Economics 312
Sample Project Report
Jeffrey Parker

## *Introduction*

This project is based on Exercise 2.12 on page 81 of the Hill, Griffiths, and Lim text. It examines how the sale price of houses in Stockton, California, are affected by house characteristics including living area, age, and lot size.

## *Data*

The project uses a dataset called stockton4.dta from the authors' collection. The data definition file provided by the authors is reproduced below:

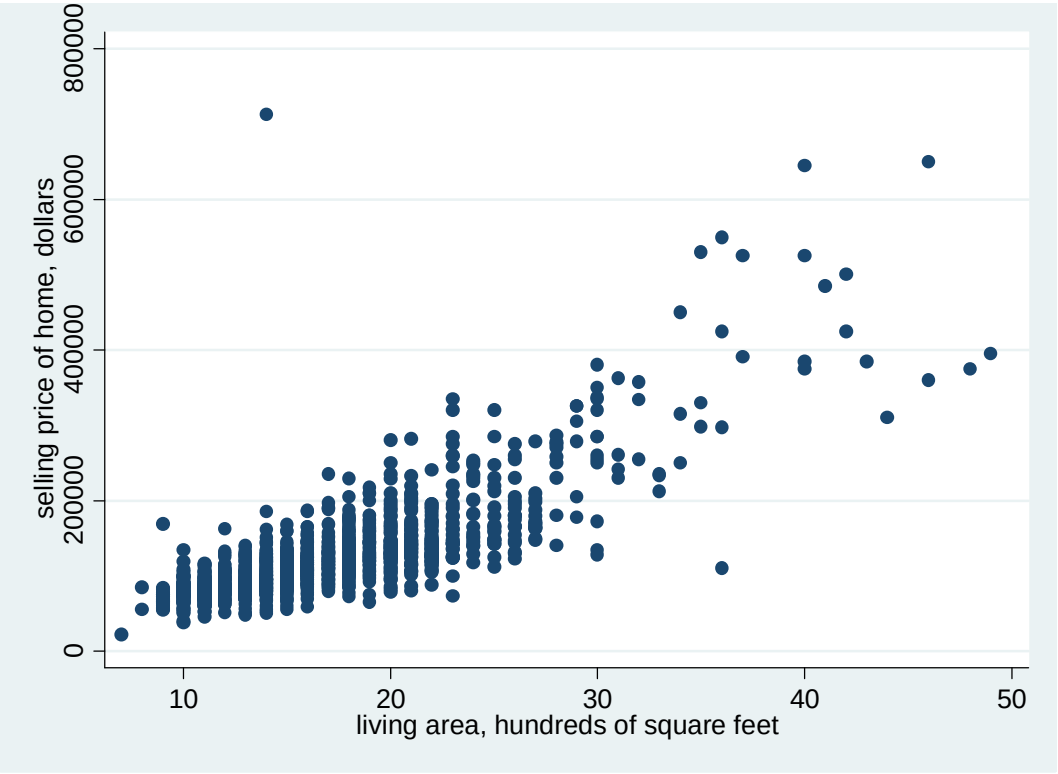
```
sprice livarea beds baths lgelot age pool

  Obs:   1500 home sales in Stockton, CA from Oct 1, 1996 to Nov 30, 1998

  This is a subset of stockton3.dat, the first 1500 observations

       sprice         selling price of home, dollars
       livarea        living area, hundreds of square feet
       beds           number of beds
       baths          number of baths
       lgelot         =1 if lot size > .5 acres, 0 otherwise
       age            age of home at time of sale, years
       pool           =1 if home has pool, 0 otherwise

Data source: Dr. John Knight, Department of Finance, University of the Pacific
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| sprice | 1500 | 123693.9 | 63250.89 | 22000 | 713000 |
| livarea | 1500 | 16.74667 | 5.461963 | 7 | 49 |
| beds | 1500 | 3.285333 | .619818 | 1 | 6 |
| baths | 1500 | 2.133 | .5253523 | 1 | 6.5 |
| lgelot | 1500 | .0633333 | .2436428 | 0 | 1 |
| age | 1500 | 21.86 | 13.11464 | 0 | 97 |
| pool | 1500 | .0653333 | .2471955 | 0 | 1 |

## *Estimating the relationship between house value and living area (Parts a–e)*

**(a)** We begin by examining the bivariate relationship between selling price and living area. The figure below shows that there is generally a positive relationship between the variables—larger houses tend to be more expensive. The relationship is plausibly linear for the smaller houses that comprise most of the observations. There are a few notable outliers, including one small house that

sold for several times the amount of the next most expensive house of similar size. The figure also shows one limitation of the data: the living-area variable appears to be rounded to the nearest hundred.




**(b)** A bivariate regression of selling price on living area yields an estimated slope of 9182. (See regression table below.) This means that an increase of 100 square feet of living area (the size of one small room) raises the expected selling price by $9,182.

The estimated intercept of the linear relationship is –30069. If the same linear relationship extended down to houses of zero size (an empty lot?), then the intercept could be interpreted as the value of such a "null" house. However, there are no houses below 700 square feet in the sample, so such an interpretation entails the dangerous practice of out-of-sample extrapolation. Moreover, given that the estimated intercept is negative, it seems likely that extrapolating the estimated linear relationship to zero living area is unreliable.

The estimated fitted line is graphed below (in part d).

