Mega HW #4
ECNS 561 (Fall 2017)
Due: 12/7/2017

<u>1.)</u> Work the following Chapter 4 problems from Wooldridge: 6, 8, 9, 11, 12

<u>2.)</u> Work the following Chapter 5 problems from Wooldridge: 1, 2, 4 (Note: If we do not get through Ch. 5, then you will not be responsible for these problems)

<u>3.)</u> Consider the following regression results from a random sample of 220 home sales

$$\widehat{Price} = \underset{(23.9)}{119.2} + \underset{(2.61)}{0.485}BDR + \underset{(8.94)}{23.4}Bath + \underset{(0.011)}{0.156}Hsize + \underset{(0.00048)}{0.002}Lsize + \underset{(0.311)}{0.090}Age - \underset{(10.5)}{48.8}Poor$$

where *Price* denotes the selling price (in \$1000), *BDR* denotes the number of bedrooms, *Bath* denotes the number of bathrooms, *Hsize* denotes the size of the house (in square feet), *Lsize* denotes the lot size (in square feet), *Age* denotes the age of the house (in years), and *Poor* denotes a binary variable that is equal to 1 if the condition of the house is reported as "poor."

a.) Is the coefficient on *BDR* statistically significantly different from zero?

b.) Typically five-bedroom houses sell for much more than two-bedroom houses. Is this consistent with your answer to a.) and with the regression more generally?

c.) A homeowner purchases 2000 square feet from an adjacent lot. Construct a 99% confidence interval for the change in the value of her house.

d.) Lot size is measured in square feet. Do you think that another scale might be more appropriate? Why or why not?

e.) The *F*-statistics for omitting *BDR* and *Age* from the regression is $F = 0.08$. Are the coefficients on *BDR* and *Age* statistically different from zero at the 10% level?

<u>4.)</u> Let e_i be the *i*th residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that $\text{plim}(e_i - \varepsilon_i) = 0$.

STATA Exercise

You are required to complete this problem **entirely in MATA** and turn in your final do file. Use the same data set we used in Mega HW #3, “KS Crime Data Set for Mega HW #3.”

For what follows, let’s only use data for 2011.

a.) Suppose we are interested in estimating the following equation

$$Property_c = \beta_0 + \beta_1 Unemployment_c + \beta_2 Pop\_Density_c + \beta_3 Democrat\_GOP_c + \varepsilon_c$$

where $Property_c$ is the property crime rate in county *c*, $Unemployment_c$ is the unemployment rate in county *c*, $Pop\_Density_c$ is the population density in county *c*, and $Democrat\_GOP_c$ is the ratio of Democratic to GOP votes in county *c*.

Solve for the OLS estimators, the t-stats, and p-values. Which coefficient estimates are statistically distinguishable from zero?

b.) Suppose we are interested in testing the joint significance of *Unemployment* and *Pop_Density*. State the null and alternative hypotheses that we are interested in testing. Calculate the F-stat in Mata. Do we reject or fail to reject the null?

Paper Replication
For this replication exercise, you will need to read the following paper,

Anderson, D. Mark. 2013. "The Impact of HIV Education on Behavior Among Youths: A Propensity Score Matching Approach." *Contemporary Economic Policy* 31(3): 503-527.

The dataset for the replication has been posted on the class webpage under the link "YRBS 2009 Data Set for Mega HW #4." Do not try to complete this exercise in Mata. Use the "canned" regression command in STATA when running OLS.

a.) Replicate the summary statistics and OLS results for the males in the sample. That is, I want you to replicate the first two columns in Table 2, the first two columns in Table 3, and columns (1), (2), (4), and (5) in Table 4. You do not need to replicate the results in Table 4 that include primary sampling unit fixed effects (i.e., columns (3) and (6)).

You are required to turn a do file that I should be able to run that replicates all of these results. I want to be able to load the data set, run the do file, and see all of the results requested. In addition, I want you to turn in two tables. The first table should be the results from your summary statistics replication. I want these condensed into one table. The second table should be your results from the OLS replication. These tables need to look professional and contain enough detail such that they should be able to "stand alone." Take this part of the exercise seriously as I will dock points for sloppiness.

In order to replicate the OLS results, you will need to incorporate three concepts into your regressions that we have not yet covered. First, you need weight your regressions with the sampling weights provided with the YRBS data. To do so, you will have to invoke the "pweight" option (consult the online STATA help manual to figure out how to include this in your regression specification). For a description of the sampling weights, refer to page 49 of the YRBS user guide for 2009 (ftp://ftp.cdc.gov/pub/data/yrbs/2009/YRBS_2009_National_User_Guide.pdf).

Second, the standard errors reported are heteroskedasticity-robust standard errors. To implement this in STATA you need to invoke the "robust" option (again, you should be able to consult the STATA manual online to quickly figure this out). Also, please read the first two sections of Chapter 8 in Wooldridge for reference on this particular topic.

Lastly, in the cases where the dependent variable is binary, OLS is referred to as a "linear probability model." Please consult section 7.5 in Wooldridge for a relevant discussion.

The answers to the remaining questions need to be typed up, not hand written. Do not hand in answers that are sloppily written either...proof read, articulate you answers carefully, and don't even think about making spelling mistakes!

b.) Consider the t-tests for the means that you replicated for Tables 2 and 3 in Anderson (2013). What is the point of conducting these tests? More specifically, what information do we get from these tests when trying to think about a causal relationship between HIV education and teen risky behavior? Be precise in your answer and carefully think about whether or not we should consider HIV education as exogenous.

c.) Consider the controls listed in Tables 2 and 3 for hours of TV watched and hours of sleep. Why are these included in the model? What is the rationale? In general, what is the tradeoff the researcher faces when potentially including "too many" control variables?

d.) Consider the regression results you replicated for Table 4 for the full male sample in Anderson (2013) (i.e., columns (1) and (2)). What happens to the coefficient estimate on HIV_ED for these three regressions when the "additional individual controls and proxies for family/school environment" are included in the model? Do you prefer the models with our without these additional controls? Explain.

e.) Interpret the coefficient estimates for the "ever had sex", "had sex within last 3 months", and "needle use" regressions for the full male sample (make sure to read the discussion in Wooldridge on how to interpret coefficient estimates from linear probability models). These are the estimates that correspond to columns (1) and (2) in Table 4 of Anderson (2013). Are the magnitudes of the coefficient estimates meaningful in size? Explain.

f.) Recall our discussion in class that we had on "falsification tests" in regression analysis. If you could collect any data you wanted on the individuals in our sample, what types of variables might make for ideal falsification tests. Justify your answer. A couple of examples will do fine.