Top Kaggler
STAT E-104 Project

Jeff Collier*

4 MAE 2015

The Challenge

Given some data on home sales, we're to demonstrate our ability at modeling using linear regression. Building on that, use the model to estimate the prices of houses represented by a smaller set of data as part of a challenge.

The only real trick to cooking is to pick good ingredients and make sure you don't screw them up. One mantra from the TV series Top Chef is if it didn't work, don't serve it. I have a feeling those may be conflicting goals modeling these data. Still, they are the guidelines that will inform my actions. That said, let's take a look at what we have in the pantry.

The Data

The data is described in Table 1. I renamed some of the original variables because LaTeX sucks at random underscores. One of the first steps to a regression analysis is to look at each variable. I'll break the data up into groups and get a taste of each of them individually.

salesprice, pl

Let's start with a taste of what we are trying to predict, *salesprice*. Some strong hints and exploratory regressions indicated that transforming *salesprice* might be a good idea. Stata's reports via detailed summary that the skewness of *saleprice* is > 0.97; this is greater than our 0.8 rule of thumb for requiring transform. Regression estimates are really the mean of a normal distribution. If the actual underlying data is skewed, our estimate are more likely to be more off. It was nigh on impossible to get exploratory regressions of *salesprice* to pass Stata's *sktest*.

We see from Figure 1 that *salesprice* is skewed towards larger values and has a few outliers. We will leave outliers in the mix for now: removing them would restrict the range of values we could predict for.

Given that it is likely we'll need to eventually transform *salesprice* we will use $pl = ln(salesprice)$ as our dependent variable going forward. We see from Figure 2* that a natural log transform gives us a roughly normal fit.

This will force us to adjust our interpretations of the coefficients†. Using the log means coefficients represent the percentage change in our dependent variable based on change in the predictor. But this is better than restricting the range of values our model can reasonably predict for.

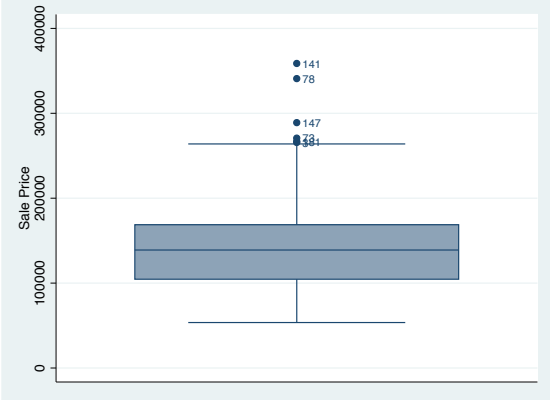

Figure 1: salesprice

*but on Kaggle they call me Maurice

*This plot gives us a line that, if the distribution were normal, the scattered points would fall on or near.

†I learned what I know of this from "http://www.biostathandbook.com/transformation.html" but all misunderstandings are my own.

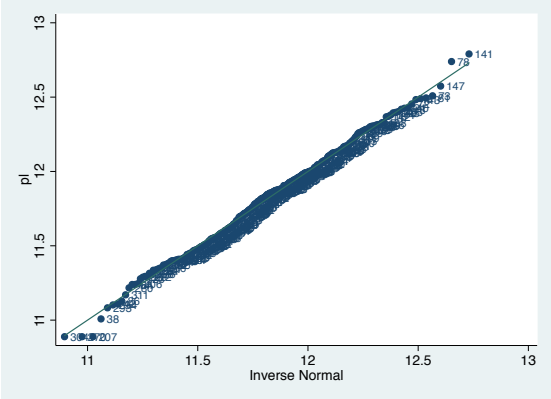

Figure 2: Transformed Sales Price

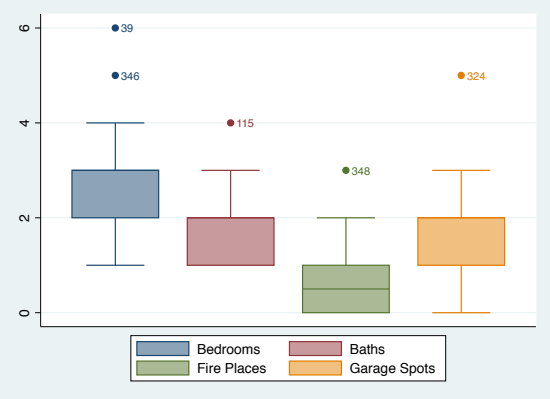

Figure 4: Bedrooms, baths, fire places, garage spots

Month Sold

Month sold would seem to be a better predictor of volume of sales, not price. Examining sales price by month in Figure 3 shows a more uniform distribution of sales price. Both the middle 50% of the price of homes sold and the remaining data all overlap to some degree, many to a large degree. I plan on ignoring this variable.

