CHAPTER IV: ANALYSIS OF RESIDUALS

In the previous chapters we describe renter and owner hedonic models and present our price estimates of those models. In this chapter we evaluate the validity of the estimated equations through an analysis of their residuals. The analysis is performed in two parts. In the first section we analyze the general pattern of estimated residuals in all owner and renter equations to see if they are consistent with our model specification. In the second section we analyze a small subset of estimated residuals that lie outside the expected pattern. These outlying residuals are examined in detail for three renter models and two owner models.

This chapter's analyses find that the observed residuals' patterns are generally consistent with the model specifications, although tests for normality of the distributions are rejected in most cases. The residuals are symmetrically clustered about zero in all owner and renter models. Typically, half the estimated residuals are within a range of .263 for renters and .328 for owners. Since these ranges are centered on zero this means that half the predicted values lie between a plus or minus 14 percent of median rent and a plus or minus 17 percent of median value. When observed residuals are plotted against the predicted value of the dependent variable they show roughly a constant variance in the renter model. However, in the owner model they show a definite tendency to cluster more tightly about zero as predicted values increase. More estimated residual get classified as

outliers than would be expected from a normal distribution of error terms in both renter and owner models. The most notable feature of these outliers is that three-fourths of them are negative in both the owner and renter models. Still, fewer than one percent of the observed residuals are classified as outliers in almost all models, and the other 99 percent are approximately symmetric. Consequently, we conclude that the validity of the model and of the t-tests and F-tests is adequately supported by the observed pattern of the residuals.

Even though the general validity of the model is supported by the estimated residuals, we think the disproportionate share of negative outliers raises questions that deserve further investigation. Our analysis of outliers in five models finds most negative outliers to have reported rents and house values at the low end of their respective rent and value distributions in spite of fairly typical distributions of those dwelling characteristics included in the model. Attempts to alter model specifications to accomodate these outliers have been largely unsuccessful in bringing the observations back into line with other residuals. As might be expected, deletion of the outliers leads to substantial reductions in equation and coefficient standard errors, and large changes in a few coefficients. The negative outliers indicate that predicted rents and values constructed from our models are likely to have a downward bias. In a forthcoming paper predicting Fair Market Rents using our equations we will be able to make a limited investigation of the severity of this bias on the predictions.

Residuals in the Fifty-Nine Renter and Homeowner Models

In this section we analyze the residuals for fifty-nine renter and fifty-nine homeowner models. For each equation, the distribution of the residuals is examined for symmetry, for clustering and for the existence of outlying observations. When the data fits the model the residuals are symmetric about zero with few outliers. It is important to examine the residuals for symmetry since standard hypothesis testing using the estimated coefficients assumes the residuals are normally distributed. While a random variable that is symmetric about its average value does not imply that the variable is normally distributed, studies of the t-test find that test to be generally robust to the normality assumption as long as the underlying distribution is bell-shaped.[1] Since the hypothesis that our estimated residuals come from a normal distribution can be rejected in most owner and renter models, the questions of symmetry become very important in interpreting our test statistics.

Values of the residuals that are significantly far from their expected value of zero are labeled outliers. An outlier is an indication that for one reason or another the data may not fit the model. There are several reasons for an inadequate fit, however. Respondents may report inaccurate values for some questions or correct responses may be incorrectly transcribed to the data tape. Also, a cluster of outliers with similar underlying data characteristics is an indication that relevant variables are omitted from the equation. The existence of a large number of outliers is contrary to the normality assumption and therefore affects statistical hypothesis testing. In addition,

1. See Theil (1971), pp. 615-16.

the existence of outliers tends to inflate the estimate of the residual variance. Since the estimate of the residual variance is used in hypothesis testing and in the calculation of confidence intervals, outliers reduce the significance level of hypothesis tests and yield wide confidence intervals. The magnitudes of estimated coefficients are likely to be disproportionately affected by outliers as well.

Since we are estimating the same model in fifty-nine SMSAs except for locational variables, it is possible to compare the distributions of the residuals across SMSAs. Two comparisons are undertaken for each tenure type. First, we compare the spread of the residuals for each equation to the average or typical residual spread. We indicate the fitted equations with an unusually large or unusually small residual spread. We also look for similarities among the distributions of residuals by location or by size of the SMSA. Second, we identify outliers for each of the estimated regressions and find a consistent pattern of negative values among the equations.