```
. reg sprice livarea
```

```
      Source |       SS       df       MS              Number of obs =    1500
-------------+------------------------------           F(  1,  1498) = 2535.97
       Model |  3.7700e+12     1  3.7700e+12           Prob > F      =  0.0000
    Residual |  2.2270e+12  1498  1.4866e+09           R-squared     =  0.6287
-------------+------------------------------           Adj R-squared =  0.6284
       Total |  5.9970e+12  1499  4.0007e+09           Root MSE      =   38557

------------------------------------------------------------------------------
      sprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     livarea |   9181.711   182.3272    50.36   0.000     8824.067    9539.355
       _cons |   -30069.2   3211.568    -9.36   0.000    -36368.85   -23769.55
------------------------------------------------------------------------------
```

**(c)** Both the scatter plot and the counterintuitive negative intercept suggest that the relationship between price and living area might be nonlinear. We next examine a restricted quadratic form in which price is a linear function of squared living area (livarea2).

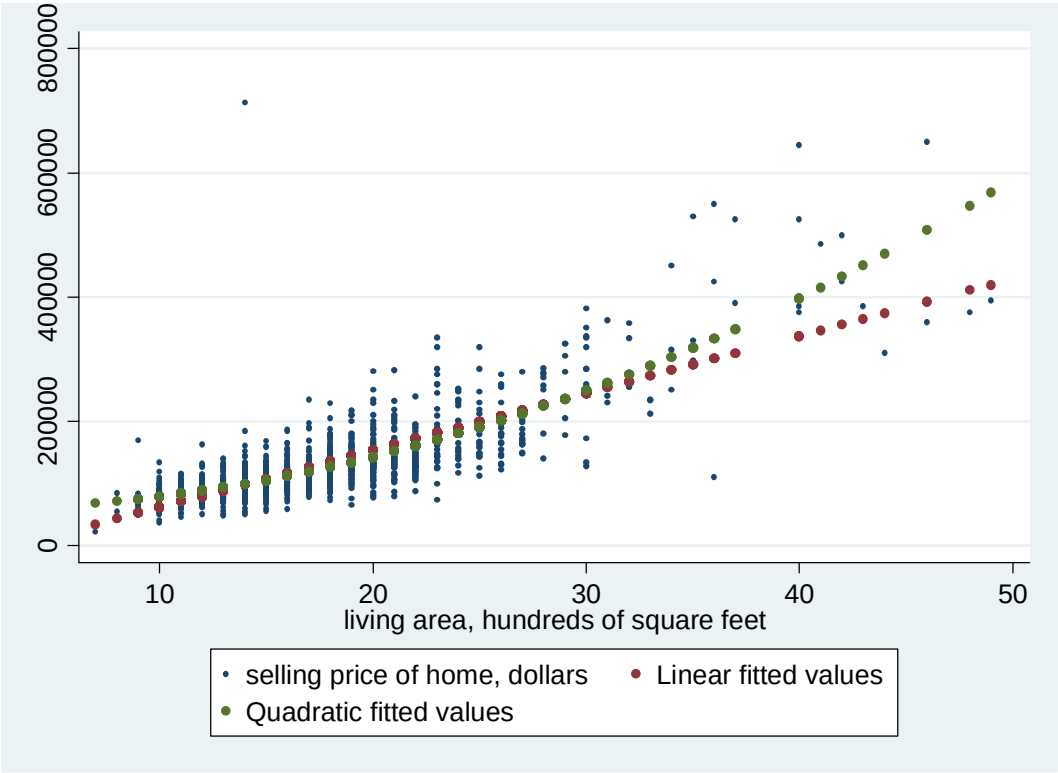
```
. reg sprice livarea2