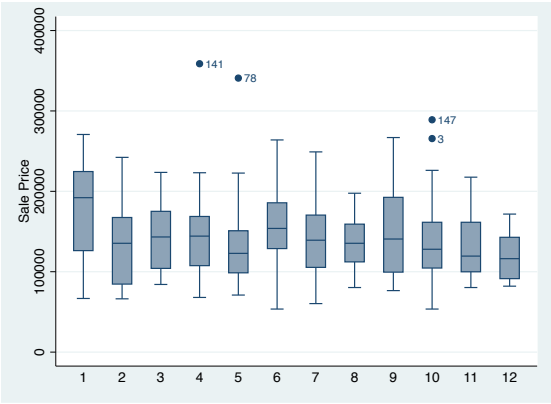

Figure 3: Sales price by month sold

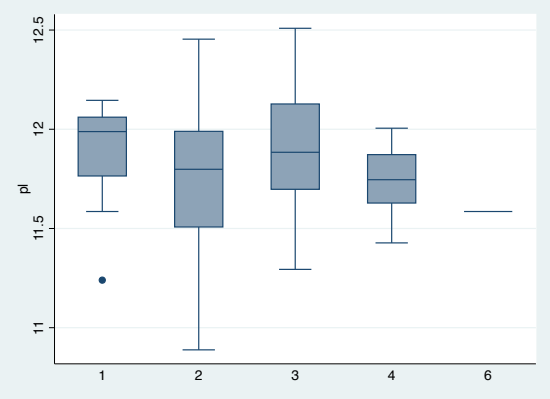

Figure 5: Variance of pl over bedrooms

The Counts

Baths, bedrooms, fire places and garage spots: these counts would all seem likely predictors of sales price, worth the trouble of mining. I expect that the sale price would increase as these values increase. But Figure 4 shows that there isn't much variance in these data. Also, the distribution fire places is skewed, so I added an indicator variable, fpp, if it has one or more. Bed rooms is also a curiosity: observe the variance of pl over each value in Figure 5. This explains why this variable had a negative coefficient in exploratory regressions.

Age

Age is a continuous variable that negatively affects the price of the house, as seen in Figure 6. Transforming by ln only seems to have traded one curve for another (Figure 7), but exploratory regressions using the two different variables show that while *age* doesn't pass Stata's ovtest and hettest for linearity and heteroscedasticity, *aget* is suitable for regression on its own.

