Sample Selection Bias in Estimating Housing Sales Prices

G. Donald Jud*
Terry G. Seaks**

Abstract. This paper explores the bias in estimates of housing price appreciation that arises because of problems of sample selection. As suggested by Haurin and Hendershott, sample selection potentially is a serious problem because studies normally are based on samples of only homes that have sold, not all houses. Using the procedure developed by Heckman, the empirical results of this study provide confirmation of the significance of the housing sample selection problem.

Introduction

A number of recent studies have explored various methodologies for the construction of quality-adjusted housing price indexes as a means of estimating the rate of inflation in housing prices. Works in this area include the papers by Case and Shiller (1989), Case and Quigley (1991), Case, Pollakowski and Wachter (1991), Clapp and Giaccotto (1992), Abraham and Schauman (1991), and Haurin, Hendershott and Kim (1991) among others. Haurin and Hendershott (1991) review this literature, and in their paper they raise a troubling question. They observe (p. 264) that real estate price indexes may be subject to sample selection bias because they:

... are based on samples of only sold properties, not all houses. The potential bias may be largest during economic downturns when few houses sell. Whether the units that sell have appreciated relatively rapidly, slowly, or by a representative amount is not generally known. Further analysis is required involving tests for sample selection bias and correction of the hedonic estimation. (emphasis added)

This paper examines exactly this question of sample selection bias in real estate price indexes.1 Estimation of the housing price index is accomplished using the assessed value method advanced by Clapp and Giaccotto (1992). We extend Clapp and Giaccotto's analysis by estimating their basic regression model as the second stage of a sample selection model in which the first stage corrects for the possible selectivity bias.

The first section of the paper sets forth the sample selection model. Section two discusses the empirical data. Estimates of the two-stage sample selection model are

*Department of Finance, University of North Carolina at Greensboro, Greensboro, North Carolina 27412.
**Department of Economics, University of North Carolina at Greensboro, Greensboro, North Carolina 27412.
Date Revised—June 1994; Accepted—June 1994.
presented in the third section of the paper. The final section discusses the implications of the findings for further studies of real estate price indexes.

Estimation of a Real Estate Price Index

The model employed here to estimate the housing price index is the assessed value (AV) method advanced by Clapp and Giaccotto (C&G). The basic equation for the AV method can be written as:

\[
\ln P_{it} = \beta_0 + \beta_1 \ln(A_{it}) + \delta_1 Y_{1it} + \delta_2 Y_{2it} + \ldots + \delta_T Y_{T_it} + \epsilon_{it},
\]

where \( P_{it} \) = transaction price of the \( i \)th house at time \( t \) (\( i = 1, \ldots, n_t \), and \( t = 1, \ldots, T \)); \( A_{it} \) = assessed value of property \( i \) at time \( t \); \( Y_T \) = a time dummy with values of 1 if the \( i \)th house sold in period \( t \) and 0 otherwise; and \( \epsilon_{it} \) = random error term. The regression coefficients \( \delta_1, \delta_2, \ldots, \delta_T \) represent the logarithm of the cumulative price index. Although random errors in assessed value (measurement error) may pose a potential problem for the AV method, C&G offer convincing evidence that measurement error bias goes to zero in large samples.

The C&G study compares the AV method to the repeat sales (RS) methods used by Case and Shiller (1989) among others. C&G (page 305) show that "the AV and RS methods give substantially the same estimates of price trends over a five- to seven-year period. However, the AV method database is relatively easy to obtain, and the data have been cleaned by state agencies." In estimating our empirical results, we employ a sample of such "clean" data that was furnished us by the Guilford County Tax Office in Greensboro, North Carolina.

