An Investigation of Residential Real Estate Property Values in the Washington, D.C., Metropolitan Statistical Area (MSA)

Stephen Widener

INTRODUCTION

Discussion of the Empirical Investigation

Local governments are constantly trying to determine what really affects the value of residential real estate in their area. It's no surprise that local governments are interested in this because a good portion of their revenue comes from taxes collected from owners of residential real estate. The Mayor, Controller, and other city officials look forward to more tax revenues each year coming from higher property values. They use these revenues to pay police officers, teachers, social workers, garbage collectors, street repair personnel, and many others. Therefore, each local jurisdiction needs to know how much to expect from property taxes. This means they must know a property's future value in order to budget future expenses. Ultimately, local governments must have a method of assessing the market value of residential real estate (hence forth referred to as "property").

Local governments employ appraisers to help them come up with projections about a property's market value. There are several classical appraisal methods considered time-tested and widely used: the Sales Comparison Approach, the Cost Approach, and the Income Approach (Betts and Ely, p. 46). Sometimes, local governments employ appraisers who use methods that are more scientific; methods commonly used in mathematics and statistics. Multiple Linear Regression (MLR) is a scientific method using several explanatory variables to predict a dependent variable. I will use MLR to find out the forces affecting property values *now*, and the formula that predicts property value in the *future*.

But what forces shall I examine? Where do I start?

Conventional wisdom generally says that macro forces play a more important role in the value of property than micro forces. For example, shifts in population, unemployment rate, the economy, or new construction are said to have an important impact. While this may be true, I question the usefulness of analyzing such forces[1]. Discussions with professionals in the housing industry lead me to believe that they are more interested in the micro forces[2] (e.g. local characteristics). In my opinion, local governments will find this kind of information more useful. For example, a local government can do something about the quality of schools in their area if they are educated about how it affects property values. What can they do about a national economic crisis, the war in Iraq, or an international trade deficit? Educating local jurisdictions about the micro forces affecting their property makes sense and is the focus of my research paper.

This paper is an empirical investigation into the micro forces driving local property values[3]. My theory is this: micro forces play a significant role in determining the value of residential real estate.

METHODOLOGY

The methodology section presents the methods and procedures used to collect and analyze the data, as well as the type of data used in the analysis.

---

[1] Understanding macro forces might be useful to real estate companies deciding where to develop a planned community. But, is it *useful* to government officials at the local level who must somehow understand their own local area and the "small changes" that *only* affect them?

[2] Mary Lee Widener – President, Neighborhood Housing Services of America.

[3] The study uses property values in a single locality in the Washington, D.C. Metropolitan Statistical Area (MSA). The data set is obtained from the Bureau of the Census's 1998 American Housing Survey. (http://www.huduser.org/datasets/ahs/ahs98metro.html - content updated 9/15/03)

Explanation of the Data

The 1998 American Housing Survey contains over 800 variables to choose from. I selected variables in the survey that I felt would best represent the micro forces affecting property value.

Then, I filtered the data in order to remove observations that were not appropriate for my study. For my observation deletions: all respondents that were not interviewed were deleted because there was no data in their records; all units except single detached/attached residences were deleted because my study does not focus on other types of property like condominiums or duplexes; time shares were deleted because they are not primary residences; all units that did not employ either a standard sewage system or septic tank were deleted because I was not interested in studying dwellings like vacation retreats and country-side log cabins; all units with less than 500 square feet were deleted because it's unlikely that single detached/attached residences are less than 500 square feet. There were 2606 observations (out of about 4,400) left in the data set after deletion.

Dependent Variable:

The dependent variable of the regression model is the estimated current value of the property (propvalue) in year 1998. (For details see Appendices A and B.)

Independent Variables:

There are two general "concepts" that affect property value: characteristics of the housing unit and characteristics of the local neighborhood. Housing unit variables refer to the physical property in which the person lives. Housing unit variables[4] are within the control of the homeowner as he/she can add or change any features to the home and potentially improve its value. Central air conditioner (airsys), number of bathrooms (baths), number of bedrooms (bedrms), basement (cellar), fireplace (fplwk), garage (garage), and unit size (unitsf) positively affect property value; while age of unit (houseage) negatively affects property value.

Unlike the housing unit variables, local neighborhood variables refer to external factors going on in the neighborhood over which the homeowner has no control. The local neighborhood setting is determined by factors like the local residents, surrounding houses, streets, schools, and shopping centers. Adequacy of neighborhood (hown), neighborhood elementary school satisfactory (sch), and neighborhood shopping satisfactory (shp) positively affect property value; while neighborhood crime (crimea) and abandoned/vandalized buildings (eaban) negatively affect property value.

(For more details about the independent variables see Appendices A and B.)

Explanation of Analysis Method(s) Used

The Analysis Method used can be described in a four-step process: Data Extraction and Migration, Data Manipulation, Data Validation, and Data Analysis.

Data Extraction and Migration:

The extraction and migration of data from the Bureau of the Census's 1998 American Housing Survey relied on the Internet and the StatTransfer software package. The data was downloaded and then expanded from a compressed format into several large text files. Then the data was migrated from the text format to STATA format using StatTransfer.

[4] The variable I used to represent the age of the house (houseage) is the only housing unit characteristic not within the control of the homeowner.

Data Manipulation:

The manipulation of the housing data was done by first dropping uneeded variables from the data set. The second step entailed merging several tables into a single STATA file. The last step entailed creating the new variables needed for analysis (e.g., propvalue and houseage).

Data Validation:

The validation process involved validating the observation counts presented in the data against an independent printout of summary statistics provided by the Census Bureau.

Data Analysis:

After validating the data, I analyzed it using the Ordinary Least Squares (OLS) Regression Method with the STATA statistics program. Some variables were kept and some were dropped depending on whether they were significant[5]. This was repeated until all variables were significant and the relationship between them and property value adhered to my expectations. This model was then tested to determine if any of the model's assumptions[6] were violated; steps were taken to address violations[7] so analysis could continue. Once the final equation was produced, its projected values were validated by plotting them against the actual property values found in the data.

RESULTS

Discussion of Findings in Comparison with Previous Results

In the spirit of conciseness, I listed the OLS regression models I worked with in the first phase of my research in Appendix F. At the end of the first phase I discovered that there was potential for heteroscedasticity in the regression model. To test for heteroscedasticity, a **`hettest`** was performed and the null hypothesis (constant variances) was rejected because of the high chi2 and low p value (meaning, there is indeed heteroscedasticity in the model).

```
hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of airsys

         chi2(1)      =   330.25
         Prob > chi2  =   0.0000
```

Additionally, I graphed the residuals against each explanatory variables to see which of the explanatory variables were causing heteroscedasticity. (For details see Appendix C.)

With the presence of heteroscedasticity, OLS regression models cannot be used because errors are no longer normally or independently distributed. In order to correct heteroscedasticity I performed multiple linear regressions using the *robust regression method*. Relevant variables were kept and irrelevant variables were dropped based on their significance (t-scores and p values).

The robust regression models appear below.

---

[5] A variable is significant to the model if its P value is less than 0.05 ($p < 0.05$).

[6] Classical Linear Regression Model (CLRM) makes certain assumptions about the data. For example, it assumes a normal distribution, homoscedasticity, no autocorrelation, and no multicollinearity. (Gujarati, p. 201)

[7] Specific information about the CLRM assumptions violated is discussed in the following section.