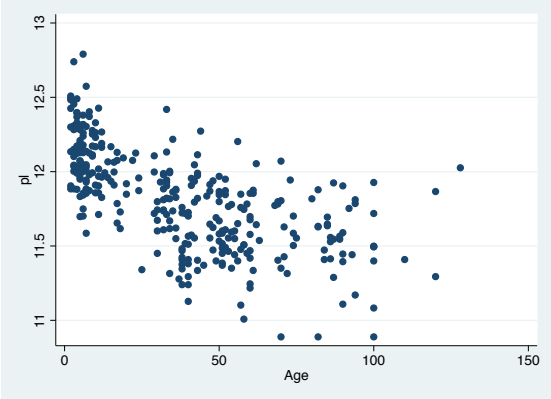

Figure 6: Curvy scatter of Age

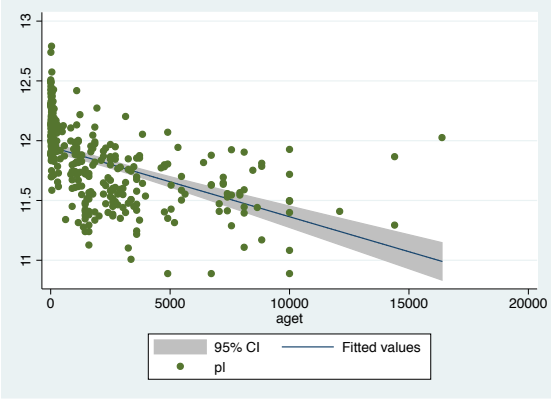

Figure 7: Curvy scatter of transformed Age

Central Air Conditioning

While most homes have this, there is clearly a differential in the Figure 8. This was confirmed by a two sample t-test. Given H_0 : $\mu_{cac=1} = \mu_{cac=0}; H_a : \mu_{cac=1} > \mu_{cac=0}$ resulted in $p < 0.0001$ which means we reject the null hypothesis; homes with central air sell for more. This could be a useful indicator variable but only five of 319 homes in modeling data don't have it. While the three "outliers" shown on this graph represent price outliers, they were culled since they won't help stata with this variable.

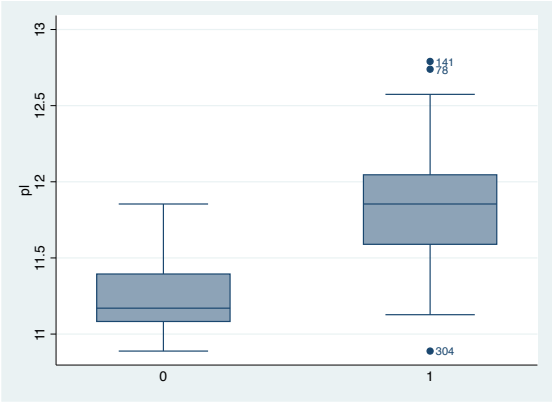

Figure 8: Central Air

Condition

I had high hopes for this but it has some serious trouble (Figure 9. Note the scores for *condition* = 5: We see nice, similar-ish variances for the other conditions, gradually creeping up in price. Perhaps someone thought the scale was 1-5, or there was some other data error. I will introduce a hack, recoding 1=3, 2=4, 3=6, 4=7 and 5=5,8,9 and call it recond, then break that up into bar1-5.

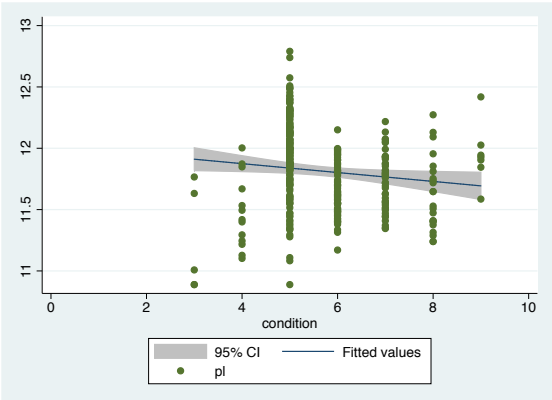

Figure 9: Condition

Basement Square Feet

I had high hopes for this one. A continuous variable with not many outliers (Figure 10), exploratory regressions show it loves to generate heteroscedasticity. It also failed Stata's ovtest, so I tried a number of transforms, partitioning the basement sizes using ttest ... , by(), etc., all to no avail. We'll see how it affects heteroscedasticity in the multiple regression model: if we see heteroscedasticity in that model, we'll pull basement and see what improvements we get.

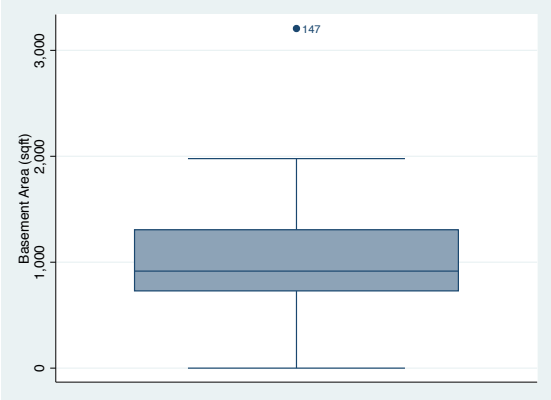

Figure 10: Basement Square Feet

Lot Size

Another continuous variable with high hopes, the box plot (Figure 11) shows some problems. Attempts at tranformations (ln, etc.) did not help. Attempts at dealing with outliers resulted in exacerbated skewness, and while not terrific at nearly 0.77, it is less than 0.8. I was able to remove four worst from these diagrams by selecting $lotSize > 19000$. One off the cuff test was a ttest of lot sizes between our modeling building set of data and our kaggle data: Stata says the μ of lot size in the Kaggle data is smaller. This supports the decision to model without the larger lots, hopefully trading overall accuracy of the model for a model that will guess better at the unknown observations I have in front of me.