      Source |       SS       df       MS              Number of obs =    1500
-------------+------------------------------           F(  1,  1498) = 2924.16
       Model |  3.9655e+12     1  3.9655e+12           Prob > F      =  0.0000
    Residual |  2.0315e+12  1498  1.3561e+09           R-squared     =  0.6613
-------------+------------------------------           Adj R-squared =  0.6610
       Total |  5.9970e+12  1499  4.0007e+09           Root MSE      =   36826

------------------------------------------------------------------------------
      sprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    livarea2 |    212.611   3.931742    54.08   0.000     204.8987    220.3233
       _cons |   57728.31    1546.67    37.32   0.000     54694.44    60762.18
------------------------------------------------------------------------------
```

The estimated relationship is once again positive, with an increase of 1 unit in the square living area variable leading to a rise of $213 in price. In order to interpret this more meaningfully, we must examine the marginal effect of living area on price, which is not constant in the nonlinear function:

$$\frac{\partial SPRICE}{\partial LIVAREA} = 2\alpha_2 LIVAREA,$$

where $\alpha_2$ is the coefficient on living area squared. Thus, for a house of 1,500 square feet of living area ($LIVAREA = 15$), the estimated marginal effect is $2 \times 212.6 \times 15 = \$6{,}378$. This would be the estimated effect on expected selling price of an increase of 100 square feet for a house of 1,500 square feet. This is a considerably smaller estimated effect of living area than the $9,181 that we obtained in the linear specification.

**(d)** The graph below shows the scatter plot (blue dots) with the linear (red dots) and quadratic (green dots) fitted values:




It appears that the quadratic function form fits the data slightly better at both the upper and lower extremes of the sample. Although extrapolation outside the sample is always risky, it is also worth noting that the quadratic model predicts a positive value ($57,728) for a lot with a house of zero area, which is more plausible than the negative prediction of the linear model.

The graph also shows that the estimated quadratic function (green) is flatter than the estimated linear function (red) for a 1,500 square foot house, as demonstrated by the smaller estimated marginal effect at that size calculated in part (c).

Because the dependent variable is the same in both models, we can compare the sum of squared residuals to provide further evidence about which model fits the data better. The SSE for the linear model is $2.2270 \times 10^{12}$ whereas the SSE for the quadratic model is $2.0315 \times 10^{12}$. This evidence supports the quadratic model as it has about 10% small sum of squared residuals than the linear model.

**(e)** To examine whether lot size affects the relationship between living space and selling price, we estimate separate quadratic regressions for houses on large lots and those on small lots. The large-lot regression is

```
. reg sprice livarea2 if lgelot==1

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F(  1,    93) =  175.70
       Model |  9.9495e+11     1  9.9495e+11           Prob > F      =  0.0000
    Residual |  5.2663e+11    93  5.6627e+09           R-squared     =  0.6539
-------------+------------------------------           Adj R-squared =  0.6502
       Total |  1.5216e+12    94  1.6187e+10           Root MSE      =   75251

------------------------------------------------------------------------------
      sprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    livarea2 |   193.8298   14.62285    13.26   0.000     164.7917    222.8679
       _cons |   113279.4   12824.64     8.83   0.000     87812.16    138746.6
------------------------------------------------------------------------------
```

For smaller lots, the regression is

```
. reg sprice livarea2 if lgelot==0

      Source |       SS       df       MS              Number of obs =    1405
-------------+------------------------------           F(  1,  1403) = 1749.57
       Model |  1.5997e+12     1  1.5997e+12           Prob > F      =  0.0000
    Residual |  1.2828e+12  1403   914323077           R-squared     =  0.5550
-------------+------------------------------           Adj R-squared =  0.5546
       Total |  2.8825e+12  1404  2.0530e+09           Root MSE      =   30238