Before proceeding to the analysis of the residuals we introduce the statistics that will be used. To avoid some of the problems caused by outliers we use order statistics to analyze the residuals. Order statistics are based on the rank of the numerically sorted data. The median, defined as the value for which half of the observations have smaller values, is a familiar order statistic. The closeness of the median value to zero gives one indication of symmetry centered about zero since half will be above and half below the median and since the mean of the residuals is constrained to be zero. To estimate the

spread or clustering of the residuals we use the difference between the third and first quartiles, called the interquartile range (IQR), of the residuals. The first and third quartiles are similar to the median of the distribution but have the property that one-fourth and three-fourths of the observations have smaller values, respectively. Unlike the usual estimate of the residual variance, the sum of squared residuals divided by the number of degrees of freedom, an outlier will not inflate the estimate of the interquartile range of the residuals. The similarity of the first and third quartiles in absolute value provides an indication of symmetry of the residuals about zero which is independent of outliers and of the median.

A residual is called an outlier if its value lies much below or above the values of most residuals. We classify a residual as a negative outlier if its value is three or more IQRs less than the first quartile of the residuals. Similarly a positive residual is considered an outlier if its value is three or more IQRs above the third quartile. This definition of an outlier is similar to the idea that a random variable does not conform to a hypothesized distribution if it is several standard deviations away from its expected value. The number obtained by computing the first quartile minus three times the IQR is called the lower fence. The corresponding upper fence is the third quartile plus three times the IQR. The concept of fences using order statistics is similar to the concept of a confidence interval. The probability of an observation appearing outside these fences, assuming a normal distribution for the residuals, is less than .0001. Out of about 250,000 residuals in the 118 equations we have estimated, at most

three outliers would be expected to occur. To reiterate, the advantage of using order statistics is that outliers do not inflate the estimate of the residual variance. Similarly, the calculation of the lower and upper fences is not significantly affected by the presence of outliers.[1]

In addition to the above order statistics, the analyses of this section makes use of stem and leaf plots for comparing renter or owner statistics among SMSAs. The method for reading these plots has been discussed in Chapter III.

We begin our residuals analysis with a listing of several statistics for renter and owner equations in every SMSA. Exhibits 13 to 18 list the number of residuals, their median value, the IQR of the residuals, the number of positive outliers and the upper fence, and finally the number of negative outliers and the lower fence. Exhibits 19 to 21 list the total number of outliers and their percentage of all residuals. The first and third quartiles are not listed but can be obtained from the lower and upper fences by adding or subtracting three IQRs as appropriate. The number of residuals equals the number of observations used to estimate the equation and does not reflect the size of the SMSA population.

The IQR, which measures the spread or clustering of the residuals, is given in column 4 of Exhibits 13 to 18. Comparison of renter and owner models in the same SMSA shows the spread to be larger for owners in most cases. This reflects the generally greater spread in reported house value than in reported rent, not the quality of fit in owner versus renter models. The variance of the logarithm of value is, in

1. Our use of the IQR in defining outliers follows the work of John W. Tukey (1977).

fact, greater than that for rent in all samples except Boston, Philadelphia, Honolulu and New York.

The distribution of IQRs for residuals in the renter models is displayed through use of a stem and leaf plot in Exhibit 22. The median IQR of the residuals is .263. The distribution looks normal except for large spreads in Pittsburgh (.366), Albany (.360) and Boston (.345). The tightest clustering occurs in Rochester and Las Vegas where both IQRs are .199. A comparable stem and leaf plot in Exhibit 23 shows the IQRs for owners to be generally higher than for renters--as noted above. The median residual spread for owners is .338. Birmingham has the widest spread with an IQR of .419 while the tightest clustering is in Paterson (IQR of .218). The IQRs for Pittsburgh of .366 for renters and .404 for owners rank among the largest for both tenure types.

When comparing IQRs among SMSAs it should be kept in mind that the IQRs reflect both goodness of fit and the underlying variation in the dependent variable. For example, the IQR for residuals in the Paterson owners equation is the lowest for all SMSAs but the Paterson R^2 statistic of .49 is also among the lowest. Paterson's low IQR is more a result of the relatively small variance in reported house values than it is a measure of the model's success. The reliability of predicted values from a model should be viewed as a function of the R^2 or F-statistic, the IQR of the residuals, and the proportion of outlying residuals. The R^2s and F-statistics are discussed in Chapter III; the outlying residuals are addressed next.