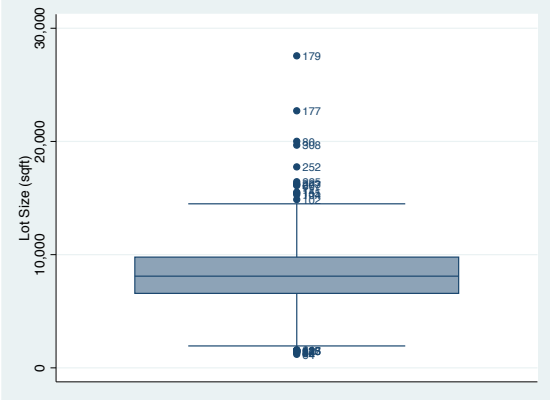

Figure 11: lotSize Box

Living Area

A continuous variable that looks good (Figure 12), exploratory regressions show it failed Stata's ovtest. You can see that it doesn't fit well to the normal distribution shown in Figure 13. I eventually transformed this using square root, which seems to generate a better model even though it still doesn't quite pass the ovtest. sktest reports both are not normally distributed; we'll serve up the less doctored one in the final dish and experiment with the transformed if it doesn't work well.

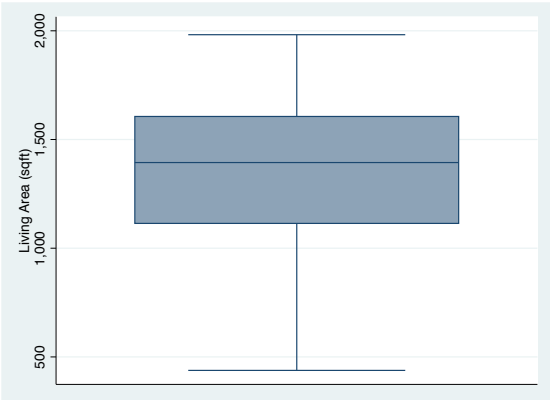

Figure 12: Living Area

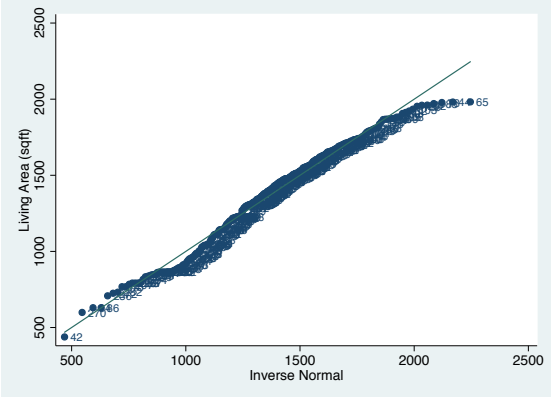

Figure 13: Living Area, normal distribution?

Deck

Everyone loves a deck, no? Alas, another continuous variable but no real correlation with price (Stata reports a correlation < 0.27). The variation on this variable is particularly screwy, in particular note the variance on "no deck" spans all prices (Figure 14).

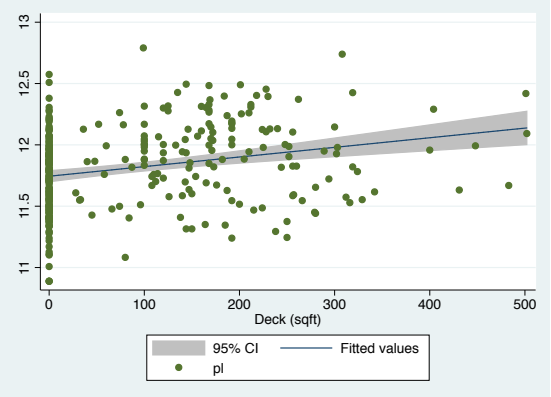

Figure 14: deck scatter

Miscellaneous

A scatter plot of $distFire$, the distance to the fire department, shows an amorphous blob. Stata reports a correlation between $distFire, pl$ of -0.0559; we can safely ignore this variable. Likewise, Stata's ttest fails to reject the null hypothesis that there's no difference for prices by *smoker*; the correlation with price is 0.0572. Pass.

