Regression II
Vanderbilt University
Prof. Sean P. Corcoran

---

**Lecture 1 In-Class Exercises**

---

1.1 **Potential outcomes and treatment effects.** (Adapted from the *Mixtape* chapter). See Lecture 1 do file for code.

(a) Input the potential outcomes $(Y_{i0}, Y_{i1})$ and treatment status $(D_i)$ for 10 cases. These represent post-intervention lifespan in years for two alternative cancer treatments (chemotherapy and surgery).

(b) Calculate the individual treatment effects $(\tau_i)$ for these cases. Note the effect varies by individual in this example.

(c) Using the known $\tau_i$ calculate the ATE, ATT, and ATU. Of course, with real data we can never calculate treatment effects since $\tau_i$ is unobservable. (We can't observe both $Y_{i0}$ and $Y_{i1}$).

(d) Calculate the *observed* $Y_i$ using the switching equation.

(e) Calculate the simple difference in the (observed) mean $Y_i$ between the treated and untreated groups.

(f) Given what you know about $Y_{i0}$ and $\tau_i$ calculate the selection bias and heterogeneous treatment effect bias. Interpret in words.

1.2 **Estimating treatment effects with randomization.** (From the *Mixtape* chapter). See Lecture 1 do file for code.

(a) Write a program that randomly assigns treatment status $D_i$ to each of the ten cases in exercise 1.1 and calculates the simple difference in (observed) means between the treated and untreated groups. This program should iterate 10,000 times and collect the resulting sample means.

(b) Examine the sampling distribution of resulting point estimates. What is its mean? What is its standard deviation?

1.3 **Simulated data with selection into treatment on $\boldsymbol{X}$.** Based on lecture notes example. See Lecture 1 do file for code.

(a) Create a simulated dataset with 10,000 observations that includes the following:

- $D_i$=1 for 40% of cases (selected at random) and $D_i = 0$ otherwise.
- A covariate $X_i$ that is related to $D_i$ in this way: $X_i = 2 + 1.5D_i + w_i$, where $w_i$ is $N(0, 1)$. In words, the covariate $X_i$ is 1.5 higher, on average, for treated cases.
- Potential outcomes $Y_0$ and $Y_1$ that depend on $X$. The $u_i$ term is $N(0, 5)$. Note the constant treatment effect of 6.

  $Y_0 = 15 + 5X_i + u_i$

  $Y_1 = 15 + 5X_i + 6 + u_i$

(b) As in 1.1, calculate the individual treatment effects $\tau_i$, ATE, ATT, and ATU.

(c) As in 1.1, calculate the observed $Y_i$ and simple difference in means for the treated and untreated group. Is the simple difference equal to the ATE or ATT?

(d) Estimate the simple OLS regression of $Y_i$ on $D_i$. How does the slope coefficient differ from the known treatment effect of 6? (This is the OVB).

(e) Now estimate an OLS regression of $Y_i$ on $D_i$ and $X_i$. Does this address the OVB?

(f) Show that OVB in part (d) is equal to the slope coefficient on $X_i$ in the "long regression" (e) times the slope coefficient from a regression of $X_i$ on $D_i$.

1.4 **RCT of private school vouchers.** 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 regression 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 would this regression 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 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. (This can also be revealing about whether the randomization "worked.") 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 $\beta_1$ (if at all)? How does it change the standard error for $\beta_1$ (if at all)?

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