Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Problem Set 1
62 total points
Due September 9, 2021

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 or .txt file. Use your last name and problem set number as the filename (e.g., *Fauci Problem Set 1.pdf*). The resulting log should include the questions below (commented), your commands, and output. Edit this file as appropriate, with any requested interpretations of your output. Graphical output can be submitted separately, preferably as a 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 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. (**41 points**)

   (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 your answers. (**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. (**5 points**)

   (c) Install the user-written .ado file called `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. (**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. Do these results conform with your answer in part (d)? Provide an interpretation in words 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) Finally, your co-author remains unsatisifed with this regression specification and recommends you also control for *mathz_1*, the student's math score in the prior grade. Estimate the multivariate regression with *totexp, econdis,* and *mathz_1*. Provide an interpretation, in words, of the three 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**)

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. Demonstrate that you understand how bootstrapping works by doing the steps below. Use the *HSLS-09 extract* dataset available on Github. This is a sample of 500 students from the High School Longitudinal Study of 2009. (**16 points**)

(a) Estimate the following simple regression which relates the student's standardized math score to a measure of their family's socioeconomic status: $x1txmtscor = \beta_0 + \beta_1 x1ses + u$. Interpret the estimated coefficient on $x1ses$ (call this $\hat{\beta}_{1,OLS}$). Is this a large effect size? Explain your rationale for assessing the effect size. (**3 points**)

(b) Now, bootstrap the sampling distribution of $\hat{\beta}_1$, using 250 replications. Each replication should be a bootstrap sample of size N=500. Do this manually by writing a loop, not with the `bootstrap` command. (Hint: the command `bsample` will be helpful here). Save your estimates from each replication, report your

bootstrapped standard error for $\hat{\beta}_1$, and a 90% percentile interval for $\hat{\beta}_1$. Give a written interpretation of these two things, and provide a histogram of your coefficient estimates. (**7 points**)

(c) Next, use the `bootstrap` command with `regress` to obtain the bootstrapped standard error. This will be much easier than doing it manually! (**3 points**)

(d) State whether the following statement is true or false. If false, explain why. (**3 points**)

Bootstrapping is a useful procedure, but relies on an assumption of normality for the underlying sampling distribution of $\hat{\beta}$.