Soc 63993, Homework #3 Answer Key:
Random Measurement Error/ Heteroscedasticity/ Outliers

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised January 22, 2015

Part I. Random Measurement error.

a. Enter the following commands into Stata. If in doubt as to what a command does (e.g. `drawnorm`), either use Stata's help command or consult the online reference manuals. Yt and Xt are assumed to be perfectly measured, whereas X suffers from random measurement error.

```
version 12
set seed 123456789
set obs 1000
drawnorm Yt Xt e, corr(1 .5 0 \ .5 1 0 \ 0 0 1)
reg  Yt Xt
gen X = Xt + e
reg  Yt X
```

Compute the reliability of X. (There are at least two or three ways to do this, and because a sample is being used the results won't be exactly identical.) Explain why the results of the two regressions differ, and why researchers should be concerned about this. Would a larger sample size take care of this problem? Why or why not? You can, of course, also enter any other commands that will help you to answer this question. If in doubt, feel free to explore, e.g. you could try creating larger or smaller samples and see what happens when you rerun the same commands.

Here are the results from running the program:

```
. version 12
. set seed 123456789
. set obs 1000
obs was 0, now 1000
. drawnorm Yt Xt e, corr(1 .5 0 \ .5 1 0 \ 0 0 1)
. reg  Yt Xt

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  1,   998) =  334.48
       Model |  264.955975     1  264.955975            Prob > F      =  0.0000
    Residual |  790.555131   998   .79213941           R-squared     =  0.2510
-------------+------------------------------           Adj R-squared =  0.2503
       Total |  1055.51111   999  1.05656767           Root MSE      =  .89002

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |   .5180475   .0283259    18.29   0.000     .4624623    .5736326
       _cons |   .0214104   .0281517     0.76   0.447    -.0338331    .0766538
------------------------------------------------------------------------------

. gen X = Xt + e
```

```
. reg  Yt X

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  1,   998) =  134.27
       Model |  125.166627     1  125.166627           Prob > F      =  0.0000
    Residual |   930.34448   998  .932208898           R-squared     =  0.1186
-------------+------------------------------           Adj R-squared =  0.1177
       Total |  1055.51111   999  1.05656767           Root MSE      =  .96551

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .2566741   .0221511    11.59   0.000     .2132061    .3001421
       _cons |   .0099768   .0305321     0.33   0.744    -.0499376    .0698913
------------------------------------------------------------------------------
```

The `version 12` and `set seed` commands make it possible for us to reproduce these results when we run the program again. With the `set obs` and `drawnorm` commands, we create a random sample of 1,000 cases drawn from a population where Xt and Yt have a .5 correlation and both have 0 correlation with the random error term, e. Since means and sds are not specified, `drawnorm` uses the default mean of 0 and default sd of 1 for every variable. Since V(Xt) = V(e) = 1, V(Xt + e) = V(Xt) + V(e) = 2. Reliability = true variance/total variance = ½ = .5. Sampling variability causes the sample to differ a little from the population, but in this case you know the population values so you don't need to use sample estimates when computing the reliability. But, if you didn't know the population values, you could estimate the reliability in the sample via something like (remember that reliability is the squared correlation between the true values and the observed values that have random measurement error):

```
. corr
(obs=1000)

             |       Yt       Xt        e        X
-------------+------------------------------------
          Yt |   1.0000
          Xt |   0.5010   1.0000
           e |  -0.0234  -0.0354   1.0000
           X |   0.3444   0.6954   0.6935   1.0000


. di "Reliability equals " .6954 ^ 2
Reliability equals .48358116
```

Or, if you want to use variance of X / variance of Xt, you get

```
. corr X Xt, cov
(obs=1000)

             |        X       Xt
-------------+------------------
           X |  1.90177
          Xt |  .953397  .988255


. di .988255 / 1.90177
.51965012
```

These two estimates differ a bit because we are using a sample rather than the entire population. In the population the approaches used above would all give the same numbers. (If in doubt, change the `drawnorm` command to `corr2data` and see how the results compare.)