Robust Regression Model 1:

```
regress propvalue airsys baths bedrms cellar crimea eaban fplwk garage houseage
hown sch shp unitsf, robust

Regression with robust standard errors                Number of obs =     372
                                                      F( 13,   358) =   39.66
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5941
                                                      Root MSE      =   66706
------------------------------------------------------------------------------
             |               Robust
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |   65380.83   15657.49     4.18   0.000     34588.62    96173.05
       baths |   30854.04   6217.716     4.96   0.000      18626.2    43081.87
      bedrms |   13244.04   5379.359     2.46   0.014     2664.929    23823.16
      cellar |   16149.89   7394.277     2.18   0.030      1608.21    30691.57
      crimea |   2046.861   10903.88     0.19   0.851    -19396.84    23490.56
       eaban |  -16786.88   17397.03    -0.96   0.335     -51000.1    17426.35
       fplwk |    21637.7    7443.42     2.91   0.004     6999.374    36276.02
      garage |   41707.86   7853.043     5.31   0.000     26263.97    57151.75
    houseage |   1180.109   244.3383     4.83   0.000     699.5903    1660.628
       hown |   2323.719    2471.84     0.94   0.348    -2537.433    7184.871
         sch |   23211.21   9802.341     2.37   0.018     3933.807    42488.62
         shp |   23524.01   12741.49     1.85   0.066    -1533.563    48581.59
      unitsf |   38.63689   5.392642     7.16   0.000     28.03165    49.24212
       _cons |    -204984   35139.58    -5.83   0.000    -274089.9   -135878.1
------------------------------------------------------------------------------
```

This regression model explains approximately 59% of the variation in property value (R-squared = 0.5941).

The significant variables with an expected sign[8] are: central air conditioner (airsys), number of bathrooms (baths), number of bedrooms (bedrms), basement (cellar), fireplace (fplwk), garage (garage), neighborhood elementary school satisfactory (sch), and unit size (unitsf). However, age of unit (houseage) is statistically significant with the *wrong* sign[9].

The insignificant[10] variables are: vandalized buildings in neighborhood (eaban), adequacy of neighborhood (hown), and neighborhood shopping satisfactory (shp) and neighborhood crime (crimea). Under normal circumstances I would investigate them, but I instead chose to focus my attention on something else that was unusual in the model.

I noticed that only 372 of 2606 observations were used by this regression model[11]. I am not confident that this model best represents the population considering it only used 14% of the available observations. I reviewed a summary of all the variables to determine which explanatory variable(s) contributed to this problem. The results are below.

[8] An expected sign means the sign of the variable is consistent with my prediction or expectation.
[9] The model shows a positive sign for this variable, but I predicted a *negative* sign; according to the Appendix D correlation matrix there is a *negative* correlation between propvalue and houseage. This means that this variable is suspicious.
[10] These variables are insignificant so I did not address their signs because they are meaningless.
[11] A small number of observations returned in a regression model can cause all kinds of problems including returning variables in the model with the wrong sign.

```
sum propvalue airsys baths bedrms cellar crimea eaban fplwk garage houseage hown
sch shp unitsf

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   propvalue |      2106    201883.7    104868.6       1000     510000
      airsys |      2606    .8576362    .3494899          0          1
       baths |      2593    1.944852    .8110716          1          6
      bedrms |      2606    3.324635     1.01644          1          9
      cellar |      2606    .6792018    .4668729          0          1
-------------+--------------------------------------------------------
      crimea |      2518    .1942017    .3956635          0          1
       eaban |      2558    .0308835    .1730359          0          1
       fplwk |      2606    .6005372     .489882          0          1
      garage |      2603    .4951978     .500073          0          1
    houseage |      2606    31.26247    21.24073          0         79
-------------+--------------------------------------------------------
       hown |      2606    7.676132    2.454394          0         10
         sch |       582     .790378     .407389          0          1
         shp |      2545     .902947    .2960882          0          1
      unitsf |      1863    2319.588    1103.571        500       5200
```

Based on this summary, neighborhood elementary school satisfactory (sch) only has 582 observations available for regression analysis (the rest of this variable's data was not reported). I cannot continue to use this one because I could be jeopardizing the accuracy of the model. Therefore, I dropped the variable and ran the regression below.

Robust Regression Model 2:

```
regress propvalue airsys baths bedrms cellar crimea eaban fplwk garage houseage
hown shp unitsf, robust

Regression with robust standard errors                Number of obs =    1518
                                                      F( 12,  1505) =  143.89
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5382
                                                      Root MSE      =   70993
------------------------------------------------------------------------------
             |               Robust
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |   16766.86    7667.76     2.19   0.029     1726.234    31807.49
       baths |   34378.25   3348.492    10.27   0.000     27810.04    40946.46
      bedrms |   6967.568   2675.764     2.60   0.009     1718.946    12216.19
      cellar |   14710.95   4293.015     3.43   0.001     6290.019    23131.87
      crimea |  -8907.955   4899.968    -1.82   0.069    -18519.44    703.5355
       eaban |  -24448.08   11608.71    -2.11   0.035    -47219.04   -1677.113
       fplwk |   25763.84   4025.511     6.40   0.000     17867.63    33660.05
      garage |   28108.48   4271.011     6.58   0.000     19730.71    36486.24
    houseage |   783.4125   116.1929     6.74   0.000     555.4953     1011.33
       hown |    4155.81   1046.343     3.97   0.000     2103.364    6208.256
         shp |    6573.29   7613.509     0.86   0.388    -8360.923     21507.5
      unitsf |    36.6492   3.045686    12.03   0.000     30.67496    42.62344
       _cons |  -97787.89   15474.98    -6.32   0.000    -128142.7   -67433.07
------------------------------------------------------------------------------
```

The number of observations increased from 372 to 1518.

This regression model explains approximately 54% of the variation in property value (R-squared = 0.5382). Although the R-squared has declined (from 0.5941 to 0.5382), this regression model is better considering the larger number of observations.

In this regression, all variables are significant *except* for neighborhood crime (crimea) and neighborhood shopping satisfactory (shp). Theoretically, neighborhood crime (crimea) and neighborhood shopping satisfactory (shp) might show up as insignificant in the model if they are irrelevant[12] or show signs of multicollinearity[13]. Based on the correlation matrix in Appendix D, there were no signs of multicollinearity so I feel these variables are *probably* irrelevant. I dropped the variables and ran the regression below.

Robust Regression Model 3:

```
regress propvalue airsys baths bedrms cellar eaban fplwk garage houseage hown
unitsf, robust
```

```
Regression with robust standard errors                Number of obs =    1537
                                                      F( 10,  1526) =  176.16
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5391
                                                      Root MSE      =   70991
------------------------------------------------------------------------------
             |               Robust
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |   16121.94   7529.687     2.14   0.032     1352.313    30891.57
       baths |   34404.04   3314.874    10.38   0.000     27901.85    40906.24
      bedrms |   7163.259   2633.439     2.72   0.007     1997.717     12328.8
      cellar |   14574.35    4263.41     3.42   0.001     6211.588    22937.11
       eaban |  -27464.87   11566.57    -2.37   0.018    -50152.93   -4776.814
       fplwk |   25634.42   3956.619     6.48   0.000     17873.43    33395.41
      garage |    28211.3   4231.424     6.67   0.000     19911.28    36511.32
    houseage |   748.5139    114.839     6.52   0.000      523.255    973.7728
        hown |   3706.791   950.5318     3.90   0.000     1842.304    5571.277
      unitsf |    36.7025   3.018745    12.16   0.000     30.78117    42.62383
       _cons |  -88600.75   12305.86    -7.20   0.000    -112738.9   -64462.56
------------------------------------------------------------------------------
```

This regression model explains approximately 54% of the variation in property value (R-squared = 0.5391). The R-squared increased by 0.0009 (from 0.5382 to 0.5391) and there is little or no change in the t-scores of other variables. These differences are miniscule and they agree with my theory that the variables I dropped previously were *in fact* irrelevant.

