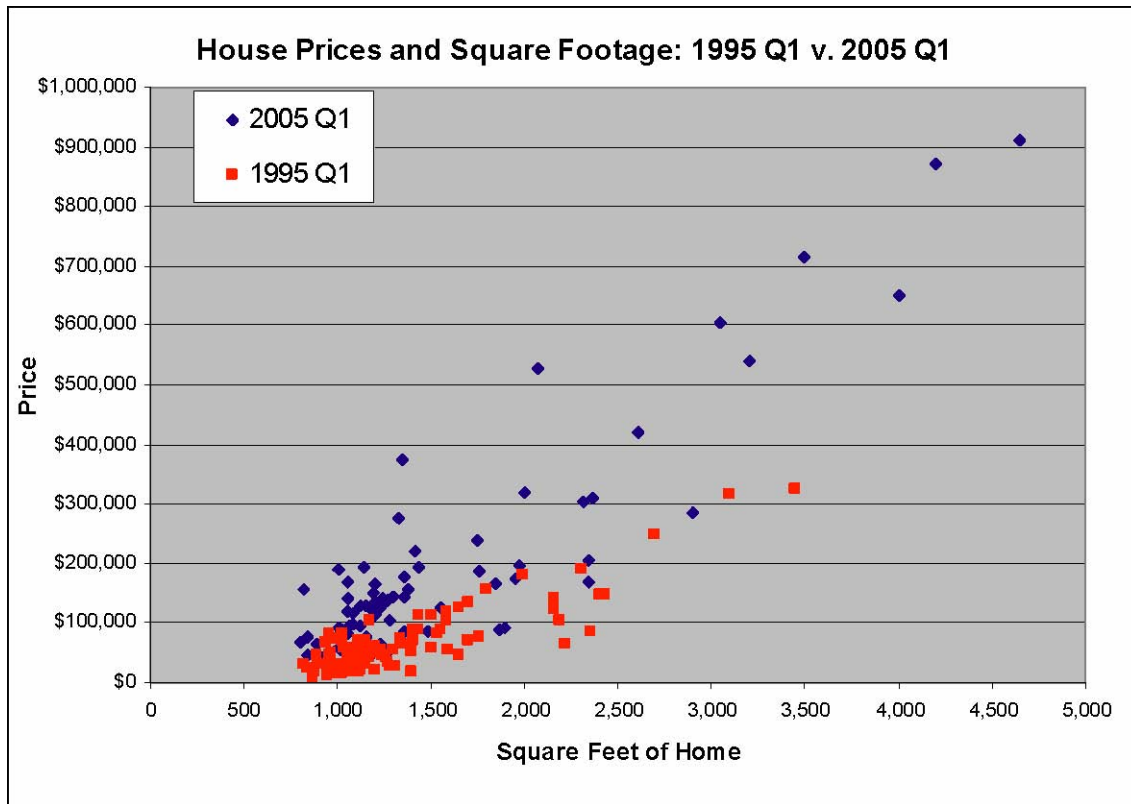
It can be seen that the line – which represents the estimated equation shown in (2) – seems to fit the data quite well, and this is consistent with the high  $R^2$  of the regression. The value of the intercept  $\alpha$  is -137,340, which is where the line intersects the left-hand vertical axis. The value of the slope  $\beta$  is 208.59, which implies that the value of a home rises by \$208.59 for every additional square foot that is added to the dwelling's total size.

Q: How is regression analysis used to compute the house price indices?

A: Very simply: I just create a variable that denotes during what time period a

house sale took place in, and the estimated coefficient on that variable indicates what the average appreciation rate was from the baseline period to that time period.

Let's continue with our previous example. The following figure shows a scatterplot of house prices and their square footage from two different periods of time: 1995 Q1 and 2005 Q1. Homes that sold in the first quarter of 1995 are color-coded in red, while those that sold ten years later in the first quarter of 2005 are color-coded in blue: Although the relationship between the size of a home and its price is still positive for both samples, it is easily observable that the homes in 1995 sold at generally lower prices than homes which sold in 2005. This can be inferred by observing that the red squares (1995 sales) typically lie below the blue diamonds (2005 sales).



To measure the rate of house price appreciation during this ten year period, the following variable is first defined:

$$= \frac{\text{Sale}_{2005}}{\text{Sale}_{1995}}$$

$$P_i =$$

0 if sale year

1 if sale year  
1995

(3)

2005

The variable "Sale2005" defined in (3) is known as a "dummy variable" because it can take on only one of two values: either a 0 or 1, depending upon which condition prevails. In this case, the variable is set to 0 if the property sold in 1995, and 1 if it sold in 2005. Adding this variable to the dataset would cause the file layout to look something like this: Having defined this variable, it is now a simple matter of adding this to the regression specification in (1) and re-estimating it. This re-estimation gives the following results:

i	P <sub>i</sub>	S <sub>i</sub>	Sale2005
1	\$85,000	1,100	0
2	\$100,000	1,250	0
3	\$125,000	1,300	1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
N	\$.750,000	.3,200	1

$$P_i = -162,592 + 171.63 \times S_i + 83,376 \times \text{Sale2005} \quad (4) \quad (-13.37) \quad (22.56) \quad (8.44)$$

No. of obs.=200, R<sup>2</sup>=0.7860

The estimated coefficient on Sale2005 gives the typical "dollar premium" that a home which sold in 2005 has over a comparable home which sold in 1995. Its value of 83,376 suggests that – in this small sample – homes which sold in 2005 sold at a price that was, on average, \$83,376 higher than homes which sold in 1995.

