Problem Set 1 – OLS and its issues
Due 2/1/18

General instructions:

- Feel free to work together, but work should be submitted individually and code should not be copy and pasted from classmates.
- Please present your answers in a clear, concise fashion. Submit a single main PDF document with your answers. In your solution packet, include relevant Stata (or other language) output (e.g., key regression output, key graphs, etc.) and well-annotated Stata .do commands. (Do NOT include pages and pages of "undigested" Stata log files in the main problem set answers.) Make clear reference to regression output and figures in your written answers.
- Please upload the PDF file to Canvas. In addition to the main answers, please also upload .do and .log files as appendices

Problem 1 – Using Monte Carlo to look at Robustness vs Efficiency

It's best to compile your numerical answers to parts 1a-1c into a nicely formatted table.

A. Use a MC analysis to determine whether the usual OLS (and usual variance estimate) is accurate in the case of heteroscedasticity. 1,000 repetitions is sufficient.

   Data generating process (DGP): Assume the true data generating process is the following: x is distributed iid N(0,1). epsilon_i is distributed N(0,1+exp(x_i)). That is, the variance of epsilon is (1+exp(x)) for that observation. Then assume that the LHS variable Y is generated according to: y = 2 + 1 * x + epsilon . Make the number of observations (N) in the dataset be 50.

   Estimation method: Use Stata's "regress" command.

   Results to look for:

   First, check whether the estimated slope on x is unbiased. To do this, first eyeball whether the average of the estimated slopes is "close to" the truth. To test this properly, perform a statistical test of whether the average of the estimated slopes equals the truth.
   Second, see if this estimation method has decent inference properties in this model. To check this, collect a t-statistic from a test of the (true) null hypothesis for each MC replication. Then, look at the percentiles of the t-stats: do they look like they are standard normally distributed? (hint, use "summarize *varname*, detail" to get some percentiles. Or use Stata's "pctile" and "_pctile" functions to get the 2.5% and 97.5%). Note that if

N (the number of observations in each data set) is small, then the distribution will look like a T-distribution and not a standard Normal. To perform a proper "size" test, check how often the true null would be rejected. If you reject the null when the t-stat is greater than 1.96 in absolute value, you hope that you will reject 5% of the time. Is this the case?

B. Use a MC analysis to determine whether the Huber-White robust standard errors give accurate inference in the model in (1a). Stata's "regress y x, robust" command computes the Huber-White robust standard errors.

C. How to the results in (1a) (1b) change as N = 5, 10, and 200 ?

D. Go back to the d.g.p. from problem (1a) and see how the properties of the FGLS estimator goes. Figure out how to estimate FGLS in Stata, and describe your method (both in "econometrics" language and in "stata" language). In order to implement FGLS you'll want a statistical model for the heteroscedasiticy. Use this model:

var(eps_i) = gamma_0 + gamma_1 * exp(x_i).

Then show how (1) the estimated slope parameter and (2) statistical inference perform with FGLS. Is the new slope estimated more efficiently? How can you tell?

E. Now change the d.g.p. further. Assume that there is some other type of heteroscedasticity. In particular, suppose epsilon_i N(0 , (1+3*abs(x_i)). Now, compare 5 estimators:

i. OLS with naïve (assuming iid) std. errors
ii. OLS with White-Robust std. errors.
iii. FGLS, assuming the **incorrect** heteroscedastiticy described in part 1a.
iv. FGLS, assuming an FGLS model based on the **correct** heteroscedastiticy described in part 1e.
v. FGLS, assuming **incorrect** model as in (C), but then doing White-Robust inference after.

For each of these estimators, discuss the results in terms of **bias** and **efficiency** of the "Beta" (estimated slope), and the "**correct inference**" properties in terms of average estimated standard error, and the likelihood of rejecting a true null hypothesis.

Problem 2 Monte Carlo, Angrist & Pishke's OLS bias formulas.

Read section 3.2 of Mostly Harmless Econometrics

A. In AP's "Mostly Harmless Econometrics", they discuss Omitted Variables Bias as one of the pitfall of OLS estimation. Equation (3.2.11) on page 60 gives a formula for the bias of a bivariate regression when key variables are omitted. Construct a Monte Carlo analysis to demonstrate the omitted variables bias. The data will be fake, but choose an application relevant to your research interests. That is, do not have variables named 'y' and 'x', instead have 'firm' and 'corporate structure' and create a DGP with logical assumptions relative to your example. Show the results qualitatively, but also check the quantitative magnitude as presented in equation (3.2.11).