The regression results differ because random measurement error in the independent variable produces a downward bias in the regression coefficient. Would a larger N help? No; look at the formula for the slope coefficient. N is not a factor:

$$\beta_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}$$

But if in doubt, try rerunning the problem with a sample of 100,000:

```
. clear all
. version 12
. set seed 123456789

. set obs 100000
obs was 0, now 100000
. drawnorm Yt Xt e, corr(1 .5 0 \ .5 1 0 \ 0 0 1)
. gen X = Xt + e
. reg  Yt Xt

      Source |       SS       df       MS              Number of obs =  100000
-------------+------------------------------           F(  1, 99998) =33039.93
       Model |  24845.6397     1  24845.6397           Prob > F      =  0.0000
    Residual |  75197.3254 99998  .751988294           R-squared     =  0.2483
-------------+------------------------------           Adj R-squared =  0.2483
       Total |  100042.965 99999  1.00043966           Root MSE      =  .86717

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |   .4993678   .0027473   181.77   0.000     .4939832    .5047524
       _cons |  -.0026293   .0027423    -0.96   0.338    -.0080041    .0027455
------------------------------------------------------------------------------

. reg  Yt X

      Source |       SS       df       MS              Number of obs =  100000
-------------+------------------------------           F(  1, 99998) =14257.19
       Model |  12483.7369     1  12483.7369           Prob > F      =  0.0000
    Residual |  87559.2282 99998  .875609794           R-squared     =  0.1248
-------------+------------------------------           Adj R-squared =  0.1248
       Total |  100042.965 99999  1.00043966           Root MSE      =  .93574

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .2503951    .002097   119.40   0.000     .2462849    .2545053
       _cons |  -.0036685   .0029591    -1.24   0.215    -.0094683    .0021312
------------------------------------------------------------------------------
```

The results differ slightly, but only because of sampling variability, not because a larger N reduces the problem caused by measurement error.

b. Briefly discuss the possible consequences of random measurement error in each of the following situations.

1. A researcher is interested in how Age affects feelings of Self-Efficacy. Age is believed to be very well-measured. Self-efficacy is measured on a scale that ranges from 0 to 100; because self-efficacy is a fairly abstract concept to most people, it is believed that this scale will suffer from at least some random measurement error.

Note that the DV, Self-efficacy, is the var thought to suffer from random measurement error. The bivariate correlation will be biased downward in magnitude (attenuated) because of this error. The slope coefficient is not biased when the DV has random measurement error; however, the standard errors will be larger, making our parameter estimates less precise and making it more difficult to determine if there is a significant relationship between the two variables. (Incidentally, a larger N would help in this case, because it would reduce the standard errors.)

2. A researcher has collected data from a sample of men and a sample of women. She believes that political attitudes will have less of an effect on the political activism of men than they do on the political activism of women. Political activism is known to be very well measured for both men and women. Political attitudes are measured by respondents' self-reports to a lengthy series of questions. During the interview process, the researcher notices that women tend to give careful thought to the questions before answering. Men, on the other hand, tend to rush through the questionnaire and finish quickly.

In this problem, the independent variable, political attitudes, may suffer from random measurement error. Random error in the IV biases correlations downward and also produces a downward bias in the estimated slope coefficient. An added problem here is that men may provide less reliable answers than the more careful women do. This could cause a greater downward bias in the estimated effects for men than it does for women. As a result, the researcher could appear to be right – the estimated effect of political activism is less for men than it is for women – when she really isn't right. The apparent differences between men and women could be an artifact of the superior measurement of women's attitudes.

Part II. Outliers/Heteroscedasticity.