Since the average house price in 1995 was \$45,000, then the average rate of appreciation during this 10-year period (in this very small sample) is 185%. A simple house price index could then be created by setting the baseline value of the index to 100 in 1995, and then the value of the index would be 285 (=100\*(1+1.85)) in the year 2005.

However, there is a major problem with the regression estimated in (4). Namely, it is implicitly assuming that the marginal price of an additional square foot (\$171.63) is remaining constant over time. This is almost certainly not true, since, as the regression results show, house values (and thus the implicit value of their attributes) generally rise over time. So, for example, an additional square foot of space that would be worth \$100 in 1995 would almost certainly be worth significantly more in 2005. By imposing the constraint that marginal prices of various attributes remain constant over time, the estimated coefficients are likely to be misleading and erroneous.

This problem can be solved by converting the regression specification from one which measures price changes in dollars to one which measures price changes in percents. For example, although the dollar price of an additional square foot (or an extra bath, or an extra bedroom, etc.) may change dramatically over time, the percent change in price is likely to be more stable. That is, the additional “bump” in price due to an additional square foot is more stable when expressed as the percent increase in price rather than the dollar change in price.

This conversion can be accomplished by taking a logarithmic transformation of the regression specification, so that it now has the following form:

$$\ln(P_i) = \ln(\alpha) + \beta_1 \times \ln(S_i) + \beta_2 \times \text{Sale}_{2005_i} \quad (5)$$

Where:  $\ln(P_i)$  = the natural log of the price of the  $i$ th house  
 $\ln(S_i)$  = the natural log of the total square footage of the  $i$ th house  
 $i$  = an index of numbers denoting each unique property

Although the equation in (5) is still linear in its specification, it is nonlinear in its variables. Logarithmic transformations<sup>2</sup> have the unique property of converting a relationship from levels to percents. Previously, the estimated  $\beta$  coefficient on square footage used to state the effect that an increase in one square foot of house size would have on a house's price in dollar terms. But now the estimated  $\beta$  coefficient states the effect that an additional square foot would have on a

house's value in percent terms.

After taking the natural logs of the data and estimating the regression in (5), the following results are obtained:

$$\begin{aligned} \ln(P_i) = & -0.5411 + \\ & 1.5966 \times S_i + 0.7913 \times \text{Sale}_{2005} \quad (6) \\ & (-0.77) \quad (16.33) \quad (11.50) \quad N.=200, \\ & R^2=0.7232 \end{aligned}$$

The sample's average rate of house price appreciation from 1995 to 2005 can be retrieved from the estimated coefficient on Sale2005. Since this equation is in logs, the coefficient must first be exponentialized and have one subtracted to recover the appreciation rate. Since  $\exp(0.7913)-1=1.206$ , these regression results imply that homes which sold in 2005 did so at a price that was, on average, 120.6% higher than homes which sold in 2005. Note that this number is significantly less than the 185% rate of appreciation implied by the results in regression (5). This is because the erroneous assumption about the marginal

<sup>2</sup> Note: it is a mathematical axiom that if  $\ln(x)=y$ , then  $e^y=x$ , where  $e$  is the exponential constant: 2.718. price of an additional square foot being constant over time has been eliminated, thus facilitating a more accurate and less biased measure of house price appreciation.

If the sample were large enough so that home sales reflected the total underlying stock of housing, you could reasonably conclude that house values appreciated by 120.6% during this period, which means they slightly more than doubled in value over the 10 year period.

Finally, a house price index can be generated by applying this appreciation rate to whatever value you would choose for the baseline period. In this simple example, the baseline period is 1995 Q1. If we set the index to a starting value of 100 for this period, then the index would have a value of 220.6 ( $=100 \times \text{Exp}(0.7913)$ ) in 2005 Q1.

Q: What variables are in the full regression specification?

A: The estimated regression which is used to generate the house price indices is a regression of the natural log of house price on the following variables:

- Building square footage
- Lot square footage
- Density (Floor Area Ratio)
- Frontage and Depth
- Number of stories
- Number of fireplaces
- Type of exterior
- Physical condition of the exterior
- Attached, Semi-detached or Detached
- Located on a corner or interior of street block
- Qualitative characteristics: garage, irregularly-shaped plot, above street-level location, central air, improvements and amenities, etc.
- Renter-occupied v. Owner-occupied
- Distance to the Central Business District (as proxied by City Hall)
- Census tract
- Season when transaction took place: winter, spring, summer, fall.
- Number of years since last transaction

The qualitative variables--such as Census tract, garage or renter occupancy – are measured as dummy variables, which take a value of “1” if the dwelling possesses such a characteristic and zero otherwise. For example, if a property is located in Census tract 230, has a garage and is owner-occupied, then the variables “Tract230”, “Garage” and “Renter” are assigned the values “1”, “1” and “0”, respectively.

Q: Are all house sales in Philadelphia used to estimate the regression?

A: No. The indices are estimated using data that has first been put through a very rigorous and extensive cleaning process.

Many transfers of real estate are not “clean” or “arms-length” transactions occurring at market prices. In addition, there are a significant number of sales that are outliers; for example, their price may be significantly out of line with prevailing market values, the property may be physically damaged or not up to code, or the recorded information may simply be erroneous due to a clerical error.