A potential problem with both the AV and the RS methods is that they may suffer from sample selection bias as Haurin and Hendershott warn. To examine the sample selection problem, we assume that there exists some latent selection variable \( z^* \) that determines whether or not a property is sold during a particular time period. Generally, \( z^* \) cannot be observed; rather, only the sign of \( z^* \) can be inferred. If a property is sold, then \( z^* \) is assumed to be positive and \( z \) takes on the value 1. If a property is not sold, then \( z^* \) is zero or negative and we observe \( z = 0 \). Thus, the probit model that governs the sale can be written as follows:

\[
\begin{align*}
z^*_{it} &= \gamma'w_{it} + u_{it}, \quad u_{it} \sim N[0,1] \\
z_{it} &= 1 \quad \text{if } z^*_{it} > 0, \\
z_{it} &= 0 \quad \text{if } z^*_{it} \leq 0, \\
\text{Prob}(z_{it} = 1) &= \Phi(\gamma'w_{it}), \\
\text{Prob}(z_{it} = 0) &= 1 - \Phi(\gamma'w_{it}),
\end{align*}
\]

where \( \Phi \) denotes the cumulative normal distribution.

The AV model of real estate prices can be simplified and written as:

\[
p_{it} = \beta'x_{it} + \epsilon_{it},
\]

where \( (u_{it}, \epsilon_{it}) \) are \( N[0,0,1,\sigma^2,\rho] \) and \( \rho \) is the correlation of \( p \) and \( z \). The crucial point
made by Haurin and Hendershott is that this equation is only observed if $z = 1$. The implication of this important observation is that estimation of the coefficient vector $\beta$ of (3) without a correction for the sample selection bias will result in regression estimates that suffer from the equivalent of omitted variable bias. That is, the estimates of $\beta$ will be biased and inconsistent if allowance is not made for the fact that the sample of houses that are sold may differ systematically from houses that are not sold.

If it is assumed that $z_u$ and $w_u$ are observed for a random sample of properties, but $p_u$ is observed only when $z_u = 1$, then the model can be written as:

$$E [p_u \mid z_u = 1] = \beta' x_u + \rho \sigma \lambda (\gamma' w_u),$$

(4)

where $\lambda (\gamma' w_u)$ is the inverse Mills ratio given by $\phi(\gamma' w_u)(1 - \Phi(\gamma' w_u))$ and $\phi$ and $\Phi$ denote the normal density and distribution functions, respectively. The presence of the variable $\lambda (\gamma' w_u)$ in equation (4) reveals the omitted variable bias that will result if equation (3) is estimated from only the houses that are sold. This is the problem Haurin and Hendershott suspect may be present, and our results below confirm it is indeed a problem.

Sample Data

Our database contains every parcel of single-family, residential property in the City of Greensboro, North Carolina. The data are drawn from the master appraisal file maintained by the Guilford County Tax Office in Greensboro, North Carolina. The database shows the appraised value of the property in 1986 and the prices paid the last three times the property sold between 1980 and 1991. If a property sold more than once during 1980–91, separate observations are included for each sale. Properties that were materially altered between the date of sale and the date of appraisal were eliminated from the database by the County Tax Office, as were some few properties selling for less than $10,000 or more than $999,999.

Exhibit 1 reports the sample means and standard deviations for the sample. There are a total of 376,671 observations. This total is comprised of every parcel of residential property in the city multiplied by the number of years the property was carried in the appraisal file. The number of properties in the file changes each year as new homes are built and others are destroyed. On average over the twelve-year period, there were 31,389 residential properties in the city. The means for the neighborhood dummy variables shown in the exhibit reflect the fraction of properties in each section of the city.

The housing price regression equations developed in the second stage of the analysis are estimated with a total of 23,095 property sales, occurring during the twelve-year sample period 1980–91. On average, there are 1,925 sales per year, representing about 6.1% of the total properties in the city. The means for the year dummy variables in Exhibit 1 show the fraction of all sales that were recorded during a particular year.