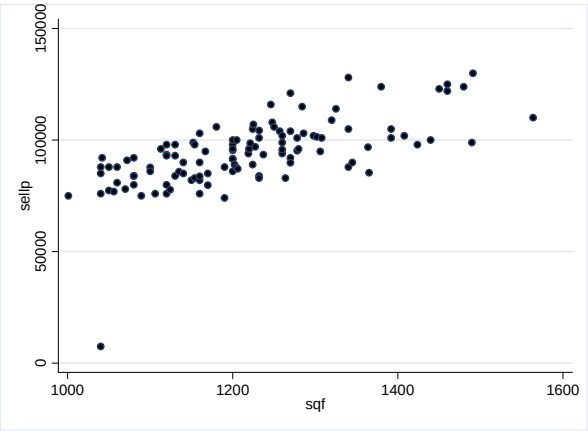
The problem on outliers and heteroscedasticity is selected from J.D.Jobson's book *Applied Multivariate Data Analysis*, pp. 169-172. This is a sample of 116 real estate sales transactions in a particular region of a large city. The variables include the dependent variable, selling price (SELLP) and the independent variable, number of square feet (SQF) of each transaction. You need to copy the file *resales.dta* from the course web page.

a. First, check the data for outliers. Your analysis should include the following. For each part, explain whether and how the analysis helps you to identify outliers.

1. A scatter plot of sellp and sqf

```
. scatter sellp sqf
```

Notice the one case off by itself in the lower left hand side:



2. An examination of the extreme values of sellp and sqf

```
. extremes sellp sqf

  +---------------------+
  | obs:   sellp    sqf |
  |---------------------|
  |   4.    7400   1040 |
  |  49.   74000   1190 |
  |   1.   75000   1001 |
  |  18.   75000   1089 |
  |   5.   76000   1040 |
  +---------------------+

  +----------------------+
  | 104.   124000   1380 |
  | 113.   124000   1480 |
  | 112.   125000   1460 |
  |  98.   128000   1340 |
  | 115.   130000   1491 |
  +----------------------+

note: 4 values of 76000

. extremes sqf sellp

  +---------------------+
  | obs:    sqf   sellp |
  |---------------------|
  |   1.   1001   75000 |
  |   2.   1040   85000 |
  |   3.   1040   88000 |
  |   4.   1040    7400 |
  |   5.   1040   76000 |
  +---------------------+

  +----------------------+
  | 112.   1460   125000 |
  | 113.   1480   124000 |
  | 114.   1490    99000 |
  | 115.   1491   130000 |
  | 116.   1564   110000 |
  +----------------------+

note: 2 values of 1460
```

Case 4 has a much lower value on sellp than any other case does.

3. The computation and examination of diagnostic statistics. At a minimum, these should include the standardized residuals and a leverage measure and the dfbetas. The predicted yhat value may also be useful.

```
. reg sellp sqf

      Source |       SS       df       MS              Number of obs =     116
-------------+------------------------------           F(  1,   114) =   87.26
       Model |  1.1063e+10     1  1.1063e+10           Prob > F      =  0.0000
    Residual |  1.4453e+10   114   126780027           R-squared     =  0.4336
-------------+------------------------------           Adj R-squared =  0.4286
       Total |  2.5516e+10   115   221874821           Root MSE      =   11260

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   83.30545   8.918024     9.34   0.000     65.63892     100.972
       _cons |  -7277.273      10905    -0.67   0.506    -28879.99    14325.45
------------------------------------------------------------------------------

. predict yhat
(option xb assumed; fitted values)
. predict rstandard, rstandard
. predict rstudent, rstudent
. predict leverage, leverage
. dfbeta
                      _dfbeta_1: dfbeta(sqf)
. predict cooksd, cooksd
. extremes rstudent sellp sqf yhat rstandard leverage _dfbeta_1 cooksd

  +-----------------------------------------------------------------------------------------+
  | obs:    rstudent   sellp    sqf       yhat   rstandard   leverage   _dfbeta_1     cooksd |
  |-----------------------------------------------------------------------------------------|
  |   4.   -8.124243    7400   1040    79360.4   -6.483429   .0283122    1.156535   .6123872 |
  | 103.   -1.911213   85400   1365   106434.7   -1.889357   .0223294   -.2263141   .0407646 |
  | 114.   -1.642972   99000   1490   116847.9   -1.630861   .0553149   -.3652734    .077868 |
  |  49.   -1.604063   74000   1190   91856.22   -1.593109   .0090839    .0346796   .0116331 |
  | 100.   -1.473081   88000   1340     104352    -1.46558   .0180848   -.1446201   .0197801 |
  +-----------------------------------------------------------------------------------------+

  +-----------------------------------------------------------------------------------------+
  |  90.    1.37313   115000   1284   99686.93   1.367828   .0114222   .0730979   .0108087 |
  | 104.   1.475247   124000   1380   107684.3   1.467694   .0252526   .1927036   .0279033 |
  |  74.   1.753662   116000   1246   96521.32   1.737914    .009142   .0402246   .0139334 |
  |  86.   2.034321   121000   1270   98520.66   2.006883   .0103714   .0855633   .0211047 |
  |  98.    2.15302   128000   1340     104352    2.11949   .0180848   .2113734   .0413687 |
  +-----------------------------------------------------------------------------------------+

. extremes _dfbeta_1 sellp sqf yhat rstandard leverage rstudent cooksd

  +-----------------------------------------------------------------------------------------+
  | obs:   _dfbeta_1    sellp    sqf       yhat   rstandard   leverage    rstudent     cooksd |
  |-----------------------------------------------------------------------------------------|
  | 114.   -.3652734    99000   1490   116847.9   -1.630861   .0553149   -1.642972    .077868 |
  | 116.   -.3473086   110000   1564   123012.5    -1.20755   .0840802   -1.210006   .0669293 |
  | 103.   -.2263141    85400   1365   106434.7   -1.889357   .0223294   -1.911213   .0407646 |
  | 109.   -.2073178   100000   1440   112682.6   -1.149461   .0397683   -1.151098   .0273602 |
  | 108.   -.2017698    98000   1424   111349.7   -1.207216   .0354558   -1.209666   .0267858 |
  +-----------------------------------------------------------------------------------------+
```