The quarterly update of the house price indices begins by taking the entire population of all real estate transfers in Philadelphia during the most recent quarter, and then applying the following set of screening filters to drop undesirable transactions from the sample:

- 1) Drop Nominal sales: sales which took place at a “nominal” price of only \$1 or \$10. These are often sales between family members.
- 2) Drop Blanket sales: sales for which multiple properties trade hands from the same seller to the same buyer for the same “blanket” (i.e. comprehensive) price. Since the individual price of each property is not observed, they are eliminated from the sample.
- 3) Drop Sheriff sales: sales for which either the grantor or grantee is listed as the “Sheriff of Philadelphia”. These sales result from foreclosures, tax liens, tax delinquencies, and seizures due to illegal activity occurring on the premises.
- 4) Drop government sales: sales where the buyer or seller is a government entity. Many of these sales are due to condemnations or eminent domain acquisitions, so the recorded price was determined by a court or appraiser rather than the market. Also, since public sector entities do not generally behave with the same profit-maximizing behavior that private individuals and investors do, the prices of these sales are often not in line with market values. An observation is deleted if any party to the transaction is one of the following entities:
  - i) Secretary of HUD or Department of HUD
  - ii) Secretary of VA or Department of VA
  - iii) City of Philadelphia
  - iv) Office of the Philadelphia District Attorney
  - v) Philadelphia Housing Authority
  - vi) Philadelphia Redevelopment Authority
  - vii) Fannie Mae
  - viii) Freddie Mac

Note: transactions where either Fannie Mae or Freddie Mac is directly involved are usually foreclosures or disposition of REO.
- 5) Drop Speculative sales: sales in which the property previously transacted within the past 3 months. These types of sales usually result from speculators “flipping” properties.
- 6) Drop Outliers: sales in which either the price or physical attributes of the house are significantly out of line with the market’s averages. The set of decision rules used to determine whether or not a sale is an outlier is too extensive to list here. However, with every quarterly update the statistical distribution of variables like price, price per square foot, building square foot, lot square foot, and floor area ratio (=building square foot/lot square foot) among others are reviewed, and those observations that lie in the extreme tails of the distribution are dropped from the sample.

Q: How much does cleaning the data affect the final sample size used to

estimate the indices?

A: Philadelphia historically has between 4,000 and 7,000 house sales per quarter (although this has been trending upwards recently). The screening process described above usually eliminates somewhere between 10 and 15 percent of observations from the population of all house transfers. The overwhelming majority of observations which are dropped are either Blanket, Nominal or Sheriff sales. The number of sales that get eliminated for other reasons, such as being government sales, speculative sales or outliers are relatively small: usually less than 5 percent of the total sample.

Q: How are the house price indices computed from the regression results?

A: The indices are recovered from estimated coefficients on the time period dummies. Since the indices are updated quarterly, there is a vector of dummy variables set to a quarterly frequency, denoting the year and quarter during which a property sold. The following table shows the regression results for the time variables from the estimation of the full hybrid hedonic specification: A unique dummy variable for each year and quarter was created to denote when each house in the dataset transacted. Since (at the time of this writing) the data covers the period from 1980 to 2005 Q1, there are 101 dummy variables. So, for example, the variable `year_qtr_2` denotes the time period 1980 Q2, while `year_qtr_101` denotes the time period 2005 Q1.

					(6)
(1)	(2)	(3)	(4)	(5)	Index
Variable	Parameter Estimate	Standard Error	t Value	Pr >  t	100
<code>year_qtr_2</code>	-0.001819	0.02053	-0.09	0.8924	99.8
<code>year_qtr_3</code>	0.04652	0.02466	1.89	0.0592	104.8
<code>year_qtr_4</code>	0.06707	0.02441	2.75	0.006	106.9
<code>year_qtr_5</code>	0.06685	0.02676	2.5	0.0125	106.9
<code>year_qtr_6</code>	0.06254	0.0247	2.53	0.0113	106.5
<code>year_qtr_7</code>	0.09661	0.02511	3.85	0.0001	110.1
<code>year_qtr_8</code>	0.12852	0.02626	4.89	<.0001	113.7
<code>year_qtr_9</code>	0.06994	0.0283	2.47	0.0135	107.2
<code>year_qtr_10</code>	0.12771	0.02636	4.85	<.0001	113.6

year_qtr_11	0.0891	0.02523	3.53	0.0004	109.3
year_qtr_12	0.103	0.02596	3.97	<.0001	110.8
year_qtr_13	0.18255	0.02593	7.04	<.0001	120.0
year_qtr_14	0.16928	0.02436	6.95	<.0001	118.4
year_qtr_15	0.2118	0.02341	9.05	<.0001	123.6
year_qtr_16	0.23445	0.02438	9.61	<.0001	126.4
year_qtr_17	0.23312	0.0256	9.11	<.0001	126.3
year_qtr_18	0.27085	0.02331	11.62	<.0001	131.1
year_qtr_19	0.26931	0.02313	11.64	<.0001	130.9
year_qtr_20	0.28017	0.0237	11.82	<.0001	132.3
year_qtr_21	0.27373	0.02469	11.09	<.0001	131.5
year_qtr_22	0.33965	0.02331	14.57	<.0001	140.4
year_qtr_23	0.35259	0.02254	15.65	<.0001	142.3
year_qtr_24	0.34291	0.02275	15.07	<.0001	140.9
year_qtr_25	0.36817	0.02346	15.69	<.0001	144.5
year_qtr_26	0.45669	0.02273	20.09	<.0001	157.9
year_qtr_27	0.47334	0.02199	21.53	<.0001	160.5
year_qtr_28	0.47407	0.02257	21.01	<.0001	160.7
year_qtr_29	0.5083	0.02344	21.68	<.0001	166.2
year_qtr_30	0.52915	0.02249	23.53	<.0001	169.7
year_qtr_31	0.58691	0.02219	26.44	<.0001	179.8
year_qtr_32	0.5786	0.02249	25.73	<.0001	178.4
year_qtr_33	0.59491	0.02369	25.11	<.0001	181.3
year_qtr_34	0.60551	0.02165	27.97	<.0001	183.2