The average value of all houses in the city is somewhat lower than the average value of the houses that sold. This is seen by comparison of the two sample means in the exhibit. The mean of the log of tax value in the first-stage sample is 11.1030 (or
$66,370), while that in the second-stage sample is 11.2670 (or $78,198). The difference is evidence of possible selectivity bias in the two samples.

### Empirical Estimates

Consistent estimates of equation (4) can be obtained using Heckman's (1979) two-step procedure. In the first stage, a probit equation is estimated by maximum likelihood to obtain consistent estimates of $\gamma$. In the second stage, least squares regression is used to obtain consistent estimates of $\beta$ and $\rho \sigma^2$ by a regression of $p$ on $x$ and $\lambda$. The results of the estimated selectivity model are shown in Exhibits 2 and 3.

The probit model (Exhibit 2) relates whether a house is sold ($z_{it} = 1$) to a constant and seven independent variables: the log of assessed value for the $i$th house, the local-area unemployment rate in period $t$, the percentage change in the real value of residential building permits in period $t$ in the local area, the national conventional mortgage rate in period $t$, and a dummy variable for whether the $i$th observation was in the northeast, southeast, or southwest section of Greensboro. The log of assessed value is included in the model to allow the probability of sale to vary with the value of the house. It is expected that the likelihood of sale is not the same for homes at all price levels. Three variables, representing 1) the local-area unemployment rate, 2) the
Exhibit 2
Probit Selection Equation

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Marginal Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.0260</td>
<td>.0957</td>
<td>—</td>
</tr>
<tr>
<td>In (Tax Val.)</td>
<td>.2188</td>
<td>.0081</td>
<td>.0252</td>
</tr>
<tr>
<td>Unemp. Rate</td>
<td>-.0416</td>
<td>.0038</td>
<td>-.0048</td>
</tr>
<tr>
<td>% Chg. Bldg. Pmnts.</td>
<td>.0042</td>
<td>.00013</td>
<td>.0005</td>
</tr>
<tr>
<td>Mortgage Rate</td>
<td>.0234</td>
<td>.0026</td>
<td>.0027</td>
</tr>
<tr>
<td>NE Neighborhood</td>
<td>-.0743</td>
<td>.0125</td>
<td>-.0086</td>
</tr>
<tr>
<td>SE Neighborhood</td>
<td>-.2341</td>
<td>.0148</td>
<td>-.0270</td>
</tr>
<tr>
<td>SW Neighborhood</td>
<td>-.0437</td>
<td>.0094</td>
<td>.0050</td>
</tr>
</tbody>
</table>

n = 376,671
log(L) = -84941
\( z^2(7) = 3.8136 \)
\( \hat{\psi}(\gamma \mu) = .115 \)

Note: The marginal effects show the estimated value of \( \hat{\partial}E(z) / \hat{\partial}w_j = \hat{\psi}(\gamma \mu) \gamma_j \) with \( w \) evaluated at the means.

Exhibit 3
Housing Price Regressions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample Selection</th>
<th>Classical Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard Error</td>
</tr>
<tr>
<td>Constant</td>
<td>1.9808</td>
<td>.2289</td>
</tr>
<tr>
<td>In(Tax Val.)</td>
<td>.7956</td>
<td>.0116</td>
</tr>
<tr>
<td>Y1</td>
<td>.0206</td>
<td>.0160</td>
</tr>
<tr>
<td>Y2</td>
<td>.0904</td>
<td>.0153</td>
</tr>
<tr>
<td>Y3</td>
<td>.1230</td>
<td>.0204</td>
</tr>
<tr>
<td>Y4</td>
<td>.1964</td>
<td>.0186</td>
</tr>
<tr>
<td>Y5</td>
<td>.2875</td>
<td>.0167</td>
</tr>
<tr>
<td>Y6</td>
<td>.4658</td>
<td>.0152</td>
</tr>
<tr>
<td>Y7</td>
<td>.5313</td>
<td>.0151</td>
</tr>
<tr>
<td>Y8</td>
<td>.5725</td>
<td>.0158</td>
</tr>
<tr>
<td>Y9</td>
<td>.5769</td>
<td>.0160</td>
</tr>
<tr>
<td>Y10</td>
<td>.5901</td>
<td>.0164</td>
</tr>
<tr>
<td>Y11</td>
<td>.6135</td>
<td>.0156</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>-.1299</td>
<td>.0501</td>
</tr>
</tbody>
</table>