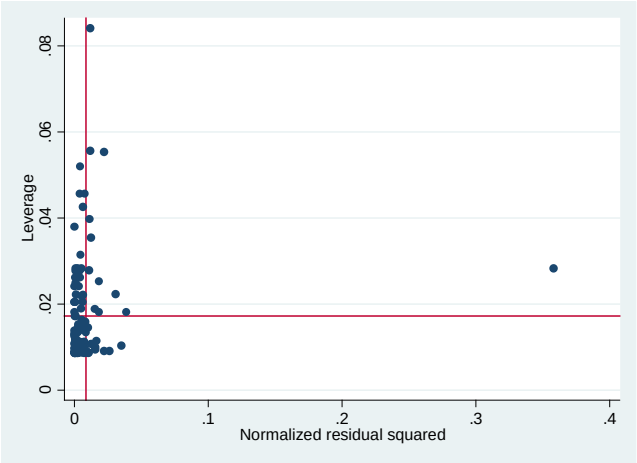
```
+-------------------------------------------------------------------------------------+
| 112.   .1905783   125000   1460   114348.7    .9683108   .0456105    .9680437   .0224047 |
| 104.   .1927036   124000   1380   107684.3    1.467694   .0252526    1.475247   .0279033 |
|  98.   .2113734   128000   1340     104352     2.11949   .0180848     2.15302   .0413687 |
| 115.   .2670693   130000   1491   116931.2    1.194392   .0556578    1.196653   .0420398 |
|***4.*** ***1.156535*** ***7400*** ***1040*** ***79360.4*** ***-6.483429*** ***.0283122*** ***-8.124243*** ***.6123872*** |
+-------------------------------------------------------------------------------------+
```

Case 4 has by far the largest standardized residual (much greater in magnitude than 3) and also has by far the largest dfbeta (which exceeds the rule of thumb of 1 or greater).

4. The leverage-versus-residual-squared plot.

You can clearly see one case (which we have already identified as belonging to case 4) has a much larger normalized squared residual. It also has slightly above average leverage.

```
. lvr2plot
```



5. Based on the above and any other analysis you do, indicate whether any of the cases appear to be outliers. [HINT: You have to be blind if you don't spot a problem right away.] If you find an outlier, discuss possible explanations for it. Coding error is always a possibility, but suggest other possible explanations as well.