Exhibits 19 to 21 list the number of residuals with values that lie beyond the calculated fences. The adjacent number in parentheses is the

percentage of observations labeled outliers, which is a better measure of model behavior since the number of total residuals varies widely. The stem and leaf plot in Exhibit 24 shows that the distribution of the percentage of outliers for the renter equations is not symmetric about the median value of .53 percent due to the long tail for high percentages. Honolulu's value of 1.87 percent outliers is clearly larger than the expected value for the distribution. Other SMSAs with large percentages of outliers are Rochester, Anaheim, Denver, Orlando, and Omaha. Four of these six cities are new rapidly growing SMSAs. Rapidly and slowly growing SMSAs are equally prevalent among SMSAs with the lowest fractions of outliers. The distribution of the percentage of outliers for the homeowner equations is given in Exhibit 25. With the exception of Louisville (1.06) and Baltimore (1.00), the distribution is symmetric about the median value of .35 percent. This distribution has a smaller variance than the corresponding distribution for renters. In addition, the owner models have fewer estimated equations with more than one percent outliers compared to the renter equations. These comparisons suggest the owner models provide better fits to the data which is surprising since the renter equations typically have better R^2s and F-statistics. The analysis of the residuals for three specific renter equations, provided later in this section, will suggest an explanation for the apparent paradox.

Columns 5 and 6 of Exhibits 13 to 18 list the number of positive and negative outliers. The number of negative outliers is greater than the number of positive outliers in 53 of 59 renter equations and 47 of 59 owner equations. Out of all outliers in the renter models, 77.7 percent

are negative and in the owner models 71.9 percent are negative. Exhibit 26 shows the classification of outliers by sign and tenure group.

The preponderance of negative outliers suggests omitted variables or bad data since one expects the same number of positive as negative outliers to arise by chance. A negative outlier implies either that the unit's rent or value is seriously under reported, over predicted or both. Long time homeowners might under report house value since values have been rising rapidly recently. The length-of-tenure variables in the homeowner model should adjust for the average under reporting of long-time occupants, but there could be wide variability in the amount of such under reporting. Renters are more likely to know their rent precisely. However, if these rents are below market levels because the tenant works for or is related to the landlord, the equations would overpredict rents.

The greater number of negative outliers in most equations also indicates a skewed distribution of residuals. To see whether this skewness also occurs in the other residuals we examine the lower and upper fences and the median. The fences, which are equidistant from the first and third quartiles, have similar numerical values if the inner half of the residuals are symmetric about zero. A larger absolute value for the upper fence indicates a downward skew for these residuals; a smaller value indicates the opposite skew. The upper and lower fences for the fifty-nine renter and owner equations appear in columns 5 and 6 of Exhibits 13 to 18. Albany renters, the first model in Exhibit 13, has upper and lower fences of 1.24 and -1.24 indicating symmetry. In the other equations the fences are generally close in absolute value.

Although the fences are generally close in size, there is a tendency for the upper fence to be greater than the lower one. Out of the 49 times the fences differ in the renter model, the upper fence is larger 47 times. Out of the 41 times the fences differ in the owner model the upper fence is larger 26 times. Thus, the inner half of the residuals are basically symmetric but to a limited extent show the downward skew also found among the distribution of outliers. Median values, reported in column 3 of Exhibits 13 to 18, are all close to zero supporting the basic symmetry of the fences, though the renter medians are disproportionately positive suggesting the same slight downward skew shown by the fences and the outliers.

Our examination of residuals among the fifty-nine renter and owner models finds the inner half of residuals to be basically symmetric about zero and appropriately clustered for most equations. Also, the proportion of outliers is greater than one percent of all residuals in only a handful of cases. For these reasons we believe the models generally fit the data well and that the t-tests and F-tests presented earlier are reliable in spite of the failure of most models to meet strict normality tests for the residuals. We hasten to add that the preponderance of negative valued outliers suggests a specification or data shortcoming needing further analysis. That is the task begun in the following section.

Residuals Analysis and Re-estimation in Five Equations

In an analysis of residuals the choice must be made between attributing an unusual observation to error and deleting it from estimation, or keeping the data point because it contains important information

about the model. In the second half of this chapter we estimate revised equations based on an analysis of the outliers, while first keeping, then deleting the remaining outliers. We point out the advantages to deleting outliers but remind the reader that a few will arise naturally from a large number of observations on a normally distributed random variable. Therefore all outliers ideally should not be deleted. The problem of course is that the valid observations are difficult to distinguish from the invalid ones.