[12] Removal of an irrelevant variable only *slightly* changes the R-squared and t-scores (or in some cases the values don't change at all). The inclusion of an irrelevant variable does not bias the regression model; however, it is committing a specification error of overfitting the model. (Gujarati, p. 413)

[13] Sometimes, a variable becomes insignificant if multicollinearity occurs (> 0.80). Mulicollinearity refers to a single perfect or near perfect linear relationshp between two or more explanatory variables. If mulicollinearity occurs, the removal of one of the correlated variables should significantly affect the t-score, p value, coefficient and R-squared.

The model now contains all significant variables, but there is still one problem. Age of unit (houseage) continues to have the *wrong* sign[14]. I suspect that the relationship between this variable and the dependent variable (propvalue) is probably being indirectly influenced by another variable that was *not included* in this model. In other words, I may have *omitted* one or more relevant variable(s) thereby "underfitting" the model and creating bias[15]. To determine whether this was true I performed an **ovtest** to check for omitted variables.

```
ovtest
Ramsey RESET test using powers of the fitted values of propvalue
       Ho:  model has no omitted variables
                 F(3, 1523) =     30.84
                   Prob > F =      0.0000
```

Based on the **ovtest**, I rejected the null hypothesis and concluded there are in fact omitted variables (because of the high F score and low p value)[16].

I suspect the issue I'm dealing with here is stemming from data that (1) wasn't collected in the survey and (2) affects the value of the properties in the survey. Can remodeling work be affecting property value? Theorectically, I would guess that it would considering that when older homes are *remodeled* it usually makes them more valuable than older homes of the same age that *aren't* remodeled. Unfortunately, my survey lacked information on whether a home was remodeled, the kind of remodeling done, or the money spent on remodeling[17]. Therefore, I decided that age of unit (houseage) was not conclusive and I dropped it. I ran another regression below.

Robust Regression Model 4:

```
regress propvalue airsys baths bedrms cellar eaban fplwk garage hown unitsf, robust

Regression with robust standard errors                Number of obs =    1537
                                                      F(  9,  1527) =  191.16
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5226
                                                      Root MSE      =   72229
------------------------------------------------------------------------------
             |               Robust
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |  -1492.596   7220.209    -0.21   0.836    -15655.17    12669.98
       baths |   29866.63   3303.456     9.04   0.000     23386.84    36346.42
      bedrms |   8300.044   2662.121     3.12   0.002     3078.245    13521.84
      cellar |   16960.85   4297.441     3.95   0.000     8531.338    25390.36
       eaban |  -23292.63   11348.59    -2.05   0.040     -45553.1   -1032.163
       fplwk |   26979.51   4073.904     6.62   0.000     18988.47    34970.55
      garage |   24210.88   4325.031     5.60   0.000     15727.25    32694.51
        hown |   4045.344   962.2617     4.20   0.000      2157.85    5932.839
      unitsf |   35.94155   2.972973    12.09   0.000     30.11001    41.77309
       _cons |  -47819.67   10661.75    -4.49   0.000    -68732.88   -26906.45
------------------------------------------------------------------------------
```

[14] The model shows a positive sign for this variable, but I predicted a *negative* sign; according to the Appendix D correlation matrix there is a *negative* correlation between propvalue and houseage. This means that this variable is suspicious.
[15] Underfitting a model is when a relevant variable is excluded; the coefficients are biased and inconsistent thereby invalidating the hypothesis. Overfitting a model is when an irrelevant variable is *included*; the coefficients are still unbiased and consistent but there is risk of making a significant variable come across as *in*significant. So, age of unit (houseage) might show up in the model with the correct sign if nothing was omitted. (Hamilton, Statistics with STATA, p. 127-128)
[16] It is not surprising to find that there are omitted variables considering the moderate R-squared value.
[17] I chose to use the 1998 survey because I wanted to study a familiar locality, Washington, D.C., and assumed that all the surveys collected the same information as stated in the survey's Codebook. Surprisingly, however, remodeling work data was not collected in 1998.

This regression model explains approximately 52% of the variation in property value (R-squared = 0.5226). Interestingly, central air conditioner (airsys) showed up in the model as *insignificant* and it had the *wrong* sign. Like the age of unit (houseage), central air conditioner (airsys) is *probably* being affected by information that was not collected in the survey. Again, remodeling work comes to mind. Could it be that most people install central air conditioner in old homes when they are remodeled? If so, this would explain the strange behavior of this variable; it is being affected by omitted information about remodeling work and is therefore inconclusive. Once again, I dropped the variable and ran the regression below.

Final Robust Regression Model:

```
regress propvalue baths bedrms cellar eaban fplwk garage hown unitsf, robust

Regression with robust standard errors                Number of obs =    1537
                                                      F(  8,  1528) =  214.16
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5226
                                                      Root MSE      =   72207
------------------------------------------------------------------------------
             |               Robust
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       baths |   29759.11   3243.151     9.18   0.000     23397.61    36120.61
      bedrms |   8302.856   2663.608     3.12   0.002     3078.142    13527.57
      cellar |   16970.11   4295.171     3.95   0.000     8545.058    25395.17
       eaban |  -23171.62   11305.14    -2.05   0.041    -45346.86   -996.3846
       fplwk |   26881.07   4076.076     6.59   0.000     18885.78    34876.37
      garage |   24182.77   4322.286     5.59   0.000     15704.53    32661.01
        hown |   4048.037   962.3582     4.21   0.000     2160.354    5935.719
      unitsf |   35.95202    2.97117    12.10   0.000     30.12401    41.78002
       _cons |  -48920.17   9805.498    -4.99   0.000    -68153.83   -29686.52
------------------------------------------------------------------------------
```