Recap

This feels like the junk food challenge on Top Chef, since much of the data is skewed or has strange associations with our dependent variable. We'll cobble together some recipes and see what we can serve up.

Kitchen Sink

Now throw it all into one large regression. After the first iteration, work to exclude residuals with an absolute value > 2 and influential observations via Cook's Distance, then rerun. We get a pretty good run, all things considered, with Listing 1. The p-value (Prob > F) is < 0.0001, so we need at least some of our variables in the model. Each variable is confirmed by its respective p-value < 0.05. $r^2 = 0.8814, r^2adjusted = 0.8765$: This model accounts for 88.14% of the variability in pl, and $r^2adjusted$ being close to r^2 tells us that this isn't just because we tossed a lot of ingredients into the blender. $s_e = 0.10754$, still within spitting distance of what I was able to achieve with some questionable experiments.

Looking to the residuals, we observe that 3.2% of our standardized residuals are greater than two standard deviations from our prediction, we'd expect up to 5%. H_0 : Residuals are normally distributed; H_a : Residuals are not normally distributed, Stata's Skewness/Kurtosis test reports a p-value of 0.1282. We fail to reject the null hypothesis, our residuals look to be normally distributed, confirming one assumption of regression. We also test H_0 : The model fits a straight line; H_a : The model fits some other line. Stata's Ramsey Reset test reports a p-value of 0.2675, we fail to reject the null hypothesis and confirm another assumption of regression.

Finally we test for heteroscedasticity. H_0 : Variance is constant across predicted values; H_a : variance is not constant, Stata's Breusch-Pagan / Cook-Weisberg test reports a p-value of

0.3771, again we fail to reject and this model looks good.

There are some troubles. First, there are a lot of variables. And a lot of this data is past its sell-by date. We also had to cull a lot of observations– nearly 70 of 320. The MAE score for modeled data was roughly 12,000, but non-modeled data rolled in over 20,000. Finally, there are a lot of variables in this model to explain and some of them (br) have nonintuitive explanations to make. I pass.

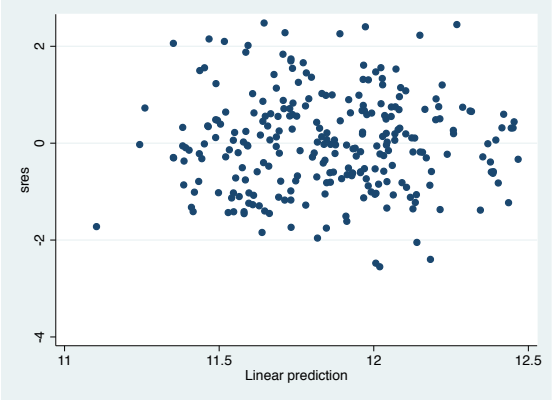

Figure 15: Residuals vs. Fit for Candidate

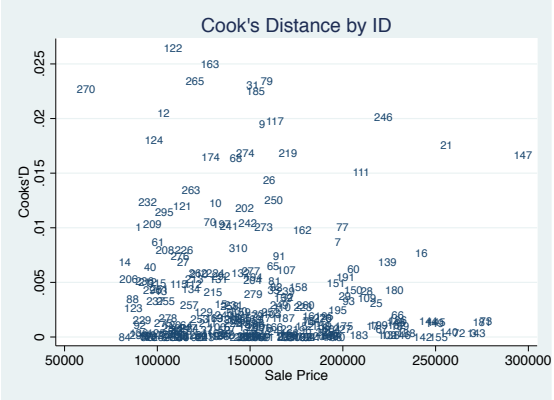

Figure 16: Cooks Distance for Candidate

Good Knife Work?

So I decided to make my own regression choices and attempt to find a more parsimonious model. One thing I was hoping to not lose so many observations to outliers in predictor variables. There was some initial success as I had 18 more observations to start, but 10 came back out after looking at Cook's Distance and large residuals (> 2.5). Still, I removed less than 20% but still a lot.