Case 4, with a sellp value of 7,400, looks a tad suspicious. One obvious possibility is that somebody left off a zero. Also, notice that it's predicted value is 79,360, very close to 74,000. But, the value might be legitimate. Property values are determined by more than just square footage. This particular property might be in terrible shape or be located in a terrible area. Perhaps the seller gave the buyer a great deal for some reason, e.g. maybe it was sold to a relative or a charity.

b. Try up to three different strategies for dealing with the outlier:

1. Robust regression (rreg) [Optional]

```
. rreg  sellp sqf, nolog genwt(w)

Robust regression estimates                           Number of obs =     116
                                                      F(  1,   114) =   98.06
                                                      Prob > F      =  0.0000

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   74.26163   7.499139     9.90   0.000      59.4059    89.11737
       _cons |   4239.196   9169.981     0.46   0.645    -13926.46    22404.86
------------------------------------------------------------------------------

. extremes w sellp sqf

  +---------------------------------+
  | obs:           w    sellp    sqf |
  |---------------------------------|
  |   4.           0     7400   1040 |
  |  98.   .48557151   128000   1340 |
  |  86.   .54778108   121000   1270 |
  | 103.   .62259287    85400   1365 |
  |  74.   .65492325   116000   1246 |
  +---------------------------------+

  +---------------------------------+
  |  15.   .99980315    84000   1080 |
  |  17.   .99980315    84000   1080 |
  |  95.   .99984885   101000   1308 |
  |  42.   .99984886    90000   1160 |
  |  20.   .99999392    86000   1100 |
  +---------------------------------+
```

Robust regression (`rreg`) uses an iterative weighting scheme that causes outliers to be weighted less heavily in the calculations than are other cases. By using the `genwt` parameter, we can see the weights it generated. Case 4 received a weight of 0, meaning that it was basically dropped from the calculations, while some other outlying cases received weights considerably less than 1. Conversely, the cases with the largest weights had very small residuals.

2. Median regression (qreg) [Optional]

```
. qreg  sellp sqf, nolog

Median regression                                     Number of obs =       116
  Raw sum of deviations  1207500 (about 94000)
  Min sum of deviations 914369.4                      Pseudo R2     =    0.2428

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   78.82883   9.938526     7.93   0.000     59.14069    98.51697
       _cons |  -1135.135   12136.36    -0.09   0.926    -25177.18    22906.91
------------------------------------------------------------------------------
```

Median regression (the default for `qreg`) estimates the median of the dependent variable given the values of X. It doesn't toss out outliers, but outliers have less impact because the median is

less affected by outliers than the mean is. The coefficient for sqf is larger than in the other options (where case 4 is dropped) but not as large as it was in the original regression.

3. OLS regression, with the outlying case deleted.

```
. reg sellp sqf if sellp!=7400

      Source |       SS       df       MS              Number of obs =     115
-------------+------------------------------           F(  1,   113) =  109.07
       Model |  8.8061e+09     1  8.8061e+09           Prob > F      =  0.0000
    Residual |  9.1237e+09   113  80741091.2           R-squared     =  0.4911
-------------+------------------------------           Adj R-squared =  0.4866
       Total |  1.7930e+10   114   157279664           Root MSE      =  8985.6

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   75.07451   7.188646    10.44   0.000     60.83251    89.31652
       _cons |   3379.622   8800.875     0.38   0.702     -14056.5    20815.74
------------------------------------------------------------------------------
```

With this approach, you are just assuming the outlier should be excluded, perhaps because it is a coding error or because you feel it does not fall in the population of interest. In this particular situation, the results are similar to `rreg`, which also wound up dropping the same case. The results are not identical, partly because `rreg` also gives other cases weights that are less than one.

Briefly explain the rationale behind each approach and discuss any important differences in the results. Discuss any other strategies you might want to try, at least if you had the necessary information and resources to do so.

The first thing I would want to do is try to double-check the coding! If case 4 is supposed to be 74,000, then change the coding accordingly. If I can't check the coding, then I might prefer just to drop the case; or, run the analysis a couple of different ways, both with and without the outlier.

But, if the coding is legitimate, I'd like to try some additional variables if possible. As noted before, the property could be in poor condition or in a poor location.

