The University of Texas at Austin
ECO 348K (Advanced Econometrics)
Prof. Jason Abrevaya
Fall 2016

PROBLEM SET #1 (due Thursday, September 15th, 8am)

1. The latest election poll, based upon a sample of 5,000 randomly sampled individuals (representative of the voting population), indicates a lead for Hillary Clinton (53.4%) over Donald Trump (46.6%).

   (a) What "true (population) parameter" does the differential 0.534-0.466=0.068 estimate?

   (b) If the (asymptotic) standard error of the 0.534 Clinton vote share is 0.007, what is a 95% confidence interval for the population parameter in part (a)?

   (c) If the (asymptotic) standard error of the 0.534 Clinton vote share is 0.007, what is a 95% confidence interval for the number of votes that she would receive out of a group of 10,000 voters?

2. Your classmate is writing their honors thesis on the effects of computer ownership on academic performance. He has collected extensive data from college students. His most basic regression specification involves a regression of **GPA** (college GPA) on **compown** (laptop-computer ownership indicator variable), **SAT** (high-school SAT score), and **female** (female indicator variable). He finds a positive and statistically significant effect of **compown** on **GPA**, with an estimated slope of 0.24 and standard error of 0.03.

   (a) Explain why we should be cautious about interpreting the 0.24 estimate as a causal effect. What might be an omitted variable here? Thinking about this omitted variable, what direction do you think the omitted-variables bias goes in this example? (i.e., is the 0.24 estimate too high or too low?)

   (b) If **SAT** is dropped from his regression model, what do you think would happen to the **compown** slope estimate? Would it go up or down?

   (c) The variable **female** is arguably randomly assigned (at birth), so your classmate is quite confident that the slope estimate on the **female** variable represents a causal effect of gender on college GPA. What do you think?

   (d) If the estimated **female** slope is 0.10 with a standard error of 0.05, which of the following tests would have a higher p-value: the test of $H_0 : \beta_{compown} = 0$ or the test of $H_0 : \beta_{female} = 0$? (You don't need to compute the p-values; just say which one is higher.)

3. Use the **fertil2.dta** dataset (one of the Wooldridge datasets on Canvas). This dataset contains 4,361 observations from Botswana's 1988 Demographic and Health Survey.

We will focus on the outcome variable **y=children** (number of living children that a woman has). Each individual in the dataset is a woman aged 15 through 49. The variable **heduc** (husband's education) is missing for unmarried women (those with **evermarr**=0) but is also missing for some married women (those with **evermarr**=1). Generate a missing-variable indicator by doing the following: **gen heducmissing = (heduc==.)**, which will be equal to 1 if **heduc** is missing and 0 otherwise. Now replace missing values of **heduc** with zeros as follows: **replace heduc=0 if heducmissing**.

(a) Run the regression of **children** on **age** (woman's age), **age** squared, **educ** (woman's education), **evermarr**, **heduc**, and **heducmissing.**

  i. What is the estimated partial effect of **age**?

  ii. What is the estimated partial effect of **evermarr**?

(b) You are concerned that married women and unmarried women are very different and should have totally separate models. Run two separate regressions of **children** on **age**, **age** squared, **educ**, **heduc**, and **heducmissing**, one for the subsample with **evermarr**=1 and one for the subsample with **evermarr**=0. Are the effects of age and education different in the two regressions? Why does Stata drop variables in the second regression?

(c) Run the regression of **children** on **age**, **age** squared, **educ**, and **electricity** (1 if home has electricity, 0 otherwise). What is the estimated partial effect of **electricity**? Should we think of this estimated partial effect as the true causal effect of electricity availability on **children**?

(d) Interact **electricity** with all of the other (three) variables in the model, and rerun the regression. What is the estimated partial effect of **electricity**? Draw a histogram of the estimated in-sample partial effects for each of the observations in the dataset. (You'll need to somehow construct the estimated partial effect for each observation and then use the **histogram** command.) What is the *average partial effect* (APE) of **electricity**?

(e) Instead of doing interactions, consider running separate regressions of **children** on **age**, **age** squared, and **educ** for the **electricity**=0 and **electricity**=1 subsamples. How do the estimates in these subsample regressions compare to those in the model with interactions in part (d)?

4. Use the **CARD.DTA** dataset (one of the Wooldridge datasets on Canvas) that we previously used in class. This dataset contains 3,010 observations.

(a) Drop the last 10 observations by doing the following: **drop if _n>3000**

(b) Run the regression (with robust standard errors) of **lwage** on **educ**, **exper**, and **expersq**. Provide a 90% confidence interval for the partial effect of **exper**, evaluated at an experience level of 10 years. (Hint: Use a partial derivative to determine the right expression and then use **lincom**.)

(c) Now separately run the same regression on three different subsets of the data: observations 1-1000, observations 1001-2000, and observations 2001-3000.

i. Focusing on the **educ** slope estimates, would there be a reason to prefer any of three estimates that you have obtained?

ii. Focusing on the (robust) standard error for the **educ** slope estimates, how do the three standard errors compare to each other? how do they compare to the standard error from the regression in part (b)?

iii. Now consider an estimator of the **educ** slope that is obtained by taking the average of the three **educ** slope estimators that you have obtained. What is the standard error for this new estimator? (Hint: The three estimators are independent of each other since you have i.i.d. observations and, therefore, the three subsamples are independent of each other.) How does the standard error compare to the one for the estimator in part (b)?

(d) Returning to the original regression in part (b), consider the **R-squared** and **Root MSE** numbers from the Stata output.

i. What are the underlying population parameters being estimated by these two quantities (estimators)?

ii. Note that Stata does not directly provide standard errors for these two estimators. Instead, we can use the bootstrap for inference. Use the bootstrap in order to provide two different 95% confidence intervals for both of the parameters from (a): (1) a symmetric confidence interval based upon the bootstrap standard error and (2) a percentile interval.