year_qtr_35	0.64819	0.02237	28.98	<.0001	191.2
year_qtr_36	0.66144	0.02316	28.56	<.0001	193.8
year_qtr_37	0.67207	0.02337	28.75	<.0001	195.8
year_qtr_38	0.69686	0.02285	30.5	<.0001	200.7
year_qtr_39	0.69633	0.02283	30.5	<.0001	200.6
year_qtr_40	0.70226	0.02286	30.72	<.0001	201.8
year_qtr_41	0.66157	0.02433	27.19	<.0001	193.8
year_qtr_42	0.71008	0.02299	30.89	<.0001	203.4
year_qtr_43	0.70071	0.02366	29.62	<.0001	201.5
year_qtr_44	0.67475	0.02397	28.14	<.0001	196.4
year_qtr_45	0.59117	0.02596	22.77	<.0001	180.6
year_qtr_46	0.64587	0.02385	27.08	<.0001	190.8

year_qtr_47	0.65654	0.02427	27.06	<.0001	192.8
year_qtr_48	0.59778	0.02569	23.27	<.0001	181.8
year_qtr_49	0.58277	0.02624	22.21	<.0001	179.1
year_qtr_50	0.63001	0.02457	25.64	<.0001	187.8
year_qtr_51	0.59568	0.02464	24.18	<.0001	181.4
year_qtr_52	0.5472	0.02601	21.04	<.0001	172.8
year_qtr_53	0.56324	0.02604	21.63	<.0001	175.6
year_qtr_54	0.54159	0.02576	21.03	<.0001	171.9
year_qtr_55	0.53693	0.02525	21.26	<.0001	171.1
year_qtr_56	0.53138	0.02626	20.24	<.0001	170.1
year_qtr_57	0.4879	0.02602	18.75	<.0001	162.9
year_qtr_58	0.51045	0.02437	20.95	<.0001	166.6
year_qtr_59	0.51289	0.02379	21.56	<.0001	167.0
year_qtr_60	0.56331	0.02082	27.06	<.0001	175.6
year_qtr_61	0.57831	0.02001	28.9	<.0001	178.3
year_qtr_62	0.61579	0.01953	31.53	<.0001	185.1
year_qtr_63	0.64837	0.01964	33.02	<.0001	191.2
year_qtr_64	0.62868	0.01985	31.67	<.0001	187.5
year_qtr_65	0.64036	0.0202	31.69	<.0001	189.7
year_qtr_66	0.67357	0.01956	34.43	<.0001	196.1
year_qtr_67	0.64242	0.01984	32.37	<.0001	190.1
year_qtr_68	0.67589	0.01954	34.59	<.0001	196.6
year_qtr_69	0.63732	0.02002	31.84	<.0001	189.1
year_qtr_70	0.65764	0.01982	33.19	<.0001	193.0
year_qtr_71	0.64755	0.01921	33.71	<.0001	191.1
year_qtr_72	0.65384	0.01972	33.15	<.0001	192.3
year_qtr_73	0.66359	0.01984	33.44	<.0001	194.2

year_qtr_74	0.71755	0.01939	37	<.0001	204.9
year_qtr_75	0.68504	0.01935	35.41	<.0001	198.4
year_qtr_76	0.68637	0.01959	35.04	<.0001	198.6
year_qtr_77	0.71068	0.01967	36.13	<.0001	203.5
year_qtr_78	0.75197	0.01918	39.2	<.0001	212.1
year_qtr_79	0.7484	0.01932	38.74	<.0001	211.4
year_qtr_80	0.75128	0.01954	38.46	<.0001	212.0
year_qtr_81	0.76969	0.01961	39.24	<.0001	215.9
year_qtr_82	0.78859	0.01916	41.15	<.0001	220.0

year_qtr_83	0.78931	0.01941	40.67	<.0001	220.2
year_qtr_84	0.80493	0.02014	39.97	<.0001	223.7
year_qtr_85	0.83397	0.01989	41.93	<.0001	230.2
year_qtr_86	0.91639	0.01956	46.84	<.0001	250.0
year_qtr_87	1.01212	0.01974	51.26	<.0001	275.1
year_qtr_88	1.0222	0.01991	51.35	<.0001	277.9
year_qtr_89	1.04038	0.02017	51.59	<.0001	283.0
year_qtr_90	1.00084	0.01908	52.46	<.0001	272.1
year_qtr_91	0.99402	0.01918	51.83	<.0001	270.2
year_qtr_92	1.02106	0.02001	51.03	<.0001	277.6
year_qtr_93	0.98146	0.01943	50.52	<.0001	266.8
year_qtr_94	1.03561	0.01924	53.83	<.0001	281.7
year_qtr_95	1.09708	0.01914	57.32	<.0001	299.5
year_qtr_96	1.11365	0.01931	57.67	<.0001	304.5
year_qtr_97	1.14165	0.01922	59.4	<.0001	313.2
year_qtr_98	1.22984	0.01896	64.88	<.0001	342.1
year_qtr_99	1.28838	0.01897	67.91	<.0001	362.7
year_qtr_100	1.32259	0.01911	69.2	<.0001	375.3
year_qtr_101	1.35016	0.01916	70.45	<.0001	385.8