n = 23,095
\( \hat{\sigma}_k = .449 \)
\( R^2 = .570 \)

percentage change in building permits in the area, and 3) the national mortgage rate, are included to capture the impact of local and national economic conditions on the probability of sale. We expect that heightened economic activity increases the probability of a housing sale. Lastly, the neighborhood dummy variables are included to allow the likelihood of sale to differ among neighborhoods in the city.
Although it is hardly surprising for such a large sample, the probit model is highly significant as shown by the value of \( \chi^2 \) for testing the hypothesis that all seven slopes in \( \gamma \) are simultaneously zero.

The small standard errors of the estimated individual coefficients further suggest that each of the independent variables significantly influences the probability that a particular property will be selected for sale. As expected, the probability of sale is not the same for houses at all price levels, but appears to rise with the assessed value of the house. This suggests that more expensive (and most likely newer) properties are more likely to sell.

Also, as expected, the economic variables indicate that increased economic activity raises the probability of sale. The positive sign on the mortgage rate variable is problematic. Our finding of a positive sign on the mortgage rate suggests that our mortgage measure is less a "price" variable and more a measure of increased economic activity. We would not suggest that higher rates increase sales, but our objective here is not so much with the individual determinants of a sale as with the overall question of whether a sale influences price in the second equation. Finally, the significance of neighborhood dummy variables underscores the importance of neighborhood in influencing the likelihood of sale.

The column of marginal effects in Exhibit 2 shows the estimated change in the probability of sale for a one-unit change in the associated independent variable. For example, a one-percentage point increase in the unemployment rate reduces the probability of sale by approximately one-half a percentage point (.0048). Given that approximately 6% of the houses were sold (23,095/376,671), the unemployment rate obviously has a relatively large impact on the probability of sale. Also one neighborhood region was associated with a substantially lower probability of sale: in southeast Greensboro, the probability of a sale was reduced by 2.7 percentage points. Property values are lowest in this neighborhood, and this result is consistent with the experience of local realtors.

The estimated price equation is presented in Exhibit 3. The year dummy variables corrected for selection bias (equation 4) are shown alongside estimates produced using the classical regression method (equation 3). The most noteworthy feature of Exhibit 3 is that it provides a direct test of the null hypothesis that the coefficient of \( \lambda(y'w) \) is zero (i.e., the nonexistence of selection bias). The small size of the standard error relative to the estimated coefficient (asymptotic t-ratio of -2.6) suggests that the null hypothesis be rejected at reasonable probability levels. There is a fairly strong negative correlation between the sale price and the selection mechanism \( \hat{\rho} = -0.289 \). Thus, our results confirm the suspicions of Haurin and Hendershott: selection bias is a potential problem in estimating real estate price indexes.

The sign of the estimated coefficient of \( \lambda \) shown in Exhibit 3 provides an estimate of the direction of the sample selection bias. A negative coefficient implies that the prices of houses that sold are lower than those of equivalent houses that did not sell. This is easily seen from equation (4) which gives the expected price conditional upon sale as \( E[p_u | z_u = 1] = \beta'x_u + \rho \sigma_{\lambda} \lambda(y'w) \). Since the estimated coefficient of \( \lambda(y'w) \) is -1.299, the expected price conditional upon sale is lower than the expected price conditional upon no sale.