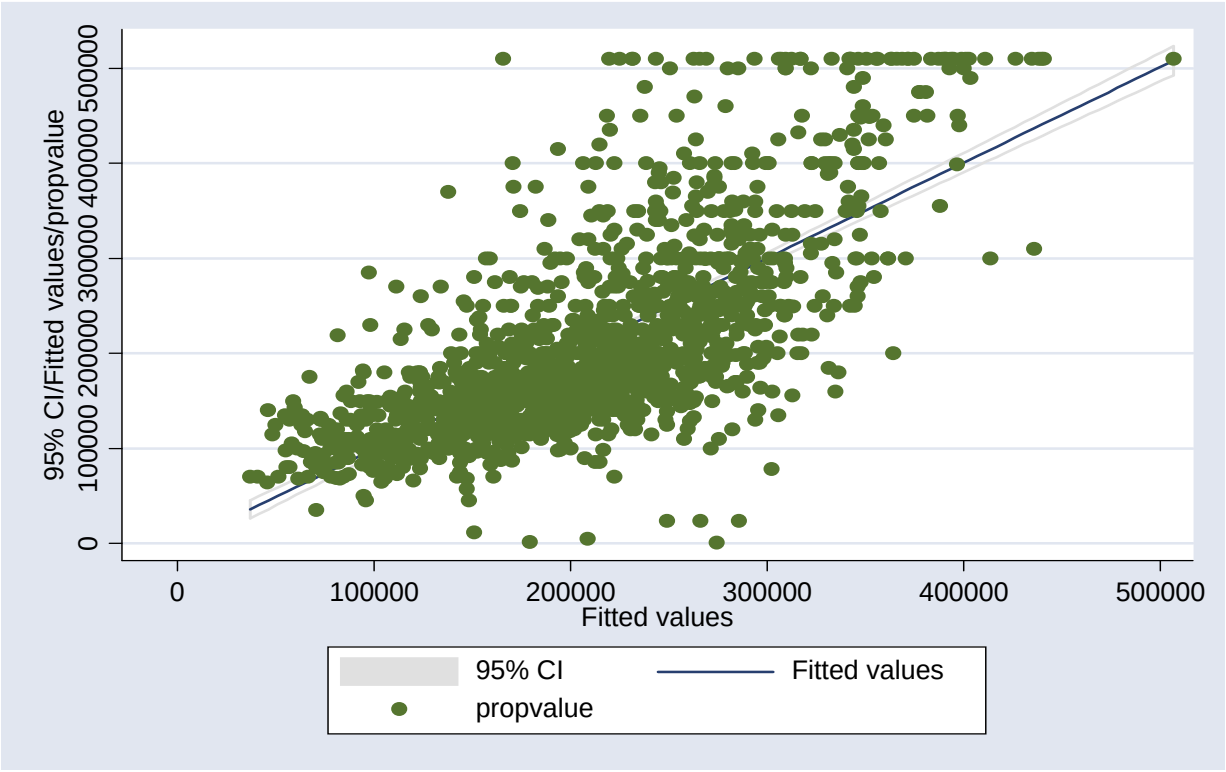
The final regression model explains approximately 52% of the variation in property value and the R-squared has remained the same (R-squared = 0.5226). According to my study:

1. All individual variables in the model are now statistically significant (based on individual t tests … high t-scores and low p-values) and all of the variables have the sign I predicted.
2. The variables I used to represent neighborhood crime (crimea) and adequate neighborhood shopping (shp) turned out to be irrelevant to a property's value.
3. The variables I used to represent a decent neighborhood public school system (sch), the age of the house (houseage), and the presence of a central air conditioner (airsys) *could be* relevant. However, I wasn't able to determine their relevance because missing & uncollected survey data resulted in those variables being inconclusive.

I performed a final test on the regression model; I graphed the fitted values (y-hat) against the dependent variable (propvalue) to visually check the validity of the projected market values produced by the model against the *actual* values in the data set.

The graph below shows the relationship between the two.

```
twoway (lfitci propvalue fittedvalue) (scatter propvalue fittedvalue)
```




```
The Final Regression Equation:

Propvalue =  –48920.17 + 29759.11*baths + 8302.856*bedrms + 16960.85*cellar –
              23292.63*eaban + 26979.51*fplwk + 24182.77*garage + 4048.037*hown +
              35.95202*unitsf
```

(For a detailed technical narration of the variables and what they mean mathematically, see Appendix E.)

LIMITATIONS

Empirical investigation does not come without its limitations. This study was limited by the following:

- The data set used for this analysis was only a survey; it is a collection of answers to questions given by homeowners. What if they really don't know when their house was built? What if everyone rates their neighborhood high because they live in it? Since some of the answers are subjective, the overall bias in the answers given during the survey could result in bias in the regression.
- This survey did not include the time and distance to work. In my opinion, the time and distance it takes for someone to get to work is an important determinant to property value because it indicates convenience to one's job. I would be interested in seeing whether time and distance to work have any effects on property value.
- The survey did not include the cost of remodeling work and the type of home improvements done on the home. As detailed in my paper, I suspect that remodeling work done on homes has some indirect effects on the variables originally used to predict property values.
- Missing values caused the data set to be incomplete. The survey that I used contained many missing values. There were some variables that I intended to use for analysis; but because missing values caused the number of observations to decline, I was unable to use them to make any concrete conclusion. I had to look for other variables as a proxy for the one(s) that I intended to use.

- Since data representing macro forces were not analyzed, my regression model only partially explained property value. Some explanation of the value of the property was left unexplained.
- I would have liked a more homogeneous data set to minimize the variance. For example, property values in the DC MSA include DC, Maryland, and Virginia, all of which probably undergo different phenomena.
- It would have been nice if the survey included city or zip code information. By including the zip code or city, I can potentially extract and include information from other public data set(s). For example, if I want to measure the quality of school, I can obtain SAT scores from a second source and merge them based on the zip code or city. Unfortunately, the Census Bureau didn't collect data this way; they removed the city and zip code information to protect the identity of the people being surveyed.

One way to reduce these kinds of limitations in the study is to use *market data* instead of *survey data*. Market data is collected using more quantitative information and they do not rely on answers to questions by homeowners. However, it is very expensive to obtain.

Another long-term solution to these kinds of limitations is to have a national database of real estate appraisals readily available for public analysis. Appraisals of real estate property are, by law, public information. Therefore, compiling them into a single database would allow information that is already public to be examined better by public officials.

CONCLUSION

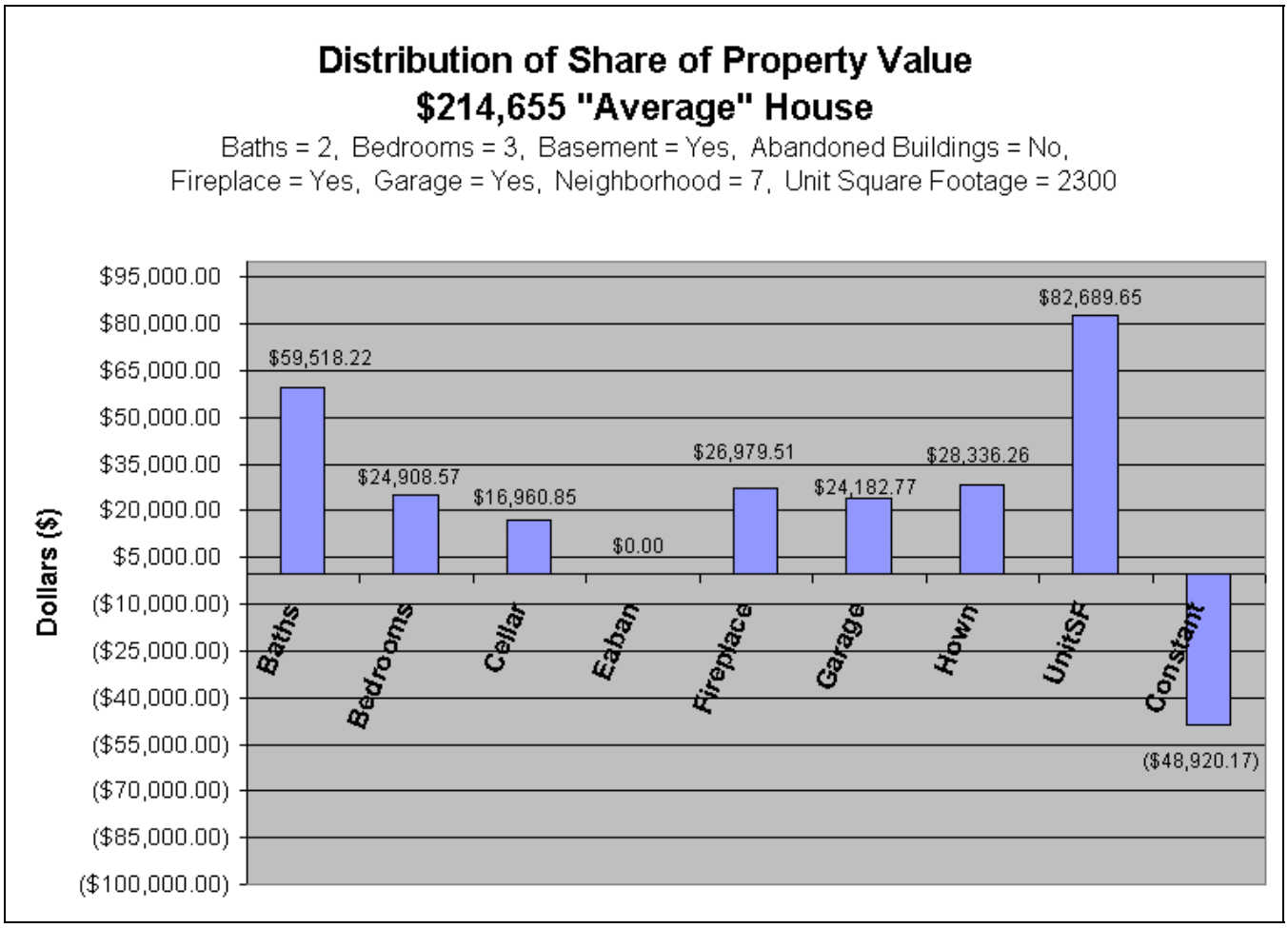
My empirical investigation into the forces driving the market value of residential real estate (property) show that the following play a role in predicting property values:

1. Bathrooms
2. Bedrooms
3. Basement
4. Abandoned/Vandalized Buildings
5. Fireplace
6. Garage
7. Neighborhood Quality
8. Unit Size

Additionally, the following forces can potentially play a role in predicting property value but I could not include them in my study:

1. Age of Unit
2. Remodeling Work
3. Central Air Conditioner
4. Time and Distance from Home to Work

The final equation illustrates some interesting phenomena in the underlying data. For example, below, is a graphical representation of the share of value determined by each variable for an "average" house. (Summary statistics on page 5 show an average house value of $201,884; the average house in the graph is slightly higher due to rounding on variables like bedrooms and bathrooms.)



It's obvious that the size (e.g., square footage) of the house is the most important feature[18] and contributes the largest amount to its total value. Popular belief states that people consider the quality of the neighborhood to be the most important factor but this study shows they are more interested in the size of the house itself; more interested in a place where the whole family can live comfortably. Perhaps people aren't thinking about the quality of the neighborhood as much as they should be? Another finding is that an additional bathroom contributes more value than an additional bedroom, probably because it is more expensive to build one.

Understanding the implications of this study on micro forces affecting residential real estate is important to local officials because it could help them improve their communities. For example, if they knew that negative attributes like abandoned buildings might cause people to buy houses somewhere else perhaps they would do more to decrease the number of these kinds of buildings in their communities. These kinds of improvements could increase property values in the community and result in higher tax revenues for the local government.

[18] The graph on this page shows that the unit's square footage contributes the most to property value; the correlation matrix in Appendix D shows that it also correlates the most to property value.

CLOSING REMARKS

After finishing this research project, I feel like I've been able to use the empirical techniques that make up Econometrics to analyze economic phenomena and forecast future trends. I've enjoyed developing the hands-on skills and working knowledge necessary to evaluate empirical studies on my own; and I've enjoyed the exposure to standard linear regression, the bread and butter tool of Econometrics. Additionally, I've enjoyed learning how time consuming preparing data for analysis can be as well as writing and presenting that information to others.

Even though I am sure someone else (possibly a professional in the field of real estate) has probably studied this before, it has still been fascinating to witness the paper's discoveries. It's fascinating to me because it shows how difficult it is to use math to make predictions (e.g. dealing with missing variables; transforming raw data into information appropriate for a regression; struggling with analysis when important information is not collected in a survey, etc.) It's fascinating to me as a member of society because one day I may buy a house and it would help to know what affects the value. Finally, this study has been fascinating to me as a researcher because of its possibilities.

Hopefully, research like this will one day be widely available to local officials making decisions about their communities; so they can use it to make better decisions that bring value to them, their constituents, the homes they live in, and their local economy.

REFERENCES

Hamilton, Lawrence C. (2003). *Statistics with STATA: Updated for Version 7*. Duxbury. Belmont, CA.

Gujarati, Damodar (1999). *Essentials of Econometrics (second edition).* Irwin/Mc Graw-Hill. New York.

Betts, Richard M. and Ely, Silas J. (1994). *Basic Real Estate Appraisal (third edition).* Prentice-Halll. Englewood Cliffs, New Jersey.

American Housing Survey. *1998 AHS Metro Survey (content updated 9/15/03).* Retrieved on October 10, 2003 from http://www.huduser.org/datasets/ahs/ahs98metro.html.

Appendix A: Description of Variables

PROPVALUE = VALUE or PVALUE (if greater than VALUE) = Current Value of Property
1:999997 $1-$999,997
999998 $999,998 or more
Long description:
Current market value of this housing unit.

VALUE: The information is collected for all owner-occupied units, but is not collected for renter-occupied units. For owner-occupied units, value represents the respondent's estimate of the property's sale price if it were for sale. For vacant units, value represents the property's sale price at the time of the interview, and may differ from the price at which the property is sold. The variable is available for all owner-occupied units and represents the value of the sample unit and its yard (VALUE). The value of the overall property for multi-family units, structures with commercial/medical establishments, and structures on more than 10 acres are recorded under the variable PVALUE.

PVALUE: This is the price that was paid at the time the property was acquired, not the estimated value at the time of the interview. If only the house is owned, but not the land, the respondent is asked for a combined estimate of the value of the house and lot at the time of purchase. If the house was a single-family unit at the time of purchase, but was split into two or more units since the purchase, the purchase price is the value of the complete structure at the time of the purchase. Purchase price includes the costs of furnishings if the property was acquired furnished. An estimate is accepted if the respondent does not know the exact purchase price. The amount reported excludes closing costs.

AIRSYS = Central air conditioner
1 Yes
0 No
Long description:
Does this heat pump/heating equipment provide air conditioning for this home?
Does this housing unit have central air conditioning?

BATHS = Number of full bathrooms in unit
0:10 Full Bathrooms

BEDRMS = Number of bedrooms in unit
0-10 Full Bedrooms

CELLAR = Unit has a basement
1 Yes (with a basement under all of the house or with a basement under part of the house)
0 No (with a crawl space, on a concrete slab or in some other way)
Long description:
Is this house built with a basement?

CRIMEA = Neighborhood has neighborhood crime
1 Yes
0 No
Long description:
The following questions are concerned with specific aspects of your PRESENT neighborhood.
Does the neighborhood have neighborhood crime?

EABAN = Abandoned/vandalized buildings within 1/2 blk
1 Yes (one or more buildings)
0 No
Long description:
Are there any vandalized or abandoned buildings within half a block of this building?

FPLWK = Unit has useable fireplace
1 Yes
0 No
Long description:
Does this housing unit have a useable fireplace?

GARAGE = Garage or carport included with unit
1 Yes
0 No
Long description:
Is a garage or carport included with this housing unit?