The variable for the very first time period, 1980 Q1, is chosen as the omitted category from the regression specification. In this way, the estimated coefficients in column 2 give the average rate of house price appreciation from 1980 Q1 to that period. Column 3 provides the standard errors of each parameter estimate. Column 4 gives the t-values of each variable, while column 5 provides the probability statistics related to the test of statistical significance given by the t-values.

The t-value of an independent variable is computed as the ratio of the coefficient to its standard error. Its value states how “statistically significant” this variable is in explaining variation in the dependent variable. The larger the t-value (in absolute terms), the more significant is the independent variable to the model. It can be seen that the t-values of the variables generally increase as the number of quarters since 1980 Q1 increases. This is an intuitive result, since the further away from 1980 Q1 the more different – and thus statistically significant – are average house values from their value in the baseline period of 1980 Q1. For example, average house values in 1981 were not all that much different than average house values in 1980, but average house values in 2005 are much higher

than average house values in 1980. Consequently, the t-value for the variable `year_qtr_101` (which denotes 2005 Q1) is much more statistically significant than the t-value for, say, `year_qtr_5` (which denotes 1981 Q1).

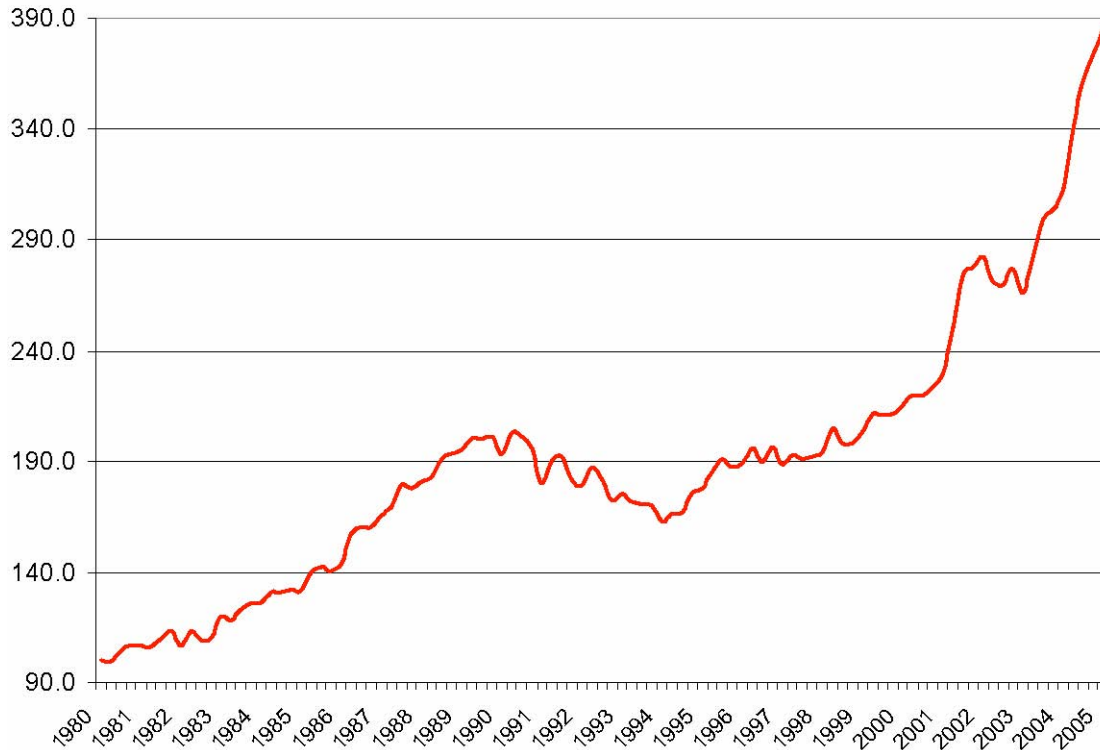
Exponentializing each of the parameter estimates and subtracting one<sup>3</sup> gives the mean house price appreciation rate since 1980 Q1 to that period. For example, since the estimated parameter coefficient for the period 2005 Q1 is 1.35016, then  $\exp(1.35016)-1=2.858$ . This implies that the typical Philadelphia home has appreciated by 285.8% since 1980, which is almost a tripling of the level of house values.

Finally, the Philadelphia house price index is computed from the parameter estimates by first setting the value of the index to a value of 100 in the base period 1980 Q1, and then multiplying 100 times the exponent of each coefficient. So, for example, the value of the index in 1990 Q1 (variable `year_qtr_41`) is  $100 \times \exp(0.66157)=193.8$ , while the value of the index in 2005 Q1 is  $100 \times \exp(1.35016)=385.8$ .

Figure 1 shows a plot of the index over time, from 1980 to 2005 Q1:

<sup>3</sup> This transformation is necessary because the dependent variable in the regression is the natural log of price, while the independent variables are dummy variables which only take values of 0 or 1. Since you can't take the log of zero, it's necessary to perform the exponential transformation. See Halvorsen and Palmquist (1980) for a more rigorous and detailed explanation of why this procedure is necessary.

