Soc 63993, Homework #3:
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.

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.

        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.

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
       2.   An examination of the extreme values of sellp and sqf
       3.   The computation and examination of diagnostic statistics. At a minimum, these should include the standardized residuals and a leverage measure the dfbetas. The predicted yhat value may also be useful.
       4.   The leverage-versus-residual-squared plot.
       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.

b.   Try up to three different strategies for dealing with the outlier:
       1.   Robust regression (rreg) [Optional]
       2.   Median regression (qreg) [Optional]
       3.   OLS regression, with the outlying case deleted.

Briefly explain the rationale behind each approach you try 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.

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

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.

d.   Try up to three different strategies for dealing with the heteroscedasticity
       1.   Regular OLS regression (i.e. do nothing about the heteroscedasticity)
       2.   Regression with Robust Standard Errors
       3.   Weighted Least Squares [Optional]

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.]