If the code was legitimate, I might be tempted to prefer `qreg`. I think it makes more intuitive sense. Also, with real estate values, I could see where outliers might skew the distribution, perhaps making the median more substantively interesting and appropriate to examine than the mean. Luckily, all three of the above methods produce similar results in this case, so even if you choose the wrong approach the error won't be too costly.

c. For the remainder of this homework, DROP the outlying case. Then do the following tests for heteroscedasticity.
   1. A visual inspection of the plot of the residual versus fitted cases.

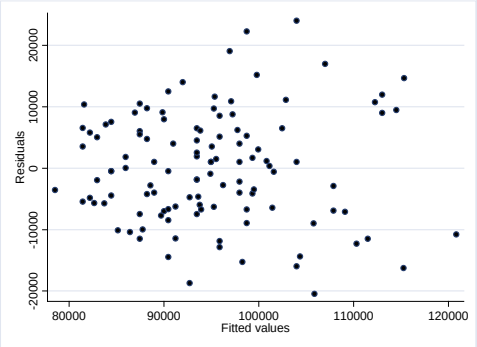
```
. drop in 4
(1 observation deleted)
```

```
. reg sellp sqf

      Source |       SS       df       MS              Number of obs =     115
-------------+------------------------------           F(  1,   113) =  109.07
       Model |  8.8061e+09     1  8.8061e+09           Prob > F      =  0.0000
    Residual |  9.1237e+09   113  80741091.2           R-squared     =  0.4911
-------------+------------------------------           Adj R-squared =  0.4866
       Total |  1.7930e+10   114   157279664           Root MSE      =  8985.6

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   75.07451   7.188646    10.44   0.000     60.83251    89.31652
       _cons |   3379.622   8800.875     0.38   0.702     -14056.5    20815.74
------------------------------------------------------------------------------

. rvfplot
```



Ideally, this would look more like a random scatter of points. You can see from this plot that the variance of the residuals does not appear to be constant. As the predicted Y increases, the magnitude of the residuals tends to increase (although eventually it seems to level off or even decline a bit). [NOTE: if there were more than 1 X, you would probably want to plot each of them against the residuals using the `rvpplot` command—or at least plot the X which you think might cause heteroscedasticity. In the bivariate case, it doesn't matter whether you use the predicted Y (i.e. the fitted values), or the observed X, because they are perfectly correlated with each other—predicted Y is computed from X and X only]

Of course, visual analyses can be deceiving, and sampling variability alone could produce the appearance of heteroscedasticity when it doesn't actually exist; hence we do the next two tests.

2. The Breusch-Pagan test and White's general test.

Breusch-Pagan tests whether the residuals are linearly related to the variables specified. In its default form (below) it tests whether, as the fitted values go up, the error variances also tend to go up (or, go down). In this particular case, since there is only one IV, the commands `hettest` and `hettest sqf` will produce identical results, but we could specify other IVs or even variables not in the model if we felt they were related to heteroscedasticity in the data.

```
. estat hettest

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

         chi2(1)      =    10.73
         Prob > chi2  =   0.0011
```

This test is consistent with our visual impressions. The significant chi-square indicates that, as the fitted values go up, the error variances also tend to go up (well actually, they could go down, but the visual inspection suggests otherwise.)

White's general test probably isn't necessary here, but it is useful when there may be non-linear relationships between the residuals and the variables specified.

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(2)      =     12.75
         Prob > chi2  =    0.0017

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      12.75      2    0.0017
            Skewness |       0.97      1    0.3253
            Kurtosis |       1.17      1    0.2793
---------------------+-----------------------------
               Total |      14.89      4    0.0049
---------------------------------------------------
```

Again, the test confirms that heteroskedasticity is present. The small increase in chi-square from Breusch-Pagan probably means we do not have to worry too much about non-linear relationships.

Based on your analyses (and any other analyses you choose to do) indicate whether heteroscedasticity appears to be a problem (and how the test supports your conclusion). If heteroscedasticity does appear to be a problem, explain why you think it occurs in this case.

All the tests indicate heteroscedasticity. Substantively, we might suspect heteroscedasticity because, with bigger and more expensive lots, there may be more variability (in absolute terms) in the range of reasonable prices. For example, a small lot might sell for between $60 and $80 thousand (a $20,000 range), a bigger lot might reasonably sell for between $150,000 and $200,000 (a $50,000 range). Also, larger square footage may be a necessary but not sufficient condition for the addition of features that disproportionately increase a house's value, e.g. a larger house may just be larger, but it might also include a swimming pool, better construction, nicer rooms.

d. Try up to three different strategies for dealing with the heteroscedasticity

1. Regular OLS regression (i.e. do nothing about the heteroscedasticity)