**Figure 1. Philadelphia House Price Index 1980-2005:  
1980Q1=100**



Since housing is such a fundamental consumer good (everyone needs to live somewhere!), it is hardly a surprise that nominal house values generally increase over time, along with the rate of economic growth and inflation. This is consistent with the fact that the price index is generally trending upward. However, closer examination can also reveal how the rate of appreciation also systematically co-varies with both the national and local economy.

For example, the rate of appreciation was rather sluggish in both the early and mid-1980s, when the national economy was emerging from a recession and the local economy was suffering from continued population and job losses, as well as the scare of the MOVE bombing. Price appreciation picked up – along with the economy – in the late 1980s, but then a national recession compounded by the city’s fiscal crisis saw actual declines in property values in the early 1990s. A recovery began in the mid-1990s, but then suffered another period of declines following the post-dot-com recession and then 9/11. Recent years have seen the most sustained period of considerable price increases as interest rates, which are at all-time historic lows combined with uncompetitive returns in other asset markets (e.g. stocks and bonds), have incited a significant nationwide boom in house price appreciation.

Q: Cleaning all this data and running a regression seems like a lot of work. Isn't it easier to just look at the trends in average sales prices over time?

A: Although it is easier to look at trends in average house prices, it can often lead to misleading and even erroneous conclusions.

There are two primary reasons why the indices estimated via regression provide a "truer" picture of house price appreciation:

- 1) There is significant seasonality in house prices over time; and
- 2) There is significant sample selection bias in house sales over time.

The "seasonality" of house prices refers to the changes in house values that are related to the season of the year. In particular, homes which sell in the winter usually do so at a price discount relative to homes which sell during the warmer weather months. The reasons for this are twofold: the winter weather inhibits both house hunting and moving, and homes for which the trees and plants are fully foliated during the warm weather months have more visual appeal than homes with bare trees and flower beds during the winter. Both of these factors lead the consumers to successfully extract a price discount from sellers as compensation for buying a house in the winter (or conversely, to pay a price premium to sellers in the summer).

This phenomenon is hardly unique to house prices. Economists have been well aware of the presence of seasonality in the time series of many commodity and asset prices. For example, crop prices in many countries have historically exhibited seasonality due to the cyclicity of the growing season. Retail prices and volume exhibit seasonality due to the presence of the holiday season in December. Gas prices exhibit seasonality due to the fact that most people disproportionately choose to vacation and travel during the warmer summer months. There is even significant evidence that stock prices tend to get a boost in January, as tax-loss selling by investors in December aims to realize capital losses which are then used to offset taxes owed on capital gains.

The sample selection bias of house sales refers to the fact that, at any given time, not every home has an equal probability of transacting. The result is that the "flow" of homes which do sell may not represent the entire "stock" of all existing homes. As a result, making inferences about all house values based upon just house sales may result in misleading conclusions.

As a simple example, consider a particular scenario. Suppose that during a given year, house values increased 6%. However, during the first half of the year, only large homes in wealthy neighborhoods transacted. Then, during the second half of the year, only small homes in lower-income neighborhoods changed hands. If you were to compare average house prices during the course of the year, you could conclude that house values fell dramatically, even though in reality they had actually increased!

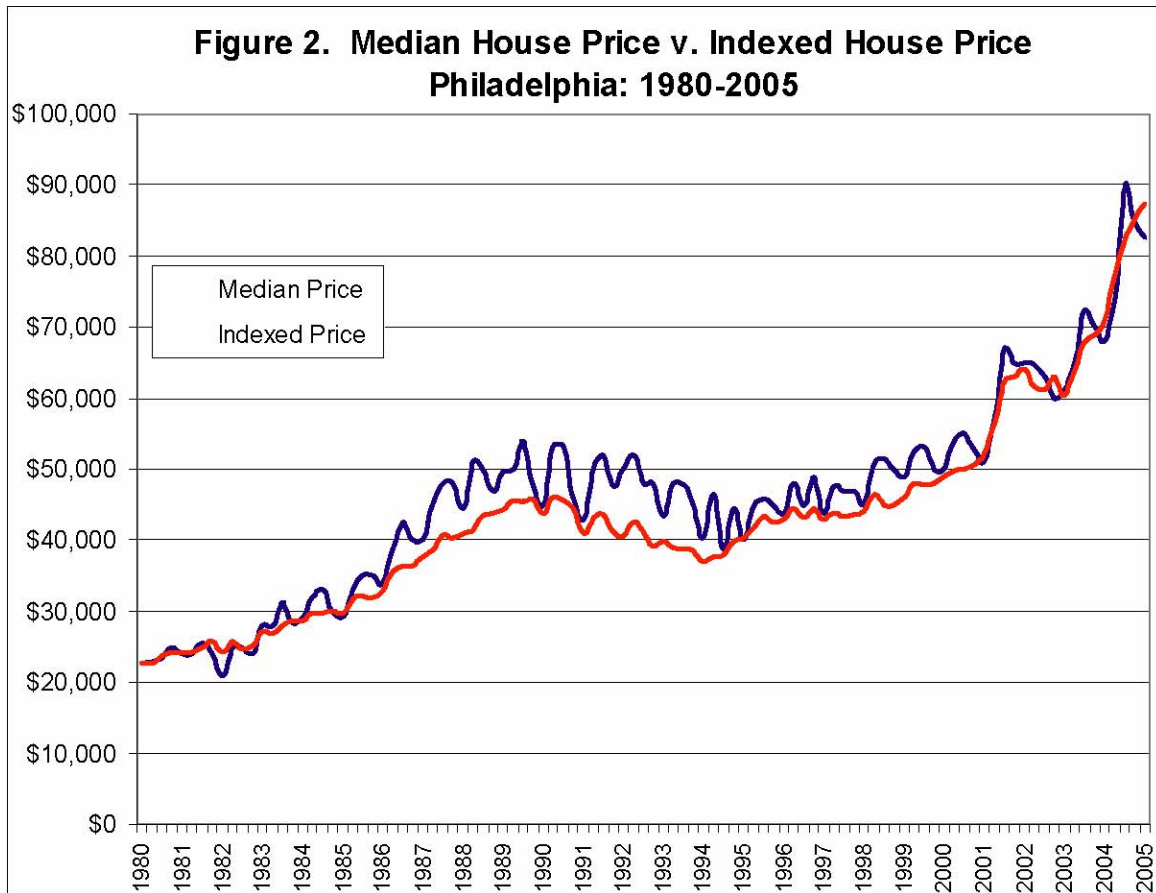
More realistically, however, it is a fairly well-established fact of housing markets that:

Homes which have appreciated relatively more rapidly in value are more likely to be on the market in the first place, and to sell quickly as well.

Homes which have lagged the market in appreciation or even gone down in value are less likely to sell, and those that do sell often spend relatively longer periods of time on the market than homes which have lead the market in appreciation.

Genesove and Mayer (2001) provide substantial empirical evidence in support of the positive correlation between price growth and sales volume in the housing market.

For proof of how the estimated indices control for these aforementioned problems, consider the following plot: This figure plots the median sales price of homes that sold in Philadelphia from 1980-2005 as the blue line. The red line illustrates the growth in house values according to the empirically estimated house price index from the regression results in Table 1<sup>4</sup>. By just casual visual observation, two differences between the different trends in house price appreciation are readily discernible:



The line representing the trend in median house prices is more volatile (or “squigglier”) than the line representing the empirically estimated house price index.

The line representing the trend in median house prices generally lies above the line representing the empirically estimated house price index.

The relative volatility of the median price trendline is due to the presence of seasonality in house prices: although the median house price is generally rising over time, the instability of the trend line over short horizons is a function of the

<sup>4</sup> The indexed rate of house price appreciation (represented by the red line in Figure 2) was computed by taking the median house price in 1980, and then applying the index to this price in order to “grow” the price by the same rate of appreciation as the index.

season when homes sold. As a result, the line oscillates around its long term trend, rising during warm weather months and then falling during cold weather ones.

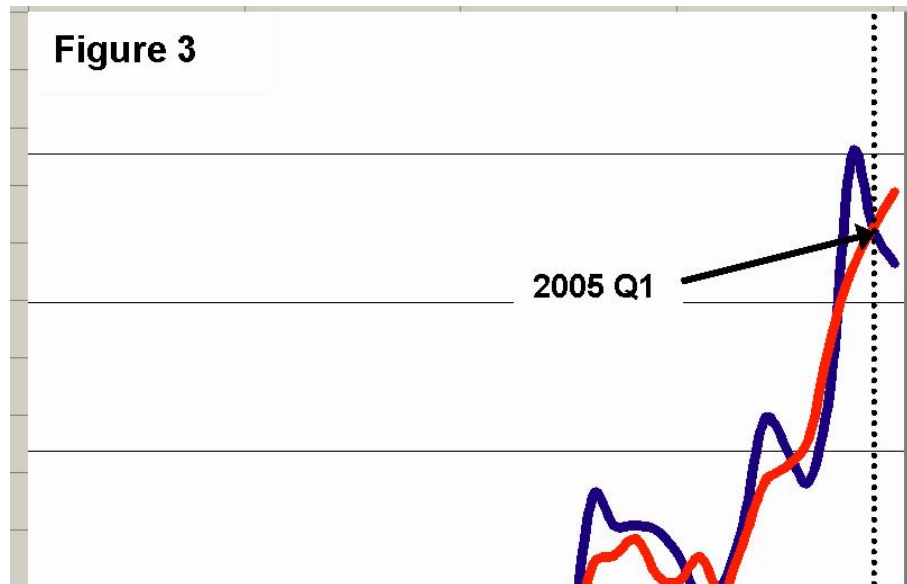
However, because the regression that estimated the house price index controls

for seasonality (by including dummy variables denoting the season of each transaction), it is a much smoother trend line. Thus the effect of “season” has been effectively eliminated so that the true rate of house price appreciation is uncovered and accurately measured.

The fact that the estimated house price index lies below the median price trendline is due to the aforementioned sample selection bias of homes that sell versus homes that do not. Homes that sell are also homes that are more likely to have recently experienced disproportionately higher rates of appreciation in value, which is why their owners are more incentivized to sell. Conversely, homes which have lagged the overall market in terms of appreciation are less likely to be for sale in the first place, and those that are sold spend disproportionately longer periods of time on the market, which translates into a lower overall turnover rate.

The result of this sample selection bias is that trends in median prices are biased upwards from true appreciation rates. But because the hybrid hedonic regression contains a vector of repeat-sales variables which control for this sample selection bias, the effect is to “deflate” this sample selection bias and uncover the true rate of house price appreciation.