------------------------------------------------------------------------------
      sprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    livarea2 |   186.8586   4.467319    41.83   0.000     178.0952    195.6219
       _cons |   62172.41   1503.058    41.36   0.000     59223.92    65120.89
------------------------------------------------------------------------------
```

Before proceeding to compare the results, we note that only 95 of the 1,500 houses in the sample are on large lots, thus we must interpret the results for this subsample with some caution. Based on our sample, additional living area appears to have a larger price effect if the house is on a large lot than if it is on a small lot. (This is consistent with my intuition because big houses fit better on larger lots.)

The estimated coefficient on living area squared is smaller *in both subsamples* than it is in the full sample. When controlling crudely for lot size, house size seems to matter less. This suggests that when we do not control for lot size, the house size variable might be picking up some of the effect that is actually due to lot size. In other words, some of the large houses may be expensive not only because the houses are large, but because the lots are large. This omitted-variable bias would be present if large houses tend to be on large lots (and lot size matters for selling price). The result reported below verifies that the average living area on large lots is 2,463 square feet compared to

1,621 square feet on smaller lots, supporting our result that failure to correct for lot size leads to an overestimation of the effect of house size.

```
. by lgelot: summarize livarea

-------------------------------------------------------------------------------
-
-> lgelot = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     livarea |      1405    16.21352    4.585679          7         49

-------------------------------------------------------------------------------
-
-> lgelot = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     livarea |        95    24.63158    9.724998          8         48
```

A logical way to compare the marginal effects of living space in the two subsamples is to evaluate each at the subsample mean. For large lots, the mean of *LIVAREA* is 24.63 and the coefficient on $LIVAREA^2$ is 193.8, so the marginal effect evaluated at the mean is $2 \times 193.8 \times 24.63 =$ \$9,549. For smaller lots, the corresponding calculation is $2 \times 186.9 \times 16.21 =$ \$6,059. Thus, an additional 100 square foot room is worth much more if added to a house on a large lot than to one on a smaller lot.

### *Other determinants of house selling price (parts f and g)*

**(f)** The age of a house may also influence its selling price. The regression below shows that the linear relationship between selling price and age is negative: a house that is one year older is expected to sell for \$627 less. The intercept of 137,404 could be interpreted as the expected selling price of a new house (*AGE* = 0). There are some houses of age zero in the sample, so this is not an out-of-sample extrapolation.

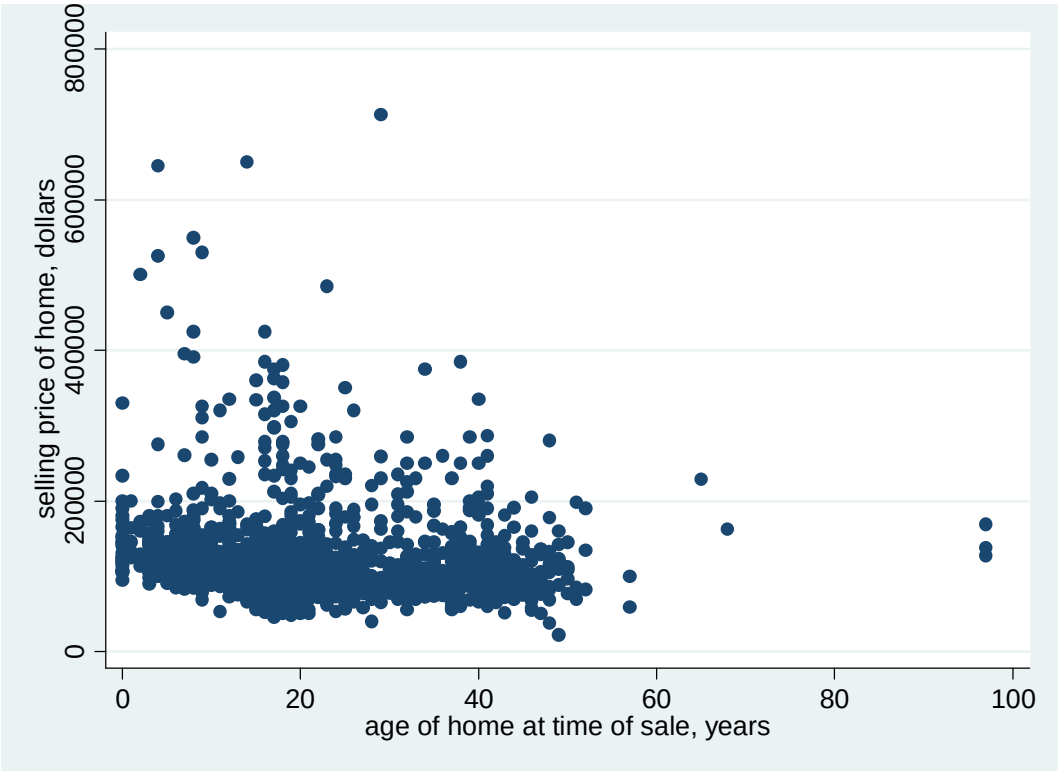
```
. reg sprice age