Our procedure for analyzing the outliers of an estimated model consists of several steps. First, we examine the plot of the residuals versus the predicted values of the dependent variable. We examine this graph for obvious patterns in the residuals. Characteristics of the outliers are then examined to determine whether there are any similarities among the observations that generated the outliers. We compare the full sample distribution of variables to the distribution of these variables for the outliers.[1] Interaction terms are introduced in an equation whenever the distribution of outliers by regressors is different from the sample distribution. The revised equation is estimated using least squares and differences in the models are noted. Finally, observations which are outliers in the revised model are deleted, and the revised equation is reestimated with the smaller sample. We note changes that occur in the estimated coefficients of

1. Only variables included in the regression are used in the comparison. Other variables available from the AHS were not used because of the cost of merging regression results--the residuals--with the original AHS user tapes. This should be a first step in future analysis of the residuals.

the interaction terms and in the other regression coefficients. In addition, we note the reduction in the standard error and related statistics.

We chose to analyze the Anaheim renter, Baltimore renter and owner, Chicago renter, and Fort Worth owner equations. These models represent a cross-section of the fifty-nine SMSAs by size, location, rate of growth, and sampling period. In addition, these models exhibit interesting patterns in their residuals. Anaheim is a small, rapidly developing SMSA in the Southwest. It has the second largest percentage of outliers for the renter equations but an equal number of positive and negative outliers. The IQR of the residuals for Anaheim is among the smallest in the distribution. Unlike Anaheim, Chicago is a large, already developed SMSA in the Midwest with mostly negative outliers. Anaheim is included in Wave 1 of the Annual Housing Survey (AHS) while Chicago is sampled during Wave 2. Baltimore represents Wave 3 of the AHS and is an older, northeastern American city. In Baltimore, as in Chicago and most of the renter equations, the number of negative outliers is much greater than the number of positive outliers.

The estimated equations for Fort Worth and Baltimore represent the owner models. Fort Worth is a rapidly growing southern city sampled during Wave 1 of the AHS. Unlike the majority of the owner equations, Fort Worth has more positive than negative outliers. Baltimore is representative of the typical owner model since it has a greater number of negative than positive outliers. Both Fort Worth and Baltimore have high IQRs of the residuals and high percentages of outliers. Chicago is the only SMSA with 15,000 observations included in the residuals analysis because of the cost of working with the larger sample.

We begin the analysis of the residuals in individual models with the Baltimore renter equation. The graph of the residuals versus the predicted rents for Baltimore appears in Exhibit 27. The variance of the residuals appears to be constant along the horizontal axis.[1] The noticeable feature of the plot is the large number (9) of negative outliers. The values of the outliers and the data that produced these residuals are listed in Exhibits 28 and 29. We find all negative outliers correspond to low values of reported rent although their predicted rents are spread across the range of other predicted rents. In Exhibit 30 we compare the natural log of rent, for the sample observations to the subsample of outliers. One hundred percent of the outliers are in the lowest two percent of the reported rent distribution! The three lowest reported rents are outliers. This means the model overpredicts rent for a large proportion of all households reporting a low rent. It is possible, but seems unlikely, that the families associated with these outliers are reporting erroneous monthly rent data. These families are nevertheless reporting rents far below their market value as judged by the hedonic equation. These families may be related to their landlord or work in lieu of paying rent. It is not possible to test these hypotheses using the AHS since the relevant questions are asked only if the respondent is paying no cash rent. There is no way to determine whether low rents also reflect such extra considerations.

1. It needs to be added that such plots for most SMSA renter equations show no strong pattern for residuals to spread out or become more concentrated as predicted rents rise. Thus, the regression model's assumption of constant variance seems adequately satisfied. The same cannot be said for the owner model residuals which show a strong tendency towards increased clustering as predicted value rises (see Exhibit 33 and the discussion in Chapter II). The owner estimates are consequently less efficient than they could be.

Exhibit 31 also compares the sample frequency to the outlier frequency for three variables included in the regression. The three are 1940, CC1 (an indicator variable for a unit located in the central city), and SFATT (an indicator variable for a single family attached dwelling). The negative outlier frequency is considerably greater than the sample frequency for these characteristics. This outcome may be due to chance or may be caused by significant interaction effects among these variables. The Baltimore renter equation is reestimated with the three combinations of interaction terms included in the regression. Each of the estimated coefficients for the interaction terms is negative and two are statistically significant at the one percent level. Exhibit 32 compares the estimated coefficients in the original equation to the equation including interaction terms. An old unit in the central city offers the tenant a 13 percent reduction in rent. An old, single-family attached unit and a central city, single-family attached unit offer 9 percent and 7 percent discounts, respectively. The inclusion of the interaction terms lowers the value and the statistical significance of the estimated regression coefficients for single-family attached and central city units. The difference in interpreting the estimated coefficients for each of the two models is important. In Model A (the original specification of the model) central city units and single-family attached units offer discounts of 7 percent and 11 percent, respectively. In Model B central city units and single-family attached units are discounted only if these units are also old. The estimated coefficients for the remaining variables in the model and their standard errors do not change.

