Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Problem Set 1
67 total points
Due September 8, 2022

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., *Fauci PS1.pdf*). The resulting log should include the questions below (commented), your commands, output, and written responses/interpretations. Graphical output 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. 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. Assume these observations are random draws from the population. (**48 points**)

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

   (a) 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. (**7 points**)

   (b) Applying the terminology used in class, is part (a) estimating a *population regression function*? Is it estimating a *conditional expectation function* (CEF)? Is it estimating a *causal* "ceteris paribus" relationship in the population? Defend your answers, *using potential outcomes terminology*. (**5 points**)

   (c) Install the user-written .ado file called `binscatter`:

```
ssc install binscatter
```

   Use this command to produce a binned scatter plot showing the relationship between math z-scores on the vertical axis and teacher experience on the horizontal axis. Bearing in mind this is sample data, do your findings suggest that the population CEF is linear? Provide an intuitive explanation for why the CEF might not be linear. (**5 points**)

   (d) Your co-author is concerned that the regression in part (a) does not have a causal interpretation. Specifically, she thinks that experienced teachers are less likely

to work with low-income students, who perform worse on tests in general. What does this say about the likely direction of omitted variables bias? Explain, using the OVB formula. (**3 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. (**7 points**)

(f) Now use the same data to demonstrate the "regression anatomy" formula below. 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. (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. (**7 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. (**7 points**)

(h) Finally, the user-written command `reganat` (short for "regression anatomy") allows you to visually compare bivariate relationships with those that account for the correlation between the X of interest and other covariates. Install this command using the syntax below, and then use it to show the relationship between *mathz* and *totexp* after "purging" *totexp* of its correlation with the other covariates *econdis, mathz_1,* and *lep*. Provide a written interpretation of what this graph shows you.

```
ssc install reganat
```

Hint: use the options `dis(totexp)` and `biline` with `reganat` to get the desired graph. `biline` superimposes the bivariate regression line for comparison. This is a pretty large dataset, so if you want to see a cleaner scatterplot, you can take a random sample of the observations using the command `sample 500, count` (for a sample of 500). (**7 points**)

2. 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**)

3. Use the syntax below to create a "toy" dataset of potential outcomes for 10 individuals. (Note: this example dataset was taken from the *Mixtape* Potential Outcomes chapter). **(14 points)**

   (a) What is the ATE? The ATT? The ATU? Show in your syntax how you calculated these values, and show that the ATE is a weighted average of the ATT and ATU. **(5 points)**

   (b) What is the simple difference in mean observed outcomes between the treated and untreated cases? If this estimator were used for the ATT, what would the selection bias be? If this estimator were used for the ATE, what would the selection bias be? **(5 points)**

   (c) Suppose D were randomly assigned. Would this guarantee that the simple differences in means gives you the ATE? Why or why not? **(4 points)**

```
clear
set obs 10

gen  y1 = 7 in 1
replace y1 = 5 in 2
replace y1 = 5 in 3
replace y1 = 7 in 4
replace y1 = 4 in 5
replace y1 = 10 in 6
replace y1 = 1 in 7
replace y1 = 5 in 8
replace y1 = 3 in 9
replace y1 = 9 in 10
```

```
gen  y0 = 1 in 1
replace y0 = 6 in 2
replace y0 = 1 in 3
replace y0 = 8 in 4
replace y0 = 2 in 5
replace y0 = 1 in 6
replace y0 = 10 in 7
replace y0 = 6 in 8
replace y0 = 7 in 9
replace y0 = 8 in 10

gen d = 1 if inlist(_n, 1, 3, 5, 6, 10)
replace d = 0 if d==.

gen y = y1*d + y0*(1-d)
```