      Source |       SS       df       MS              Number of obs =    1500
-------------+------------------------------           F(  1,  1498) =   25.77
       Model |  1.0141e+11     1  1.0141e+11           Prob > F      =  0.0000
    Residual |  5.8956e+12  1498  3.9357e+09           R-squared     =  0.0169
-------------+------------------------------           Adj R-squared =  0.0163
       Total |  5.9970e+12  1499  4.0007e+09           Root MSE      =   62735

------------------------------------------------------------------------------
      sprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -627.161   123.5524    -5.08   0.000     -869.515   -384.8069
       _cons |   137403.6   3149.347    43.63   0.000       131226    143581.2
------------------------------------------------------------------------------
```

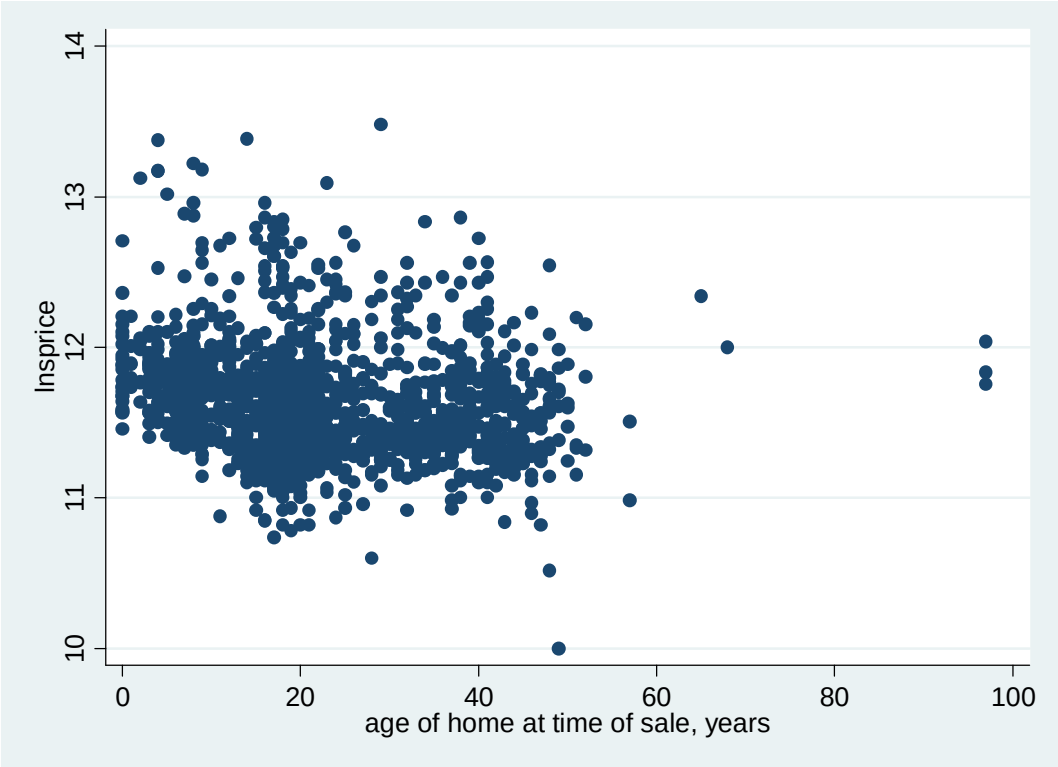
An alternative functional form for the relationship would be to regress the log of selling price on age. This is done in the table below.

```
. g lnsprice = ln(sprice)

. reg lnsprice age

      Source |       SS       df       MS              Number of obs =    1500
-------------+------------------------------           F(  1,  1498) =   41.45
       Model |  5.84157999     1  5.84157999            Prob > F      =  0.0000
    Residual |  211.122211  1498  .140936055           R-squared     =  0.0269
-------------+------------------------------           Adj R-squared =  0.0263
       Total |  216.963791  1499   .14473902           Root MSE      =  .37541

