EC508: Econometrics
Causality and the Multiple Regression Model

Jean-Jacques Forneron

Spring, 2023

Boston University

What, precisely, is a causal effect?

- "Causality" is a complex concept!
- In this course, we take a practical approach to defining causality:
- **A causal effect is defined to be the effect measured in an ideal randomized controlled experiment.**

Ideal Randomized Controlled Experiment

- **Ideal**: subjects all follow the treatment protocol – perfect compliance, no errors in reporting, etc.!
- **Randomized**: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- **Controlled**: having a control group permits measuring the differential effect of the treatment
- **Experiment**: the treatment is assigned as part of the experiment: the subjects have no choice, so there is no "reverse causality" in which subjects choose the treatment they think will work best.

Back to class size:

- Imagine an ideal randomized controlled experiment for measuring the effect on Test Score of reducing STR...
- In that experiment, students would be randomly assigned to classes, which would have different sizes.
- Because they are randomly assigned, all student characteristics (and thus ui) would be distributed independently of STRi.
- Thus, $\mathbb{E}(u_i|\text{STR}_i) = 0$ - that is, LSA #1 holds in a randomized controlled experiment.

How does our observational data differ from this ideal?

- The treatment is not randomly assigned
- Consider PctEL – percent English learners – in the district. It plausibly satisfies the two criteria for omitted variable bias: $Z = PctEL$ is:
  1. a determinant of Y; and
  2. correlated with the regressor X.
- Thus, the "control" and "treatment" groups differ in a systematic way, so $corr(STR, PctEL) \neq 0$

Return to omitted variable bias

Three ways to overcome omitted variable bias

1. Run a randomized controlled experiment in which treatment (STR) is randomly assigned: then PctEL is still a determinant of TestScore, but PctEL is uncorrelated with STR. (This solution to OV bias is rarely feasible.)
2. Adopt the "cross tabulation" approach, with finer gradations of STR and PctEL – within each group, all classes have the same PctEL, so we control for PctEL (But soon you will run out of data, and what about other determinants like family income and parental education?)
3. Use a regression in which the omitted variable (PctEL) is no longer omitted: include PctEL as an additional regressor in a multiple regression.

The Population Multiple Regression Model (SW 6.2)

- Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 \times X_{1,i} + \beta_2 \times X_{2,i} + u_i, i = 1, \ldots, n$$

- Y is the dependent variable
- X_1, X_2 are the two independent variables (regressors)
  $(Y_i, X_{1,i}, X_{2,i})$ denote the ith observation on Y, X_1, and X_2.
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding X_2 constant
- β_2 = effect on Y of a change in X_2, holding X_1 constant
- u_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 \times X_{1,i} + \beta_2 \times X_{2,i} + u_i, i = 1, \ldots, n$$

- Consider changing X_1 by ΔX_1 while holding X_2 constant:
- Population regression line before the change:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Population regression line, after the change:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

Interpretation of coefficients in multiple regression, cont'd

- Before: $Y_i = \beta_0 + \beta_1 \times X_{1,i} + \beta_2 \times X_{2,i}$
- After: $Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$
- Difference: $\Delta Y = \beta_1 \Delta X_1$
- As a result:
  - $\beta_1 = \frac{\Delta Y}{\Delta X_1}$ holding X_2 constant
  - $\beta_2 = \frac{\Delta Y}{\Delta X_2}$ holding X_1 constant
  - $\beta_0 =$ predicted value of Y when $X_1 = X_2 = 0$

The omitted variable bias, illustration

- With two regressors, the OLS estimator solves:

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_{1,i} + b_2 X_{2,i}])^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators** of β_0, β_1 and β_2.

Example: the California test score data

- Regression of TestScore against STR:

$$\text{Test Score} = 698.9 – 2.28 \times \text{STR}$$

- Now include percent English Learners in the district (PctEL):

$$\text{Test Score} = 686.0 – 1.10 \times \text{STR} – 0.65 \times \text{PctEL}$$

- What happens to the coefficient on STR?
- It becomes closer to 0 by 1.18, about half of the initial estimated effect

Application to the California Test Score in R

```
# packages to compute standard errors
library(sandwich)
library(lmtest)

library(foreign)
data = read.dta('caschool.dta')
data$score = 0.5*(data$math_scr + data$
   read_scr)
linear_model = lm(score~str+el_pct,data=
   data)

# compute standard errors, t-statistics
coeftest(linear_model, vcov. = vcovHC)
```

Table 1: Coefficients, Standard Errors, t-statistics and p-values

	Estimate	Std. Error	t-value	Pr($>\|t\|$)
(Intercept)	686.032244	8.812242	77.8499	$<$ 2e-16 ***
str	-1.101296	0.437066	-2.5197	0.01212 *
elpct	-0.649777	0.031297	-20.7617	$<$ 2e-16 ***