Both of these effects have serious implications over short-term horizons. For example, suppose you were interested in knowing what the rate of house price appreciation was during the first quarter of 2005. To find out, you examine the changes in the two indices during this period: The above figure is a tight, close-up screen shot of the indices from Figure 2, with the dashed line delineating where the fourth quarter of 2004 ends and the first quarter of 2005 begins. Note that the two trend lines in house values move in opposite directions during the first quarter of 2005.



If you were to go by the change in median house prices (the blue line) you would conclude that house prices had declined during the first quarter. However, the estimated hedonic index (red line) indicates the exact opposite: house prices went up.

The reason for these contrary results is precisely because of the presence of the two aforementioned problems that plague analyses which only look at trends in mean or median prices. First, house price data is subject to seasonality. Since the first quarter of the year contains the winter months, prices are lower – on average – during this periods than they were in the relatively warmer months of the previous quarter. Second, house price data is subject to sample selection bias. Further investigation of the home sales data revealed that a disproportionate percentage of home sales in that period were from neighborhoods of Philadelphia which are historically lower-valued, but are currently gentrifying and experiencing revitalization. Hence, although they are still priced relatively lower than most Philadelphia homes, they are currently undergoing a relatively higher rate of price appreciation.

When combined, these two effects explain the contradictory results. The presence of seasonality explains why the trend in median house prices declined from the 4<sup>th</sup> quarter of 2004 to the 1<sup>st</sup> quarter of 2005, while sample selection bias in home sales explains why the empirically estimated index shows a positive increase in home values. The empirically estimated index would generally be considered the “truer” picture of trends in house values by most statisticians because it controls for these other factors which have nothing to do with the fundamental drivers of house values. However, if you were only to examine trends in mean or median house prices, you would be led to the exact opposite – and erroneous--conclusion that the estimated indices suggest!

Q: Can I use the indices to predict the current value of a house?

A: Yes. You can do this by simply applying the index to the original purchase price of the house.

In general, if you want to know the change in house prices between period "t" and period "t-k", where "t" is the current period and "t-k" is the period when the home last sold, you can use this formula:

$$\text{Pct. Change} = (I_t - I_{t-k}) / I_{t-k} \quad (7)$$

Where:  $I_t$  = the value of the index in time period t  $I_{t-k}$  = the value of the index in time period t-k

For example: the value of the citywide index in 1980 is 100, while the value of the index in the first quarter of 2005 is 385.8. The percent change in these two numbers from 1980 to 2005 Q1 is  $(385.8 - 100) / 100 = 2.858$ , or 285.8%. This implies that the average Philadelphia home appreciated in value by 285.8% during this 26-year time period. Then, add 1 to this number ( $1 + 2.858 = 3.858$ ) and multiply this times the purchase price of the property. Since the median house price of Philadelphia homes in 1980 Q1 was \$22,700, then the estimate of this home's current value would be  $\$22,700 \times 3.858 = \$87,577$ .

This same algorithm can be applied for any two time periods covered by the indices. For example, if you wanted to know the 1-year change in house values from 2004 Q1 to 2005 Q1, it is  $(385.8 - 312.6) / 312.6 = 23.4\%$ . This implies that the average Philadelphia home appreciated in value by 23.2% over the course of this particular 1-year time period. So, the indices would predict that a home which sold for \$64,000 in 2004 Q1 (the median sales price then) would be worth \$78,987 one year later.

However, be aware that this technique implicitly assumes that the dwelling has remained in constant-quality condition since its last sale. That is, the indices are only appropriate for making predictions if the dwelling is in essentially the same condition now as it was when it last sold. The indices do not make any allowances for either major improvements to a property or any significant depreciation. Naturally, if the owner added a new pool or patio, or conversely, if the property suffered fire or water damage, then the indices will not capture these effects. But if the property has been adequately maintained with no substantial changes (either positive or negative), then the implicit assumption of constant-quality will have been maintained. Under this assumption, the value of

the property will then simply move with that of the overall market; both locally and nationally, and the HPIs are a reasonable estimate of its change in value.

## References

Bailey, M.J., Muth, R.F. and Nourse, H.O. 1963. "A Regression Method for Real Estate Price Index Construction." Journal of the American Statistical Association. 58. 933-942.

Case, Bradford, Henry Pollakowski and Susan Wachter. 1991. "On Choosing Among House Price Index Methodologies." AREUEA Journal. 19. 286-307.

Case, K.E. and Shiller, R.J. 1987. "Prices of Single Family Homes since 1970: New Indexes for Four Cities." New England Economic Review. Federal Reserve Bank of Boston, Sep. 45-66.

Case, K.E. and Shiller, R.J. 1989. "The Efficiency of the Market for Single-Family Homes." The American Economic Review. 79. 125-137.

Court, Andrew T. 1939. "Hedonic Price Indexes with Automotive Examples." Dynamics of Automobile Demand. New York: The General Motors Corporation.

Genesove, David and Christopher Mayer. 2001. "Loss Aversion And Seller Behavior: Evidence From The Housing Market," The Quarterly Journal of Economics. 116(4). 1,233-1,260.

Halvorsen, Robert and Raymond Palmquist. 1980. "The Interpretation of Dummy Variables in Semilogarithmic Equations." American Economic Review. 70(3). 474-475.

Rosen, Sherwin M. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." Journal of Political Economy. 82(1).