HOUSEAGE = 1998 – BUILT
Year unit was built
1990:2001 1990-2001
1985 1985-1989
1980 1980-1984
1975 1975-1979
1970 1970-1974
1960 1960-1969
1950 1950-1959
1940 1940-1949
1930 1930-1939
1920 1920-1929
1919 1919 or earlier
Long description:
Year of survey - Year this housing unit was built.

HOWN = Rating of neighborhood as place to live
1:10 Rating (10 is best, 1 is worst)
Long description:
How would you rate your neighborhood on a scale of 1-10?

SCH = Neighborhood public elem. school satisfactory
1 Yes
0 No
Long description:
Is the public elementary school for this area satisfactory?

SHP = Neighborhood shopping satisfactory
1 Yes
0 No
Long description:
Do you have satisfactory neighborhood shopping, that is, grocery stores or drug stores?

UNITSF = Square footage of unit
99 99 square feet or less
100:99997 100-99,997 square feet
99998 98,998 square feet or more
Long description:
Thinking about all the rooms you mentioned earlier, as well as the hallways and entry ways in this housing unit, about how many square feet is that? (Include: Finished attics. Exclude: Unfinished attics, carports, and attached garages. Also exclude porches that are not protected from the elements and heated.)

Appendix B: Summary of Variables and Expected Signs

Dependent Variable:

The dependent variable of the regression model is the estimated current value of the property (propvalue) in year 1998.

Independent Variables:

The independent variables that were used to explain property value are listed below.

Housing Unit Variables:

Central air conditioner (airsys):
Homes with heat pump/heating equipment (e.g. central air conditioner) enhance the living conditions in the housing unit. The expected sign is positive.

Number of full bathrooms in unit (baths):
More full bathrooms can accommodate more people in the household living comfortably. The expected sign is positive.

Number of bedrooms in unit (bedrms):
More full bedrooms can also accommodate more people in the household living comfortably. The expected sign is positive.

Unit has a basement (cellar):
A basement under part of or the entire house (excludes crawl space) provides additional storage and can be converted into more livable area in the housing unit. The expected sign is positive.

Fireplace (fplwk):
A fireplace is a well-sought-after feature in a house because it provides comfort; it therefore adds value to the housing unit. The expected sign is positive.

Garage or carport (garage):
A garage or carport included with this housing unit is an added feature protecting the homeowner's means of transportation and/or providing additional storage; it therefore adds value to the housing unit. The expected sign is positive.

Age of unit (houseage):
Older houses tend to have more wear and tear than newer houses. These problems if not fixed will decrease the value of the unit. Houseage is defined as 1998 (year of survey) minus year the housing unit was built. The expected sign is negative.

Square footage of unit (unitsf):
Unit square feet measures the size of the housing unit; the larger the unit, the greater the value. The expected sign is positive.

Local Neighborhood Variables:

Neighborhood has neighborhood crime (crimea):
The existence of crime in the neighborhood usually devalues the property because of increases in risk/cost and decreases in safety and public confidence. The expected sign is negative.

Abandoned/vandalized buildings within 1/2 block (eaban):
Vandalized or abandoned buildings are signs of a neighborhood's quality. Whether a neighborhood is good or bad can easily be determined by observing the buildings in it. Buyers observing negative features like vandalized buildings will buy in another neighborhood or try to negotiate a lower price thus reducing the value of the property. The expected sign is negative.

Rating of neighborhood (hown):
Each homeowner was asked to rate their neighborhood quality on a scale from 1 (worst) to 10 (best). The expected sign is positive.

Neighborhood public elem. school satisfactory (sch):
Satisfaction reflects respondent's happiness with the neighborhood's public elementary school. School satisfaction is an important determinant and adds value to the neighborhood, especially for buyers who have a family. The expected sign is positive.

Neighborhood shopping satisfactory (shp):
Satisfaction in local shopping (grocery stores or drug stores) indicates the homeowner is content with the neighborhood's shopping. Most buyers purchase homes for the long term therefore convenient shopping is an added value. The expected sign is positive.

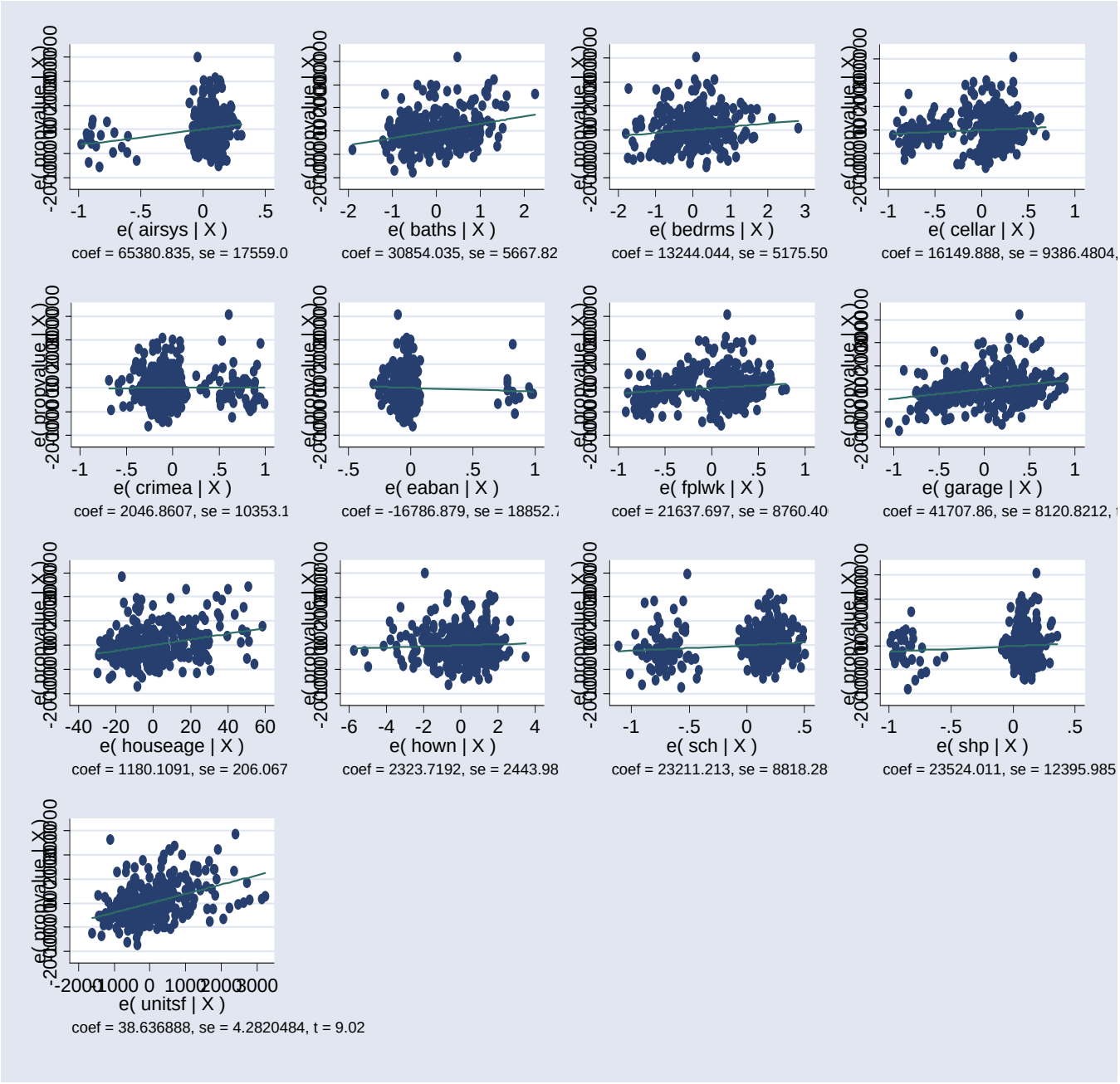
Table of Expected Signs and Transformations:

Concept	Variable	Transformation	Expected Sign
Dependent Variable: Property Value	propvalue	Propvalue = if pvalue > value, use pvalue, otherwise use value	
Central air conditioner adds value to the housing unit.	airsys	Yes = 1 No = 0	+
Bathrooms add value to the housing unit.	baths	None	+
Bedrooms add value to the housing unit.	bedrms	None	+
Basement adds value to the housing unit.	cellar	Yes = 1 No = 0	+
Neighborhood crime increases risk; thus, lowering the property value.	crimea	Yes = 1 No = 0	-
Abandoned/vandalized buildings lower the quality of the neighborhood.	eaban	Yes = 1 No = 0	-
Fireplace adds value to the housing unit.	fplwk	Yes = 1 No = 0	+
Garage adds value to the housing unit.	garage	Yes = 1 No = 0	+
Age of unit increases wear and tear; thus, reducing value of housing unit.	houseage	Houseage = 1998 - Year built	-
Adequacy of neighborhood indicates quality neighborhood.	hown	Yes = 1 No = 0	+
Satisfaction with school indicates quality of neighborhood.	sch	Yes = 1 No = 0	+
Satisfaction with shopping indicates convenience.	shp	Yes = 1 No = 0	+
Square footage of unit adds value to the housing unit.	unitsf	None	+

Appendix C: Graph of Residual against Explanatory Variables

These scatter diagrams plot the residuals against each explanatory variables. As shown below, some of the plots indicate that there is heteroscedasticity in the OLS regression model (e.g. airsys, crimea, eaban).

```
avplots
```



Appendix D: Correlation Matrix

```
correl propvalue airsys baths bedrms cellar crimea eaban fplwk garage houseage hown shp unitsf
(obs=1518)

             | propva~e   airsys    baths   bedrms   cellar   crimea    eaban    fplwk   garage
-------------+---------------------------------------------------------------------------------
   propvalue |   1.0000
      airsys |   0.1218   1.0000
       baths |   0.5260   0.2059   1.0000
      bedrms |   0.4487   0.0978   0.4548   1.0000
      cellar |   0.3136   0.0471   0.2091   0.2567   1.0000
      crimea |  -0.1168  -0.0511  -0.0592  -0.0726  -0.0628   1.0000
       eaban |  -0.0696  -0.0638  -0.0609  -0.0463  -0.0352   0.1134   1.0000
       fplwk |   0.4100   0.1660   0.3343   0.2759   0.2148  -0.0841  -0.0530   1.0000
      garage |   0.4044   0.0963   0.2779   0.2852   0.1612  -0.0949  -0.0713   0.3441   1.0000
    houseage |  -0.0588  -0.3977  -0.3032  -0.1051  -0.0059   0.0883   0.0763  -0.1220  -0.2007
        hown |   0.2069  -0.0238   0.0950   0.0992   0.0746  -0.1881  -0.0632   0.1362   0.1206
         shp |   0.0293   0.0890   0.0225   0.0232   0.0217  -0.0275  -0.0205   0.0277  -0.0136
      unitsf |   0.6334   0.1032   0.4791   0.4957   0.3343  -0.0838   0.0082   0.3495   0.3888

             | houseage     hown      shp   unitsf
-------------+------------------------------------
    houseage |   1.0000
        hown |   0.0049   1.0000
         shp |  -0.0043  -0.0112   1.0000
      unitsf |  -0.1692   0.1625  -0.0091   1.0000
```

Appendix E: Interpretation of Intercept and Coefficients

The Final Regression Equation:

```
Propvalue = – 48920.17 + 29759.11*baths + 8302.856*bedrms + 16960.85*cellar –
            23292.63*eaban + 26979.51*fplwk + 24182.77*garage + 4048.037*hown +
            35.95202*unitsf
```

The intercept of – 48920.17 indicates that if baths, bedrms, cellar, eaban, fplwk, garage, hown, and unitsf variables assumed zero values, the average property value will be –48920.17. This intercept value does not make any economic sense; unless we interpret this figure as "negative" property value. In other words, the government pays 48920.17 to those who do not own a home.

The baths coefficient of 29759.11 indicates that as the number of bathrooms increases by 1, the average property value goes up by 29759.11.

The bedrms coefficient of 8302.856 indicates that as the number of bedrooms increases by 1, the average property value goes by 8302.856.

The cellar coefficient of 16960.85 indicates that if the unit has a basement, the average property value goes up by 16960.85.

The eaban coefficient of –23292.63 indicates that if there are vandalized buildings in the neighborhood, the average property value falls by 23292.63.

The fplwk coefficient of 26979.51 indicates that if the unit has a working fireplace, the average property value goes up by 26979.51.

The garage coefficient of 24182.77 indicates that if the unit has a garage, the average property value goes up by 24182.77.

The hown coefficient of 4048.037 indicates that as the neighborhood rating increases by 1, the average propety value goes up by 4048.037.

The unitsf coefficient of 35.95202 indicates that as the housing unit size increases by 1 square footage, the average property value goes up by 35.95202.

Appendix F: OLS Regression (with Heteroscedasticity)

OLS Regression Model 1:

```
regress propvalue airsys baths bedrms cellar crimea eaban fplwk garage houseage
hown sch shp unitsf
```

```
      Source |       SS       df       MS              Number of obs =     372
-------------+------------------------------           F( 13,   358) =   40.31
       Model |  2.3319e+12    13  1.7938e+11           Prob > F      =  0.0000
    Residual |  1.5930e+12   358  4.4497e+09           R-squared     =  0.5941
-------------+------------------------------           Adj R-squared =  0.5794
       Total |  3.9249e+12   371  1.0579e+10           Root MSE      =   66706
------------------------------------------------------------------------------
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |   65380.83   17559.04     3.72   0.000        30849    99912.67
       baths |   30854.04   5667.827     5.44   0.000     19707.62    42000.45
      bedrms |   13244.04   5175.503     2.56   0.011     3065.835    23422.25
      cellar |   16149.89    9386.48     1.72   0.086    -2309.681    34609.46
      crimea |   2046.861   10353.14     0.20   0.843    -18313.75    22407.47
       eaban |  -16786.88    18852.8    -0.89   0.374    -53863.03    20289.27
       fplwk |    21637.7   8760.401     2.47   0.014     4409.384    38866.01
      garage |   41707.86   8120.821     5.14   0.000     25737.35    57678.37
    houseage |   1180.109   206.0678     5.73   0.000     774.8537    1585.365
       hown |   2323.719   2443.987     0.95   0.342    -2482.656    7130.094
         sch |   23211.21   8818.289     2.63   0.009     5869.056    40553.37
         shp |   23524.01   12395.98     1.90   0.059    -854.0863    47902.11
      unitsf |   38.63689   4.282048     9.02   0.000     30.21576    47.05802
       _cons |    -204984   32605.74    -6.29   0.000    -269106.9   -140861.2
------------------------------------------------------------------------------
```

Note: This OLS Regression Model 1 indicates that basement (cellar) is statistically insignificant; however, Robust Regression Model 1 indicates that basement (cellar) is statistically significant.