Notice from Exhibit 1 that the mean log tax value for the houses that sold of 11.267 ($78,200) exceeds the mean log tax value of all houses 11.103 ($66,400). This
Exhibit 4
Housing Prices in Greensboro
(1980=100.0)

suggests that the higher priced homes were those that were more often offered for sale, and this is consistent with the pattern of neighborhood sales activity most often observed in Greensboro. However, the negative estimated coefficient of \( \lambda(y'w_n) \) implies that the mean of the houses that sold would have been even higher were it not for the selection process that determined which homes were offered and sold.\(^{10}\)

Some measure of the magnitude of the selection bias can be obtained by comparing the size of the dummy variable coefficients estimated using the sample selection model to those produced by the classical regression procedure. In almost every case, the classical regression coefficients are larger than those using the sample selection procedure (see Exhibit 4).

Looking at Exhibit 5, column 7 shows the absolute difference between the classical regression index (column 2) and the selectivity-corrected index (column 5). The average yearly difference between the two indexes is 2.2 index points. Over the twelve-year period, the classical regression index estimates an inflation rate of 85.3%, while the corrected index indicates a rate of 84.7%, a difference of .6 percentage points.

Of course, the magnitude of the difference depends on where the comparison begins and ends. The yearly differences between the annual rate of price increase calculated using the classical regression index (column 3) and the rate tabulated with the selectivity-adjusted index (column 6) are shown in column 8. The mean value of the percentage differences shown in column 8 is less than .05%, indicating that the errors
### Exhibit 5

**Housing Price Indexes, 1980–91**

<table>
<thead>
<tr>
<th>Year</th>
<th>Classical Regression</th>
<th></th>
<th></th>
<th>Sample Selection</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>.0000</td>
<td>100.0</td>
<td></td>
<td>.0000</td>
<td>100.0</td>
<td></td>
<td>.00</td>
</tr>
<tr>
<td>81</td>
<td>.0323</td>
<td>103.3</td>
<td>3.3</td>
<td>.0206</td>
<td>102.1</td>
<td>2.1</td>
<td>.12</td>
</tr>
<tr>
<td>82</td>
<td>.0878</td>
<td>109.2</td>
<td>5.7</td>
<td>.0904</td>
<td>109.5</td>
<td>7.2</td>
<td>-.3</td>
</tr>
<tr>
<td>83</td>
<td>.1614</td>
<td>117.5</td>
<td>7.6</td>
<td>.1230</td>
<td>113.1</td>
<td>3.3</td>
<td>4.4</td>
</tr>
<tr>
<td>84</td>
<td>.2304</td>
<td>125.9</td>
<td>7.1</td>
<td>.1964</td>
<td>121.7</td>
<td>7.6</td>
<td>4.2</td>
</tr>
<tr>
<td>85</td>
<td>.3142</td>
<td>136.9</td>
<td>8.7</td>
<td>.2875</td>
<td>133.3</td>
<td>9.5</td>
<td>3.6</td>
</tr>
<tr>
<td>86</td>
<td>.4834</td>
<td>162.2</td>
<td>18.4</td>
<td>.4658</td>
<td>159.3</td>
<td>19.5</td>
<td>2.8</td>
</tr>
<tr>
<td>87</td>
<td>.5434</td>
<td>172.2</td>
<td>6.2</td>
<td>.5313</td>
<td>170.1</td>
<td>6.8</td>
<td>2.1</td>
</tr>
<tr>
<td>88</td>
<td>.5821</td>
<td>179.0</td>
<td>3.9</td>
<td>.5725</td>
<td>177.3</td>
<td>4.2</td>
<td>1.7</td>
</tr>
<tr>
<td>89</td>
<td>.5860</td>
<td>179.7</td>
<td>.4</td>
<td>.5769</td>
<td>178.1</td>
<td>.4</td>
<td>1.6</td>
</tr>
<tr>
<td>90</td>
<td>.5999</td>
<td>182.2</td>
<td>1.4</td>
<td>.5901</td>
<td>180.4</td>
<td>1.3</td>
<td>1.8</td>
</tr>
<tr>
<td>91</td>
<td>.6170</td>
<td>185.3</td>
<td>1.7</td>
<td>.6135</td>
<td>184.7</td>
<td>2.4</td>
<td>.6</td>
</tr>
</tbody>
</table>

