Regression II
Vanderbilt University
Prof. Sean P. Corcoran

73 total points
Due August 28, 2025

Problem Set 1

Instructions: Answer the following questions in a Stata do-file, and submit your resulting log file via email to `sean.corcoran@vanderbilt.edu`, preferably as a PDF. Use your last name and problem set number as the filename (e.g., *Smith PS1.pdf*). The resulting log should include the questions below (commented), your commands, output, and written responses/interpretations. Graphical output (where applicable) can be submitted separately, or combined with your log into one PDF file. Working together is encouraged, but all submitted work should be that of the individual student.

---

1. Use the Stata syntax below to create a dataset of potential outcomes (Y_0, Y_1) for 600 students. The data include four "types" of students, indicated by the X variable. The indicator variable $D = 1$ if the student participated in an educational intervention (and $D = 0$ otherwise). **(28 points–4 each)**

```
clear all
set seed 3791
set obs 100
   gen x  = 1
   gen y0 = 25
   gen y1 = 35
   gen d  = runiform()<=0.20
set obs 250
   replace x = 2 if d==.
   replace y0= 50 if d==.
   replace y1= 90 if d==.
   replace d = runiform()<=0.80 if d==.
set obs 450
   replace x = 3 if d==.
   replace y0= 40 if d==.
   replace y1= 60 if d==.
   replace d = runiform()<=0.50 if d==.
set obs 600
   replace x = 4 if d==.
   replace y0= 30 if d==.
   replace y1= 45 if d==.
   replace d = runiform()<=0.40 if d==.
```

   (a) Use this dataset to calculate the ATE, ATT, and ATU (show your syntax). How do they compare? Show that the ATE is a weighted average of ATT and ATU.

   (b) How would you describe (in words) the four student "types" in this dataset, in terms of their potential outcomes, treatment effects, and propensity to be

treated? Use the switching equation to create the *observed* Y in your dataset. Speculate on the direction of selection bias and heterogeneous treatment effect bias (if any) if you were to use the simple difference in observed means ($Avg(Y|D = 1) - Avg(Y|D = 0)$) to estimate the ATE.

(c) What is the simple difference in mean *observed* outcomes between the treated and untreated cases? Given what you know about potential outcomes for these students, calculate the selection bias and heterogeneous treatment effect bias.

(d) As an alternative to the naïve estimator in (c), calculate the difference in mean *observed* outcomes separately for each student type. Then, take the simple average of the these four differences. How does it compare to your answer in (c)? To the (known) ATE? ATT? Why is this better (or is it) than the mean in (c)?

(e) As another alternative, calculate the *weighted* average of the four differences found in part (d), using the number of students of each type as weights. How does it compare to your answer in parts (c)-(d)? To the (known) ATE? ATT? Why is this better (or is it) than the means in (c)-(d)?

(f) Estimate an OLS regression of Y on D and include dummy variable indicators for student type (use Stata factor variables, and exclude the first type). What is your estimated coefficient on D? How does it compare to the (known) ATE? ATT? ATU? To your earlier treatment effect estimates?

(g) Suppose D were randomly assigned to the students in this dataset. Will this guarantee that the simple difference in means equals the ATE? Why or why not?

2. Suppose you conduct a randomized controlled trial in which 50% of your study population is assigned to the treatment condition and 50% is untreated. Unfortunately, 1/3 of your treated subjects fail to comply and do not actually receive the treatment. Explain (using potential outcomes terminology) why the ATE cannot be estimated in this case. **(5 points)**

3. For the following questions use the Stata dataset on Github called *LUSD4_5.dta*. This dataset consists of 47,161 observations of 4th and 5th graders from a large urban school district ("LUSD") in 2005 and 2006. For now, keep only 5th grade observations from 2005. NOTE: I also recommend keeping only observations that have nonmissing *mathz, totexp* and *econdis*. (**35 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta, clear
```

(a) You are interested in the causal effect of having a more experienced teacher (where experience is measured in years). Apply the concept of potential outcomes and counterfactuals to explain the causal effect you care about. (**4 points**)

(b) Estimate a simple regression relating student z-scores in math (*mathz*) to their teachers' years of experience (*totexp*). Interpret the slope and intercept in words. Is the coefficient for teacher experience statistically significant? Is the estimated coefficient *practically* significant? (Hint: consider a one standard deviation change in the explanatory variable). Explain. (**5 points**)

(c) Do you think the coefficient being estimated in part (b) represents either an ATE or ATT for a population of interest? Why or why not? (**4 points**)

(d) Your co-author is concerned that the regression in part (b) does not have a causal interpretation. Specifically, she thinks that experienced teachers are less likely to work with low-income students, who (for other reasons) tend to perform worse on tests on average. What does this say about the likely direction of omitted variables bias? Explain, using the concepts of potential outcomes and the OVB formula. (**4 points**)

(e) Using these variables (*mathz, totexp,* and *econdis*, an indicator variable for economically disadvantaged students), demonstrate the omitted variables bias formula shown in class ($\beta_s = \beta_\ell + \pi_1\gamma$), where the parameters are as defined in the lecture notes. Are these results consistent with your answer in part (d)? Provide an interpretation of the auxiliary regression coefficient π_1. (**5 points**)

(f) Another formula that is useful in regression is called "regression anatomy," below. It looks similar to—but is not the same as—the OVB formula. In this expression, β_1 is the coefficient on teacher experience from the "long" regression on teacher experience and *econdis*. $\tilde{X}_{1i}$ is the estimated residual after regressing teacher experience on *econdis*. $C()$ is covariance and $V()$ is variance. Show that this formula holds in your data. (Hint: you can easily get the covariance using `corr`).

$$\beta_1 = \frac{C(Y_i, \tilde{X}_{1i})}{V(\tilde{X}_{1i})}$$

This formula has a simple interpretation: the multivariate regression coefficient on X_1 (here, teacher experience) can be written as the *simple* regression coefficient from a regression of Y on $\tilde{X}_{1i}$, teacher experience that has been "purged" of all correlation with the other explanatory variables in the model. (**5 points**)

(g) Your co-author remains unsatisfied with the regression specification in (e) and recommends you also control for *mathz_1*, the student's math score in the prior grade, and *lep* (Limited English Proficient). Estimate the multivariate regression with *totexp, econdis, mathz_1*, and *lep*. Provide an interpretation, in words, of the four regression coefficients. How did the two regression coefficients on *totexp* and *econdis* change from the case in which these were the only two explanatory

variables? What happened to their standard errors? Provide some intuition behind both changes. (**4 points**)

(h) Add an interaction term to the regression in part (g), between *lep* and *totexp*. Interpret the estimated coefficient on the interaction. (**4 points**)

4. A researcher estimates a bivariate regression of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ but confides to a colleague that she believes this regression model suffers from omitted variables bias. The colleague suggests that the researcher construct $\hat{\epsilon_i} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ and then run a regression of $\hat{\epsilon_i}$ on x_i–that is, a regression of the form $\hat{\epsilon_i} = \gamma_0 + \gamma_1 x_i + \nu_i$–and then test the null $H_0 : \gamma_1 = 0$ to see if ϵ_i and x_i are correlated. Is this a good idea, or not? Explain. (**5 points**)