B. Angrist and Pishke also discuss (pg 67, equations 3.2.13 - 3.2.15) the bias that can result from "over conditioning" or "bad control": conditioning on an outcome (endogenous) variable. Construct a Monte Carlo analysis to demonstrate the "conditioning on an outcome variable" bias. Show the results qualitatively, but also check the quantitative magnitude as presented on page 67 and the equations there.

Problem 3 – Experimental data – Density Plots

For this problem, download the stata dataset "kenya.dta". The data are from a project that Michael Kremer, Ted Miguel, and Rebecca Thornton did in rural Kenya. The latest version of the paper is at:
http://www.mitpressjournals.org/doi/abs/10.1162/rest.91.3.437

This problem examines the relationship between merit awards and academic performance, as measured by school exams, among Kenyan schoolchildren. In early 2001, Grade 6 girls in a random subset of "treatment" schools (variable name "treat") were offered a large cash award if they scored in the top 15% of all treatment school girls. These girls are cohort 1 (indicator variable "c1"). There is also a cohort 2, who were grade 6 girls who faced the incentive in 2002. The winners in each cohort are given by the binary variables ("winn01" "winn02")

The dataset contains test score information from late 2000, the year before the program, late 2001, the first year of the program, and 2002, the second year of the program. The test score outcomes ("test00", "test01" "test02") were normalized such that the test distribution in the comparison schools is mean zero with a standard deviation of one (for all students in that grade, not just those in this sample – thus the mean need not equal zero in the sample).

The goal of this exercise is to estimate overall program treatment effects using various parametric and non-parametric methods. Another important issue for the analysis is whether girls at the bottom of the baseline test score distribution were harmed by the program – perhaps due to demoralization or diversion of teacher attention to high-achieving classmates.

There is also some long-term (2005-2006) follow-up data. in_school measures whether the girl is still in school. Educ_attain is educational attainment upon follow-up. Vocab and Math are normalized test scores.

A. Present summary statistics for the key variables in the dataset. Do baseline 2000 test scores differ on average across the treatment and comparison students? Do 2001 test scores differ on average across treatment groups? Do 2002 test scored differ on average? Do the 2005-2006 test scored differ? [HINT: Use the STATA commands **summarize, detail** and **ttest** (with the "by" option)]

B. Estimate treatment impact in regression way
Add controls for test00
Regress difference on treatment
Which of these is preferred? Why?

C. Present a scatterplot graph of test01 against test00. Split the symbols in your graph, so that we can see who the “treated” and “control” observations are. (To do this, learn about stata’s “msymbol” graphing option. You may need to have two graph commands in the same command line to get two different symbols.)

Kernel Density Estimation

D. Plot the kernel density of 2000 test scores in the following ways:
   (i) Epanechnikov kernel with the "optimal" bandwidth
   (ii) Epanechnikov kernel with bandwidth equal to 0.05
   (iii) Gaussian kernel with optimal bandwidth

Also, experiment with many different bandwidths. Try to find the one that's the smallest, but still "does justice" to the data. Try to find the largest one that you think still gives a good representation of the data.

Characterize the distribution of 2000 test scores. In this application, does kernel density estimation appear to be more sensitive to the bandwidth or to the kernel? [HINT: Use **kdensity** and **graph twoway**.]

E. Plot the kernel density of 2000 test scores with Epanechnikov kernel and optimal bandwidth separately for treatment and comparison students, and place them in one figure. Do the same for 2001 scores, and for 2002 scores. Describe any shifts in the distributions through time.

Problem 4 – Measurement Error in Stata

Let's figure out whether measurement error in RHS variables can make coefficients go **up**. We can do this using "proof by Stata" with a single "large N" sample. Later we can embed this in a monte carlo.

A. Revisit the classical errors-in-variables problem. Let's assume a single regressor, x_true, which is distributed as a standard normal variable. But we see instead x_observed, which has a random (N(0,1)) measurement error. Note here that the variance of the truth and the variance of the measurement error are the same, giving a "reliability ratio" of 0.50. Generate y = 1 + 1 * x_true + epsilon. (For this, use a N(0,1) distribution for epsilon.) Now do a regression on the "unobtainable" x_true, to confirm that things work as they should. Next, regress on x_observed. Is there attenuation bias?

B. In addition to attenuation bias, we should see "variability of beta hat" go up, and "accuracy of estimated standard errors" go down. To check these, build an MC exercise.

C. Next, let's add in an additional regressor. Let's assume that we measure x2 without error. Further, experiment with whether x1_true and x2_true are correlated or not, and with their relative variances. y = 1 + 1 * x_true + 1* x2_true + epsilon. See what happens to beta1 and beta2 when you regress on x1_observed and x2_true. Can you find a specification where one of the betas goes up? Can you find a specification where none of the betas go up?