------------------------------------------------------------------------------
    lnsprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    -.00476   .0007394    -6.44   0.000    -.0062103   -.0033097
       _cons |   11.74597   .0188462   623.26   0.000     11.70901    11.78294
------------------------------------------------------------------------------
```

Again, the relationship between selling price and age is negative. The estimated coefficient of –0.00476 can be interpreted as reflecting a 0.476% reduction in selling price for each additional year of age.

The first scatter plot below shows the relationship between selling price and age. It is a large cluster of points without an obvious linear relationship and with numerous outliers. Moreover, all of the outliers are above the cluster. The second plot shows the relationship between log price and age, which seems slightly more linear and where the outliers occur on both sides. Based on visual fit, I'd prefer the log model.


selling price of home, dollars
0
200000
400000
600000
800000
0
20
40
60
80
100
age of home at time of sale, years


lnsprice
10
11
12
13
14
0
20
40
60
80
100
age of home at time of sale, years

**(g)** Finally, we estimate the relationship between price and the dummy variable indicating lot size. We established above that controlling for lot size by splitting the sample had a large effect on the estimated coefficient for living area, which suggests that lot size may be an important determinant of selling price. Because of the binary nature of the lot size variable, we are restricted to a rather crude way of characterizing the relationship between lot size and selling price, but our model should indicate the direction effectively. The results, shown in the table below, demonstrate the expected relationship. Lots larger than ½ acre are expected to sell for $133,797 more than those on smaller lots. The estimated intercept term of $115,220 is the expected selling price of a house on a small lot, where *LGELOT* = 0.

```
. reg sprice lgelot

      Source |       SS       df       MS              Number of obs =    1500
-------------+------------------------------           F(  1,  1498) =  541.83
       Model |  1.5930e+12     1  1.5930e+12           Prob > F      =  0.0000
    Residual |  4.4041e+12  1498  2.9400e+09           R-squared     =  0.2656
-------------+------------------------------           Adj R-squared =  0.2651
       Total |  5.9970e+12  1499  4.0007e+09           Root MSE      =   54221

------------------------------------------------------------------------------
      sprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lgelot |   133797.3   5747.992    23.28   0.000     122522.4    145072.3
       _cons |     115220   1446.546    79.65   0.000     112382.5    118057.5
------------------------------------------------------------------------------
```

## *Conclusions and validity assessment*

The sample of 1500 houses in Stockton yields estimates of the effects of living area, lot size, and age of home that confirm intuition. Larger and newer houses and larger lots seem to be associated with higher selling prices.

The assumptions of the least-squares model seem reasonable for this application, with some qualifications:

- **Autocorrelation** between the error terms of nearby houses could be a problem: there may be unobserved neighborhood characteristics that would affect all houses in an area in a similar way. Correcting for this, if possible, might give more efficient estimators, though the OLS coefficient estimators are still unbiased.
- **Reverse causality** does not seem to be a worry here. The characteristics of the house that we have used on the right-hand side are determined before the sale, so it is unlikely that random variation in the selling price would have any influence on them.
- **Homoskedasticity and normality**. It is more problematic to assume that all houses in the sample have equal error variance. Big houses that are expected to sell for $500,000 would logically have a larger error variance than smaller houses with expected price of $150,000. These considerations of the distribution of the error term point toward using the

log of selling price as the dependent variable because I think it is more likely that *percentage* deviations from expected selling price are symmetrically distributed with constant variance than the *dollar* deviations. This means that the assumptions of homoskedasticity and normality are more plausible when using log(price) than when using price. Failure to account properly for heteroskedasticity leads to inefficient, but still unbiased, coefficient estimators.

- **Omitted variables.** Because we have identified several characteristics that seem to be associated with selling price, it would be useful to include all of them in the same regression equation using multiple regression. This would mitigate possible issues of omitted-variables bias such as the one we identified when estimating the effects of living area separately for large and small lots. Such omitted-variable bias manifests itself as correlation between the regressor and the error term that includes the effect of the omitted variable.
- **Functional form.** The fact that the quadratic (for living area) and logarithmic (for age) functions fit better than linear models suggests that some additional exploration of nonlinear forms might be appropriate.

With respect to external validity, we have no information on how the sample was drawn, so it is difficult to assess whether the sample accurately represents real estate in Stockton, California. The sample is now more than a decade old, so any applications to the modern housing market would need to be updated for inflation. Moreover, it is possible that the value attached to specific house characteristics may be different now than in 1996–98.

It is difficult to know how the effects of home characteristics on prices in Stockton might differ from those in other cities. One would expect some differences in effects across cities associated with variations in demographics, topographical features, and population density, so it would be unwise to apply specific results from this study to other urban areas.