OLS Regression Model 2:

```
regress propvalue airsys baths bedrms cellar crimea eaban fplwk garage houseage
hown shp unitsf
```

```
      Source |       SS       df       MS              Number of obs =    1518
-------------+------------------------------           F( 12,  1505) =  146.14
       Model |  8.8387e+12    12  7.3656e+11           Prob > F      =  0.0000
    Residual |  7.5852e+12  1505  5.0400e+09           R-squared     =  0.5382
-------------+------------------------------           Adj R-squared =  0.5345
       Total |  1.6424e+13  1517  1.0827e+10           Root MSE      =   70993
------------------------------------------------------------------------------
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |   16766.86   6913.839     2.43   0.015     3205.082    30328.64
       baths |   34378.25   2945.714    11.67   0.000     28600.11    40156.39
      bedrms |   6967.568   2444.154     2.85   0.004     2173.257    11761.88
      cellar |   14710.95   4416.312     3.33   0.001     6048.167    23373.72
      crimea |  -8907.955   4918.438    -1.81   0.070    -18555.67    739.7655
       eaban |  -24448.08   11842.78    -2.06   0.039    -47678.18    -1217.97
       fplwk |   25763.84   4530.171     5.69   0.000     16877.72    34649.96
      garage |   28108.48   4210.754     6.68   0.000     19848.91    36368.04
```

```
    houseage |   783.4125   102.0916     7.67   0.000     583.1556    983.6693
        hown |    4155.81   1056.758     3.93   0.000     2082.935    6228.685
         shp |    6573.29   6276.142     1.05   0.295    -5737.623     18884.2
      unitsf |    36.6492   2.239074    16.37   0.000     32.25716    41.04123
       _cons |  -97787.89   14424.08    -6.78   0.000    -126081.3   -69494.45
------------------------------------------------------------------------------
```

Note: The difference between Robust Regression Model 2 and OLS Regression Model 2 are the changes in the F value, standard errors, t-scores, p-values, and the confidence intervals. The R-squared, intercept, and coefficients did not change. Both models indicate that neighborhood crime (crimea) and neighborhood shopping satisfactory (shp) are insignificant.

OLS Regression Model 3:

```
regress propvalue airsys baths bedrms cellar eaban fplwk garage houseage hown
unitsf
```

```
      Source |       SS       df       MS              Number of obs =    1537
-------------+------------------------------           F( 10,  1526) =  178.51
       Model |  8.9965e+12    10  8.9965e+11           Prob > F      =  0.0000
    Residual |  7.6906e+12  1526  5.0397e+09           R-squared     =  0.5391
-------------+------------------------------           Adj R-squared =  0.5361
       Total |  1.6687e+13  1536  1.0864e+10           Root MSE      =   70991
------------------------------------------------------------------------------
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |   16121.94   6709.384     2.40   0.016     2961.354    29282.53
       baths |   34404.04   2922.118    11.77   0.000     28672.25    40135.84
      bedrms |   7163.259   2417.969     2.96   0.003     2420.365    11906.15
      cellar |   14574.35   4385.971     3.32   0.001     5971.183    23177.52
       eaban |  -27464.87   11775.47    -2.33   0.020    -50562.68   -4367.061
       fplwk |   25634.42   4484.236     5.72   0.000      16838.5    34430.34
      garage |    28211.3    4182.96     6.74   0.000     20006.34    36416.26
    houseage |   748.5139   101.1766     7.40   0.000     550.0541    946.9738
        hown |   3706.791   954.7946     3.88   0.000     1833.942    5579.639
      unitsf |    36.7025   2.227563    16.48   0.000     32.33309    41.07191
       _cons |  -88600.75   12407.83    -7.14   0.000      -112939   -64262.54
------------------------------------------------------------------------------
```

Note: Both Robust Regression Model 3 and OLS Regression Model 3 indicate that age of unit (houseage) is positive, when it should be negative.

OLS Regression Model 4:

```
regress propvalue airsys baths bedrms cellar eaban fplwk garage hown unitsf
```

```
      Source |       SS       df       MS              Number of obs =    1537
-------------+------------------------------           F(  9,  1527) =  185.73
       Model |  8.7207e+12     9  9.6897e+11           Prob > F      =  0.0000
    Residual |  7.9664e+12  1527  5.2171e+09           R-squared     =  0.5226
-------------+------------------------------           Adj R-squared =  0.5198
       Total |  1.6687e+13  1536  1.0864e+10           Root MSE      =   72229
------------------------------------------------------------------------------
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      airsys |  -1492.596   6382.117    -0.23   0.815    -14011.24    11026.04
       baths |   29866.63    2906.86    10.27   0.000     24164.77    35568.49
      bedrms |   8300.044    2455.17     3.38   0.001     3484.182    13115.91
      cellar |   16960.85   4450.384     3.81   0.000     8231.338    25690.36
       eaban |  -23292.63    11967.1    -1.95   0.052    -46766.33    181.0639
       fplwk |   26979.51   4558.697     5.92   0.000     18037.54    35921.48
      garage |   24210.88   4220.209     5.74   0.000     15932.86     32488.9
        hown |   4045.344   970.3315     4.17   0.000     2142.021    5948.668
      unitsf |   35.94155   2.263998    15.88   0.000     31.50068    40.38243
       _cons |  -47819.67      11310    -4.23   0.000    -70004.44   -25634.89
------------------------------------------------------------------------------
```

Note: Both Robust Regression Model 4 and OLS Regression Model 4 indicate that central air conditioner (airsys) is statistically insignificant when age of unit (houseage) is removed from the regression model. Additionally, this OLS Regression Model 4 indicates that vandalized buildings (eaban) is statistically insignificant; however, Robust Regression Model 4 indicates that vandalized buildings (eaban) is statistically significant.

Final OLS Regression Model:

```
regress propvalue baths bedrms cellar fplwk garage hown unitsf
```

```
      Source |       SS       df       MS              Number of obs =    1551
-------------+------------------------------           F(  7,  1543) =  237.02
       Model |  8.7262e+12     7  1.2466e+12           Prob > F      =  0.0000
    Residual |  8.1155e+12  1543  5.2595e+09           R-squared     =  0.5181
-------------+------------------------------           Adj R-squared =  0.5159
       Total |  1.6842e+13  1550  1.0866e+10           Root MSE      =   72523
------------------------------------------------------------------------------
   propvalue |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       baths |   29459.52   2862.439    10.29   0.000     23844.84     35074.2
      bedrms |   8865.766   2448.626     3.62   0.000     4062.779    13668.75
      cellar |   17223.55   4444.826     3.87   0.000     8505.016    25942.09
       fplwk |   28171.38   4529.653     6.22   0.000     19286.45     37056.3
      garage |   25001.69   4205.818     5.94   0.000     16751.97    33251.41
        hown |   3804.516   905.3954     4.20   0.000     2028.581    5580.452
      unitsf |   35.23598   2.259606    15.59   0.000     30.80376    39.66821
       _cons |  -48540.75    9919.08    -4.89   0.000    -67997.05   -29084.44
------------------------------------------------------------------------------
```

Note: This Final OLS Regression Model does not include vandalized buildings (eaban), whereas the Final Robust Regression Model includes vandalized buildings (eaban).