Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Lecture 1 In-Class Exercise

1. **Randomized controlled trials.** In a well-known study, Howell and Peterson (2006) evaluated the effects of a private school voucher in NYC from the School Choice Scholarships Foundation (SCSF). This program provided scholarships of up to \$1,400 for 1,300 children from low-income families to attend a private elementary school. There were more applicants to the program than vouchers, so a random lottery was used to award the scholarships. Ultimately, 1,300 families received the voucher and 960 didn't.

   (a) Let $D_i = 1$ if the student was offered a voucher and $D_i = 0$ if not. Suppose we wanted to estimate the simple population regression function below, where Y_i represents student achievement after three years of the program:

   $$Y_i = \beta_0 + \beta_1 D_i + u_i$$

   Under what conditions does this PRF describe "differences in average potential outcomes for a well-defined population" (our criteria for causal interpretation)? Do those conditions hold here? How would you describe the relevant population? What is our *estimand* of interest (ATE, ATT, ATU, something else)?

   (b) Read the following dataset from Github which contains a subsample of 521 African-American students who participated in the lottery:

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

   (c) Use `ttest` and the simple regression model above to estimate the effects of the voucher (*voucher*) on student achievement after three years of the program (*post_ach*). Is the estimated effect statistically significant? Practically significant? (The outcome variable is a composite measure of reading and math achievement, expressed as a national percentile score).

   (d) Randomization (e.g., via lottery) in theory should prevent omitted variables bias. However, in finite samples, there may be *incidental* (chance) correlation between treatment assignment and other predictors of the outcome. The first step in the analysis of any RCT is to "check for balance" between the treated and untreated group on a host of baseline predictors. The only other variable in this dataset is a measure of baseline achievement, *pre_ach*. How does this measure differ between the treated and untreated group? (You can compare both means and other features of the distribution).

(e) Add the *pre_ach* measure to the regression function below (as X). What purpose does this serve? How does this additional covariate change your point estimate for β_1 (if at all)? How does it change the standard error for β_1 (if at all)?

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

2. **Simulated data 1** This problem will estimate population regression functions using data from a known population that we define ourselves. Draw a $N = 100$ random sample of three independent $N(0,1)$ variables: x_1, x_2, and u. The relevant command in Stata is `drawnorm`. From these, generate two outcome variables: $y_1 = 10 + x_1 + u$ and $y_2 = 10 + x_1 + 2x_2 + u$. Note: if you want to be able to replicate work done with randomly generated values in Stata, put the `set seed #` command at the beginning of your do-file. You will then get the same set of random numbers every time you run your program.

(a) What is the population mean of y_1, $E[y_1]$? What is the population variance of y_1, σ^2_{y1}? What is the conditional expectation function $E[y_1|x_1]$? Is it linear? What is the *conditional variance* of y_1 given x_1? Note: these questions can be answered without use of the data.

(b) What is the population mean of y_2, $E[y_2]$? What is the population variance of y_2, σ^2_{y2}? What is the conditional expectation function $E[y_2|x_1]$? Is it linear? Note: these questions can be answered without use of the data.

(c) Regress y_1 on x_1 (i.e., estimate the model $y_1 = \beta_0 + \beta_1 x_1$ using OLS). Note the slope coefficient and its standard error. Do the intercept and slope equal the known population intercept and slope? Why or why not?

(d) Regress y_2 on x_1 (i.e., estimate the model $y_2 = \tilde{\gamma}_0 + \tilde{\gamma}_1 x_1$ using OLS). Note the slope coefficient and its standard error. If you are interested in an unbiased estimate of the slope on x_1 in the population regression function for y_1, will your slope estimator suffer from omitted variables bias? Why or why not?

(e) Now regress y_2 on x_1 and x_2 (i.e., estimate the model $y_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2$ using OLS). Why does $\hat{\gamma}_1$ differ from $\hat{\tilde{\gamma}}_1$, even though we know the population correlation between x_1 and x_2 is zero?

(f) Compare the estimated standard errors on $\hat{\tilde{\gamma}}_1$ from part (d) and $\hat{\gamma}_1$ from part (e). How and why did it change?

(g) Now modify x_2 to purge it of any sample correlation with x_1. Call this variable x_{2a}. Hint: you are looking for variation in x_2 that is orthogonal to ("not explained" by) x_1.

(h) Generate a new y_2 (call it y_{2a}) using x_{2a} in place of x_2. Repeat parts (d) and (e). What changed, and why? Why does the standard error on $\hat{\gamma}_1$ change with the inclusion of x_{2a}, when we know x_{2a} is uncorrelated (by construction) with x_1?

(i) Return to part (c). Compare the reported standard error for $\hat{\beta}_1$ to the *population* standard error for $\hat{\beta}_1$. Hint: you know the population σ^2.

(j) Start with an empty dataset and recreate your random variables x_1, u, and y_1, but this time draw a $N = 10,000$ random sample. Repeat part (i). Now how do your reported $\hat{\beta}_1$ and standard error for $\hat{\beta}_1$ compare to the population β_1 and standard error for $\hat{\beta}_1$?

3. **Simulated data 2.** This problem is similar to #2, but we will assume x_1 and x_2 come from a *bivariate normal* distribution, so that we know x_1 and x_2 are correlated. The relevant command in Stata is `drawnorm`, but we need to specify a correlation matrix for the distribution (call this $\mathbf{C}$). σ^2_{x1} and σ^2_{x2} will continue to be 1, but assume they have a correlation of 0.5. Continue to use $N = 100$. Create the outcome variable $y_2 = 10 + x_1 + 2x_2 + u$. See the syntax below for the `drawnorm` command and its correlation matrix.

```
clear
matrix C = (1, .5 , 0 \ .5, 1, 0 \ 0, 0, 1)
drawnorm x1 x2 u, n(100) corr(C)
```

(a) What is the population variance of y_2? How does this compare with your answer in question #2 part (b)?

(b) For fun, use the user-written Stata command `tddens` to visualize the bivariate distribution of (x_1, x_2) as a "heat map".

(c) Regress y_2 on x_1. Note the slope coefficient and its standard error. If you are interested in an unbiased estimate of β_1 (the slope coefficient on x_1 in the population), does this regression suffer from omitted variables bias? Why or why not? If so, in what direction is the bias?

(d) Now regress y_2 on x_1 and x_2. What changed, and why?

(e) Apply the "regression anatomy" formula. That is, show that $\hat{\beta}_2$ is equal to the slope coefficient from a simple regression of y_2 on $\tilde{x}_2$, where $\tilde{x}_2$ is the residual from a regression of x_2 on x_1. Equivalently, $\hat{\beta}_2 = Cov(y_2, \tilde{x}_2)/Var(\tilde{x}_2)$.

(f) Demonstrate the omitted variables bias formula by showing the coefficient in the "short" regression (part c) is equal to the coefficient on x_1 in the "long" regression (part d) + the product of β_2 (the coefficient on x_2 in the "long" regression) and π (the coefficient from a regression of the omitted x_2 on the included (x_1)).