% Chg. 1980–91: 85.3%  
Ave. Difference, 1981–91 2.2  .0%  
Ave. Difference without regard to sign, 1981–91 2.2  1.0%

in the inflation rate tend to cancel out over the sample period. However, a better measure of the bias in the annual inflation rate is found by calculating the mean value of column 8 without regard to sign. This average of the absolute values in column 8 is 1.0%, which indicates that ignoring the sample selection problem leads to an average error of 1 percentage point in annual estimates of housing price inflation. From a practitioner’s standpoint, a seemingly small difference in the rate of housing price appreciation can make a substantial difference in the estimated value of the housing stock or the tax base.

### Conclusions

This paper provides evidence of the bias imparted to estimates of housing price appreciation by ignoring the problem of sample selection. The bias arises because price indexes are normally based on samples of only houses that have sold, not all houses. Our empirical results provide confirmation of the significance of the bias problem as suggested by Haurin and Hendershott.

To analyze the sample selection problem, we employ the two-stage procedure developed by Heckman. Estimates of the first-stage sample selection equation confirm that the likelihood of sale is influenced by area-wide economic conditions as well as neighborhood effects. Second-stage estimates of the price index equation reveal the significance of the potential bias. Using our sample drawn from house sales in Greensboro, North Carolina during 1980–91, we find that the standard regression procedure used to estimate yearly price changes yields values that bias the yearly rates
of price appreciation by an average of 1 percentage point over the eleven-year period of our sample. While the cumulative effect of the errors tends to cancel out over time, in any given year the differences in our sample range from .1% to over 4%.

Notes

1The problem of selectivity bias is explored also in a recent paper by Gatzlaff and Haurin (1993) using data for the Miami area.
2Greene (1993, ch. 22) provides an excellent treatment of the sample selection model.
3Because \( \text{Prob}(z_u^* > 0) = \text{Prob}(\gamma' w_u + u > 0) = \text{Prob}(u > -\gamma' w_u) \) and the normal distribution is symmetric, it follows that \( \text{Prob}(z_u^* > 0) = \text{Prob}(u < \gamma' w_u) = \Phi(\gamma' w_u) \) which gives the probability \( z_u = 1 \).
4In 1990, Greensboro had a population of 183,521.
5LIMDEP 6.0 was used for all the computations. The analysis of 376,671 observations proved to be easier than we might have expected. The entire computing time on a VAX 6000 was just under 30 minutes of cpu time.
6This four-part breakdown of the city is widely used by city officials and the local realtors.
7An analogy might be drawn to the problem of multicollinearity in classical regression. A wrong sign on an individual coefficient is not necessarily indicative of an equation that forecasts badly. Here we are mainly concerned with using the predicted probabilities from the probit model to correct the housing price equation for selection bias.
8Because probit fits a nonlinear model, the marginal effect is not constant as in classical regression and caution is necessary in interpreting marginal effects for dummy variables where mean values may be difficult to interpret. See Greene (1993, p. 641) for details.
9While this paper was under review, we became aware of Gatzlaff and Haurin (1993). Using a data set from the Miami SMSA, they were able to estimate a different selection equation by year and found that the coefficient of \( \lambda \) often varied significantly year by year.
10The authors are indebted to Professor Gerald Makepeace (University of Hull) for discussions on this point. Interested readers are referred to Dolton and Makepeace (1987) for an excellent discussion of the interpretation of the selectivity coefficient.

References


The authors gratefully acknowledge helpful discussions with Gerald Makepeace, Christopher Ruhm, and two anonymous referees. We are also indebted to Mr. Jenks Crayton of the Guilford County Tax Office for help in obtaining and interpreting county tax data. All errors are our responsibility.