```
. reg sellp  sqf

      Source |       SS       df       MS              Number of obs =     115
-------------+------------------------------           F(  1,   113) =  109.07
       Model |  8.8061e+09     1  8.8061e+09           Prob > F      =  0.0000
    Residual |  9.1237e+09   113  80741091.2           R-squared     =  0.4911
-------------+------------------------------           Adj R-squared =  0.4866
       Total |  1.7930e+10   114   157279664           Root MSE      =  8985.6

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   75.07451   7.188646    10.44   0.000     60.83251    89.31652
       _cons |   3379.622   8800.875     0.38   0.702     -14056.5    20815.74
------------------------------------------------------------------------------
```

2. Regression with Robust Standard Errors

```
. reg sellp  sqf, robust

Regression with robust standard errors                Number of obs =     115
                                                      F(  1,   113) =   85.66
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4911
                                                      Root MSE      =  8985.6

------------------------------------------------------------------------------
             |               Robust
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   75.07451   8.111358     9.26   0.000     59.00445    91.14458
       _cons |   3379.622   9595.886     0.35   0.725    -15631.56     22390.8
------------------------------------------------------------------------------
```

When heteroskedasticity is present, OLS coefficient estimates are unbiased but the estimates of the standard errors are biased. The use of robust standard errors does not change the coefficient estimates. But, the assumption that errors are independent and identically distributed is dropped, causing the standard errors and hence the T values and confidence intervals to change. As a result, in this particular case the effect of sqf is a little less significant and its confidence interval is a little larger, but our main conclusions do not change much.

3. Weighted Least Squares [Optional]

```
. reg sellp  sqf [aw = 1/sqf^2]
(sum of wgt is   7.9476e-05)

      Source |       SS       df       MS              Number of obs =     115
-------------+------------------------------           F(  1,   113) =  107.50
       Model |  8.0127e+09     1  8.0127e+09           Prob > F      =  0.0000
    Residual |  8.4231e+09   113  74540296.6           R-squared     =  0.4875
-------------+------------------------------           Adj R-squared =  0.4830
       Total |  1.6436e+10   114   144173552           Root MSE      =  8633.7

------------------------------------------------------------------------------
       sellp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sqf |   75.44355   7.276581    10.37   0.000     61.02733    89.85976
       _cons |   2931.524   8753.001     0.33   0.738    -14409.75     20272.8
------------------------------------------------------------------------------
```

We have previously determined that as sqf goes up, the error variances tend to go up. With WLS, estimates which are less precise (i.e. have larger error variances) are weighted less heavily than estimates which are more precise (have smaller error variances). The result (assuming we do the weighting correctly) is parameter estimates that are more efficient and that have unbiased standard errors, i.e. they will vary less from sample to sample than regular OLS regression or regression with robust errors would.

For 2 (and 3 if you use it), briefly explain the rationale for each method. Indicate whether methods 2 and 3 change the conclusions you would reach using OLS regression without any attempt to deal with heteroscedasticity. [HINT: The differences between the three methods are not too dramatic in this case.]

Alas, in this case, even though the various tests all indicated heteroscedasticity, the harms it caused were obviously slight. Both robust standard errors and weighted least squares led us to almost the exact same conclusion as did regular OLS. This is consistent with Allison's claim that, unless heteroskedasticity is marked, it tends not to have much of an effect.

Of course, it could be that we have not approached this problem in the optimal fashion either. There are undoubtedly important variables omitted from the model. It may be that we should transform the variables in some way, e.g. use the log of selling price rather than selling price. If we had more information or had more of a theory about how square footage is related to selling price, we might be able to come up with a better solution than we have here.