Technical Analyis

Technically, the model is sound, noted in Listing 2. The p-value (Prob > F) is < 0.0001, so we need at least some of our variables in the model. Each variable is confirmed by its respective p-value < 0.05. $r^2 = 0.8367, r^2adjusted = 0.8327$: This model accounts for 83.67% of the variability in pl, and $r^2adjusted$ being close to r^2 reassures us again. $s_e = 0.12475$, still within spitting distance of our sink model and what I was able to achieve with some questionable experiments.

5.4% of our standardized residuals (Figure 15) were greater than two, casting some niggling doubt on the predictive power of our model.

H_0 : Residuals are normally distributed; H_a : Residuals are not normally distributed, Stata's Skewness/Kurtosis test reports a p-value of 0.5008. We fail to reject the null hypothesis, our residuals look to be normally distributed, confirming one assumption of regression. We also test H_0 : The model fits a straight line; H_a : The model fits some other line. Stata's Ramsey Reset test reports a p-value of 0.5875, we fail to reject the null hypothesis and confirm another assumption of regression.

Finally we test for heteroscedasticity. H_0 : Variance is constant across predicted values; H_a : variance is not constant, Stata's Breusch-Pagan / Cook-Weisberg test reports a p-value of 0.2974, again we fail to reject and this model looks good.

Interpretation

Our model predicts the natural log (x) of $salesprice$:

$x = 11.01223 - 0.1221579 \times aget + 0.0205386 \times sqftt + 0.1141497 \times bar4 + 0.0000216 \times lotsize + 0.0466525 \times fpp + 0.1355746 \times garage$

Finally, $salesprice = e^x$.

It can be interpreted as follows: The sale price of a brand new house on an infinitesimally small lot with no living space, panned for its condition, no fireplace, no garage would be $e^{11.01223} = \$60,610.90$, on average. For every 1% increase in age, while all other variables remaining the same, the price of the house decreases 12.2%, on average. As square feet increase by squares $(1, 4, 9, 16, 25, \ldots)$, while all other variables remaining the same, the price of the house increases 2.05%, on average. For each unit increase in condition, while all other variables remaining the same, the price of the house increases 11.41%, on average. For each additional square foot in the size of the lot, while all other variables remaining the same, the price of the house increases 0.0022%, on average. Fireplaces will add 4.66% to the price of the house, on average, all other variables remaining the same. Each additional garage space adds 13.56% to the price of the house, on average, all other variables remaining the same.

Creating the the CI for this estimate is a little more difficult. It's still $\pm 1.96 \times 0.12475$ but this is to the natural log, the actual CI will be asymmetric because you adding/subtracting to the exponent of e. The lower bound will be closer to the estimate than the upper bound.

Notes and Caveats

I did predict for a 100 year 200 sqft house on a 500 sqft lot with nothing and got a prediction $\$42,844$.

I would like to research and understand better the idea of a percentage increase for things like "if fireplaces" and garage spots. This seems to run a counter to what I've seen.

This model predicts meh for observations from the model, with a MAE of $14,224$, and less than meh for observations not from the model, a MAE of 28036 from the model data that I dropped, and $17,269$ for the Kaggle data.

The model seems rotated counter-clockwise. I found the median sale price and counted the large residuals ($> 2, < 2$) by price greater and price less than. I see that for prices less than the median, gross underestimates outnumber the overestimates 15-4. For prices greater, the overestimates win 13-5[‡]

Armed with this knowledge and some healthy skepticism with regard to overfitting, I would prefer this model to the stepwise model shown earlier.

[‡]Using all modeling building data, whether I actually included it in my regression model or not.