The final step in the residual analysis is to reestimate the regression coefficients while deleting the outliers observed in Model B. The same observations that produced outliers in Model A produce outliers in Model B indicating that the interaction terms failed to accomodate the original outliers. Model C in Exhibit 32 lists the results of deleting the outliers and reestimating the regression coefficients. The estimated coefficients of CC1DAGE (an indicator variable for an old, central city unit) and of DAGESFAT (an indicator variable for an old, single-family attached unit) remain statistically significant at the one percent level. The discount for an old, central city unit is reduced from 12.9 percent to 9.7 percent, however. This means the statistical significance of the interaction terms included in Model B is not produced by the outliers alone. Two additional changes occur in the estimated coefficients going from Model B to Model C. The estimated coefficient of the four or more bedrooms variable increases by 60 percent and remains statistically significant. The estimated regression coefficient of the indicator variable for black head of household swells by a factor of 5 but still does not become strongly significant. The standard error in the estimated equation goes from 0.2555 in Model A to 0.2300 in Model C, a reduction of 10 percent. Model C shows larger values for levels of statistical significance for most variables as well as higher R^2 and F-statistics because of the smaller residual variance. The confidence interval around predicted rents would be similarly reduced.

The residuals for the Baltimore owner model show a pattern similar to the residuals of most other owner equations. The plot of the

residuals versus the predicted values in Exhibit 33 shows a tendency for the variance of the residuals to decline with increasing values of the predicted value of the dependent variable. The plot also shows that 16 of 20 outliers are negative (see also Exhibits 34 and 35). The negative outliers for this model occur in the low range of reported house values even though their predicted values are spread throughout much of the range of all predicted values. A striking 81 percent of the negative outliers appear in the lowest one percent of the distribution for reported house value (see Exhibit 36). All of the negative outliers occur in the lowest 2.1 percent of the distribution for reported house values. It seems likely that many of these homeowners are understating the market value of their homes.

The outliers in the Baltimore owner equation exhibit a greater percentage of old, central city, and single-family attached units as did the outliers in the Baltimore renter equation (see Exhibit 37). The same interaction terms used in the Baltimore renter equation produce statistically significant estimated coefficients in the owner model. The estimated coefficients for the interaction terms in the owner equation are at least 75 percent larger than they were in the Baltimore renter equation. Exhibit 38 lists the estimated coefficients for Models A, B, and C for Baltimore owners. Deleting the outliers from Model B leaves the estimated coefficients for the interaction terms statistically significant at the one percent level while raising the t-statistics for most coefficients. The estimated coefficients for single-family attached units and central city units decrease in value from Model A to Model C but remain statistically significant at the one

percent level. The standard error of the equation is reduced 12 percent from Model A to 0.3124 in Model C which causes a 6.4 percent increase in the R^2 statistic, to 0.7339.

The procedure for analyzing the residuals in the Anaheim renter, Chicago renter, and Fort Worth owner models is the same as that used to analyze the residuals in the two Baltimore equations. The estimated coefficients for Models A, B, and C for Anaheim, Chicago, and Fort Worth are listed in Exhibits 39, 40, and 41, respectively.

We summarize the results of the analysis, beginning with Anaheim. Four out of seven of the negative outliers are in the lowest one percent of the reported rent distribution and all the negative outliers are in the lowest 6 percent of the rent distribution. All the postive outliers are in the upper 10 percent of the rent distribution while the four largest rents produce outliers. The inclusion of an indicator variable for a large dwelling interaction is suggested by examining the data on the residuals. The estimated coefficients for a dwelling having three or more bathrooms and four or more bedrooms is .192 and significant at the 10 percent level. The estimated coefficient of the large dwelling interaction variable shows no change after deleting the outliers. The estimated coefficient for single-family detached units (SFDET) is 0.0159 and statistically insignificant in Model B while it is 0.0403 and significant at the one percent level in Model C. The standard error in the Anaheim renter model is reduced by 11 percent after deleting outliers.