Listing 1: Stepwise Multiple Regression, Kitchen Sink

```
. stepwise , pr(0.05) : regress pl  aget bar2 bar3 bar4 basement baths br cac recond
    deck distFire fp garage lotSiz
> e monthSold smoker sqftt if model
                      begin with full model
p = 0.8812 >= 0.0500  removing monthSold
p = 0.8667 >= 0.0500  removing bar2
p = 0.1947 >= 0.0500  removing deck
p = 0.1711 >= 0.0500  removing distFire
p = 0.2234 >= 0.0500  removing baths
p = 0.0698 >= 0.0500  removing cac
p = 0.0734 >= 0.0500  removing smoker

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F( 10,   239) =  177.66
       Model |  20.5458051    10  2.05458051           Prob > F      =  0.0000
    Residual |  2.76395908   239  .011564682           R-squared     =  0.8814
-------------+------------------------------           Adj R-squared =  0.8765
       Total |  23.3097642   249  .093613511           Root MSE      =  .10754

------------------------------------------------------------------------------
          pl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aget |  -.1058091   .0094444   -11.20   0.000    -.1244141   -.0872041
     lotSize |   .0000131   2.45e-06     5.32   0.000     8.22e-06    .0000179
        bar3 |   .0695827    .029167     2.39   0.018     .0121255    .1270399
        bar4 |   .1370819   .0247909     5.53   0.000     .0882453    .1859184
    basement |   .0001652   .0000215     7.67   0.000     .0001227    .0002076
      garage |   .0908664   .0150552     6.04   0.000     .0612085    .1205242
          br |  -.0603909   .0134081    -4.50   0.000     -.086804   -.0339778
          fp |   .0285014   .0128621     2.22   0.028     .0031639     .053839
      recond |   .0359796   .0130071     2.77   0.006     .0103564    .0616029
       sqftt |    .027752   .0023803    11.66   0.000      .023063     .032441
       _cons |   10.67145   .1089018    97.99   0.000     10.45692    10.88598

                     Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
        sres |    250      0.7106         0.0471         4.11         0.1282

Ramsey RESET test using powers of the fitted values of pl
       Ho:  model has no omitted variables
                 F(3, 236) =      1.32
                  Prob > F =      0.2675

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

         chi2(1)      =     0.78
         Prob > chi2  =   0.3771
```

Listing 2: Manual Multiple Regression

```
. regress pl aget sqftt bar4 lotSize fpp garage if model

      Source |       SS       df       MS              Number of obs =     258
-------------+------------------------------           F(  6,   251) =  214.27
       Model |   20.007548     6  3.33459133           Prob > F      =  0.0000
    Residual |   3.9062121   251  .015562598           R-squared     =  0.8367
-------------+------------------------------           Adj R-squared =  0.8327
       Total |  23.9137601   257   .09304965           Root MSE      =  .12475

------------------------------------------------------------------------------
          pl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aget |  -.1221579   .0096961   -12.60   0.000     -.141254   -.1030619
       sqftt |   .0205386   .0023146     8.87   0.000     .0159801    .0250972
        bar4 |   .1141497   .0249036     4.58   0.000      .065103    .1631965
     lotSize |   .0000216   2.61e-06     8.27   0.000     .0000165    .0000268
         fpp |   .0466525   .0175641     2.66   0.008     .0120608    .0812443
      garage |   .1355746   .0166735     8.13   0.000     .1027369    .1684123
       _cons |   11.01223   .0885233   124.40   0.000     10.83789    11.18657

                   Skewness/Kurtosis tests for Normality
                                                      ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
        sres |    258      0.2688         0.6997         1.38         0.5008

Ramsey RESET test using powers of the fitted values of pl
       Ho:  model has no omitted variables
                F(3, 248) =      0.64
                 Prob > F =      0.5875

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

         chi2(1)      =     1.09
         Prob > chi2  =   0.2974
```

Table 1: Data Description

Variable	Original	Description
ID		data driven observation ID
salesprice		Selling price in dollars
pl	new transform	natural log of salesprice
sqft	live_area_sqft	Square footage of living space
sqftt	new transform	$\sqrt{sqft}$
basement	bsmt_sf	Square footage of basement
lotSize	lot_sq_ft	Size of lot in square feet
lotSizet	new transform	ln of lotSize
deck	deck_sf	Size of deck in square feet
dp	new dummy	$deck > 50$ square feet?
age		Age of the house in years
aget	new transform	natural log of age
distFire		Distance from the closest fire station
baths		Count of bathrooms
bedrooms		Count of bedrooms
fp	fireplaces	Count of fireplaces
fpp		Has one or more fireplaces
garage	garage_cars	Count of spaces in garage
condition		Rated conditon 1-10
recond	recode	Recoded condition, 1-5
bar$n$	recode	Indicators for recode
cac		Boolean, central air
smoker		Boolean, previous owner smoked
monthSold	month_sold	Month of year sold 1-12