The outliers in Chicago are mostly negative with 20 of 21 negative outliers corresponding to the lowest 5 percent of the rent distribution.

An indicator variable for the presence of an elevator in a structure of more than fifty units produces an estimated coefficient of 0.11 and is statistically significant at the one percent level. The inclusion of the interaction terms reduces the magnitude and the statistical significance of the fifty or more units indicator variable. The Chicago standard error decreases 10 percent by including the interaction terms and deleting outliers.

The Fort Worth owner equation exhibits more positive outliers than negative ones. All positive outliers occur in the highest quartile of the reported house value distribution. The four smallest reported house values appear as the four negative outliers! An interaction term for old units with owners having a long length of tenure produces an insignificant estimated coefficient before and after deleting outliers. Deleting the outliers reduces the standard error in the Fort Worth owner model by eight percent to 0.3081.

In this section we have found the disproportionate number of negative outliers to be associated with very low reported rents and values. All negative outliers have reported rents or values in the bottom 6 percent of their respective distributions, and most of them have occurred in the lowest one percent. Examination of characteristics unique to the outliers has suggested interaction terms for inclusion in the models. Addition of these terms mostly fails to bring the outliers back into line, although the variables frequently do well even when the outliers have been deleted. While there appears to be room for improving the hedonic specification by including interaction terms, new information is needed to explain the outliers. Deletion of outliers reduces

standard errors and increases sigificance levels substantially as well as altering specific coefficients. Some outliers should be expected even in a complete model, however, so that dropping all outliers as we have done probably overstates the reduction in residual variance that better information could achieve. It is clear though that a better understanding of the negative outliers could lead to important improvements in AHS-based hedonic models.

One implication of these findings is that predictions of rents and values from the equations reported in Chapter III will tend to be biased downward. Inclusion of observations with largely unexplained and very low rents pulls down the average predicted rent in the sample, and is likely to pull down predicted rents for most dwelling specifications. Exclusion of the outliers will not necessarily avoid the downward bias. That depends on the source of the bias. If it is something that effects only the outlying observations then this bias can be avoided by deleting the observations. However, if the source of bias effects other dwellings as well, deletion of the outliers would not eliminate the bias. In a following paper we will examine the effect of eliminating outliers on predictions of rents and values. In the remaining paragraphs of this paper we suggest ways to search for the source of the negative outliers and to measure their impacts on prediction.

The most likely source of the negative outliers in our opinion is that reported rents and values understate actual market prices. Renters could receive reduced rent because they work for the landlord or are related to him. In the Demand Experiment of the Experimental Housing Allowance Program, where these questions were asked, 5.2 percent of

respondents in Phoenix worked in lieu of full rent and another 2.2 percent paid reduced rent because they were related to the landlord.[1] The Demand Experiment data could be analyzed to see if hedonic models like those estimated for this project produce a similar majority of negative outliers. If so, those outliers could be examined to determine how many receive subsidized rents for the above reasons. Estimation of the model excluding reduced rent for these reasons could be used to give an idea of the prediction error from this source. Earlier work by Ozanne, Andrews and Malpezzi (1979) found that models for evaluating AHS hedonics can be constructed from the Demand Experiment data and that these models give substantial discounts to tenants that are related to their landlord. Thus, this seems like a promising avenue of analysis.

Long-time homeowners may have widely varying ideas about the current value of their homes. Since values have been rising rapidly in many places, some of these homeowners could substantially under-report the value of their homes, even relative to the average for long time occupants. Perhaps characteristics like age of the survey respondent, when interacted with length of tenure, would characterize some of these outliers.

Other sources of under reporting would not be as easy to identify. Renters may receive subsidized housing but not report it, or know it. A few homeowners actually do get very low priced housing from urban homesteading programs.

---

1. Percentages supplied by James Zais from user tapes of Demand Experiment tenants survey. Analyses of market outcomes performed on the experimental data commonly exclude these non-market rents, e.g., Cronin (1979).

Under reporting is not the only potential cause of the negative outliers. Omitted physical, neighborhood or locational information could also be involved. A simple first step in investigating this possibility would be to examine variables omitted from the hedonic but included in the AHS. Future metropolitan AHS user tapes will identify dwellings located in the same sampling cluster. This information could be used to test whether neighborhood location is associated with the low rents and values. In this section we have suggested hypotheses that could account for the observed outliers and ways to test the hypotheses. It remains for future work to investigate them.