Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 9

1. Bias and Efficiency

2. OLS Assumption 1

3. Omitted Variable Bias

Where are going...

Today:

- Bias and Efficiency
- Introduction to Multivariate regression
  - OLS Assumption 1 & omitted variable bias
- Reading:
  - NONE
- Homework:
  - Homework Assignment #9 posted on D2L

Next Week

- Multivariate regression cont.
- Reading:
  - Empirical manuscripts

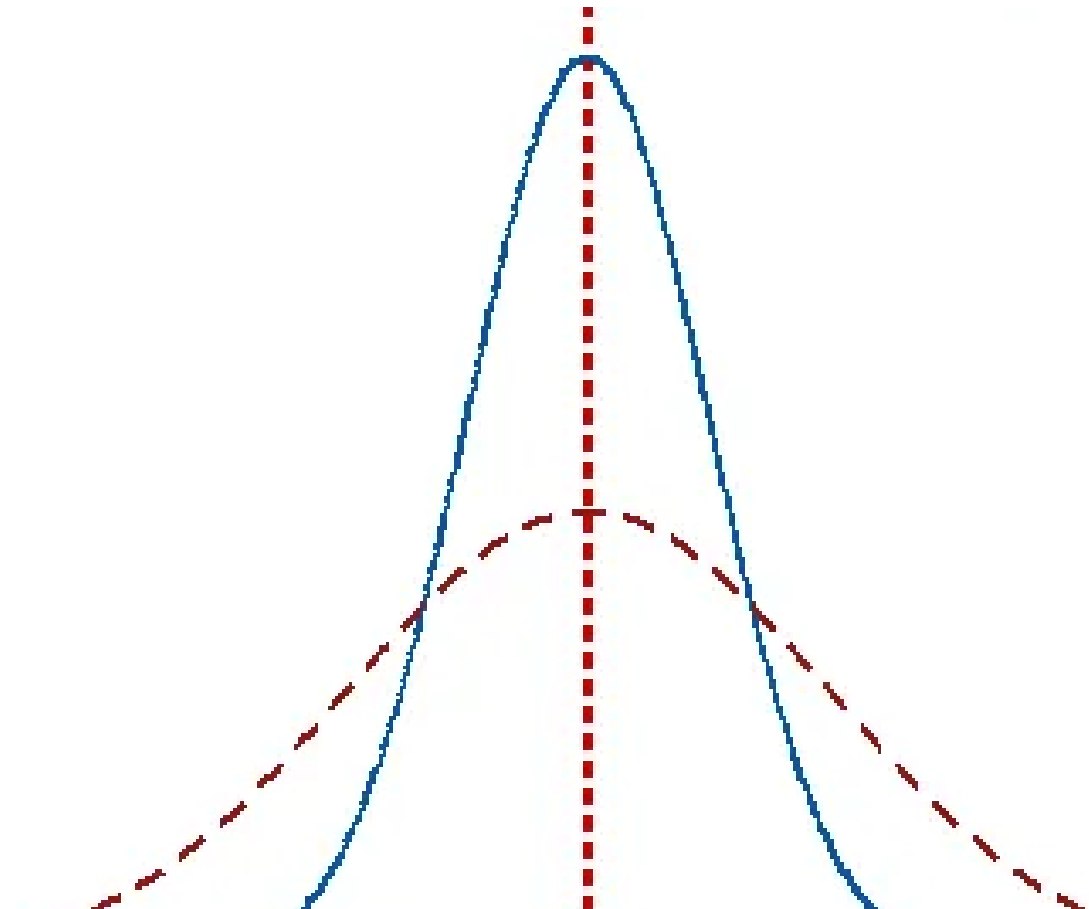
Following week

- Read empirical quantitative work

Bias and Efficiency

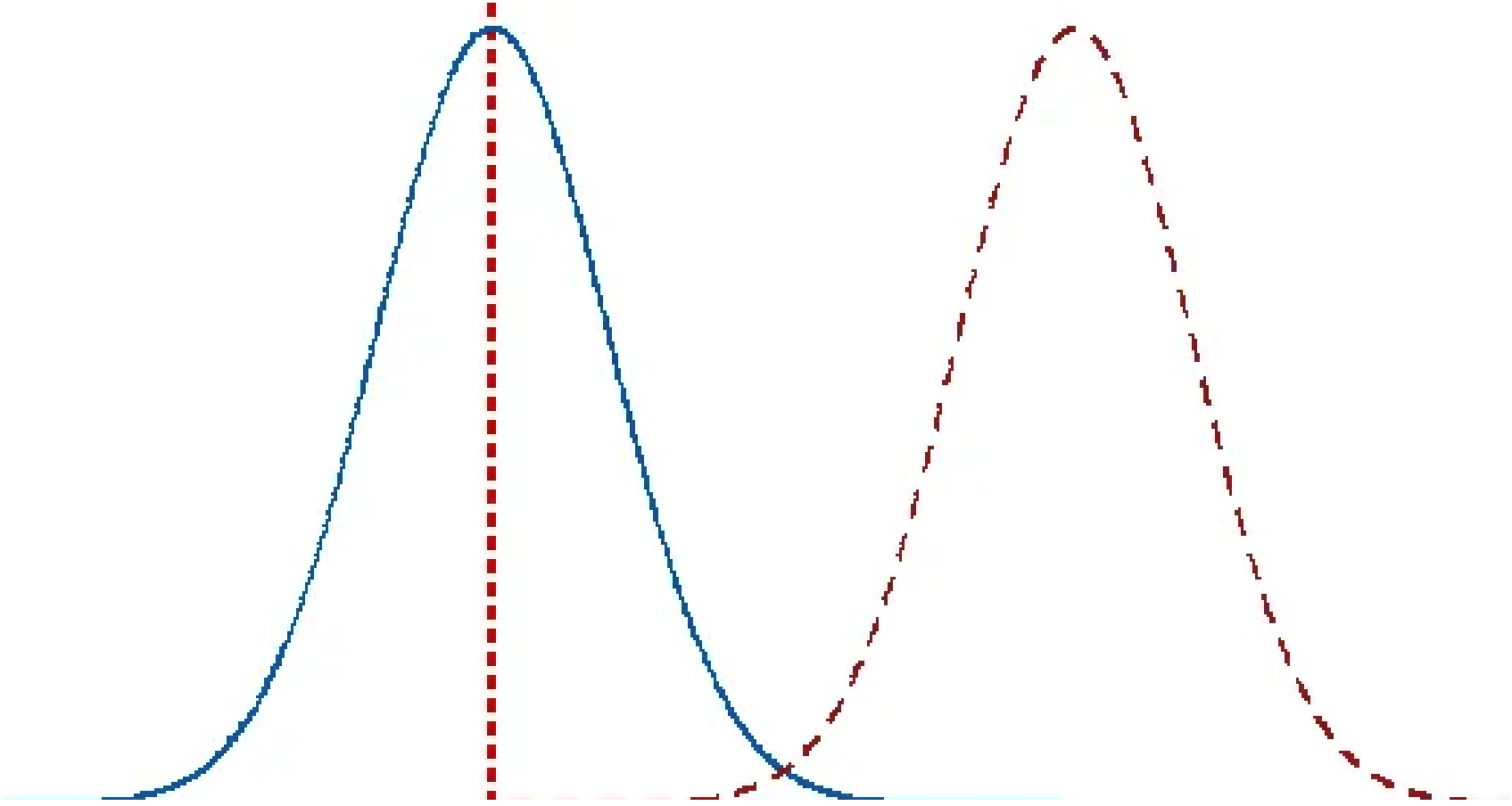
Efficiency, also called "precision"

- Desirable properties of our point estimates (e.g. $\hat{\beta}$ or $\bar{Y}$)
  - Efficient
  - Unbiased
- **Efficiency**
  - Definition: how close your point estimate(s) is to the population parameter
  - Standard Error: on average, how far away is a point estimate from one random sample from the value of the population parameter
  - *Therefore*, an efficient point estimate is one with a low standard error (in other words, the sampling distribution of β_1 has low variance or is "tight" around the population parameter)



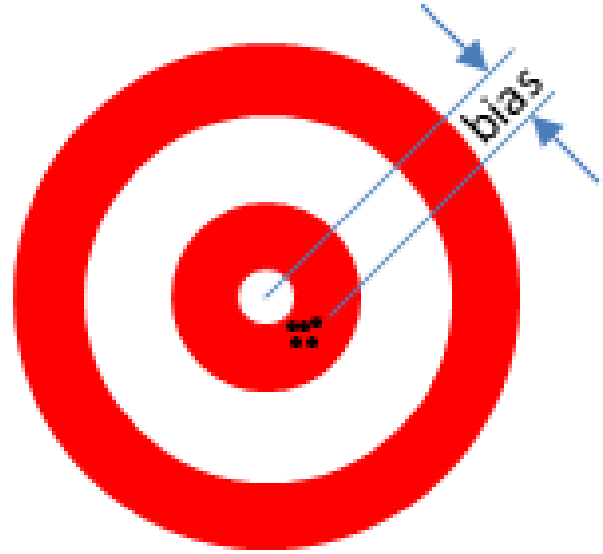
Bias

- **Bias**: consistently overestimates or underestimates population parameter in repeated random samples
  - There are many different types of bias!
- *Sampling Bias*:
  - Estimate of population parameter is biased because you fail to take a random sample
  - Example: goal is to estimate high school graduation rate; take a random sample of 10th graders and see if they graduate within three years.
- *Omitted variable bias*:
  - Bias in estimate of β_1 due to omitting necessary "control" variables in your regression model



Bias and Efficiency (or "Precision")

unbiased, precise



biased, precise

unbiased, imprecise

biased, imprecise

OLS Assumption 1

Prior to assumptions: define causal effect

- Asking causal inference research questions
  - "What is the effect of X on Y"
  - Goal: to estimate the "causal effect" of X on Y
- What is a "causal effect"?
  - Stock and Watson define it as "what would happen in a randomized experiment"
  - Causal effect is the average effect of being in the "treatment" group as opposed to the "control" group on the value of Y if people were randomly assigned to groups
- How do you know if you are asking a causal inference research question?
  - As yourself, what is the relevant randomized experiment?
  - That is, how would your question be designed as a randomized experiment?

Today's example & defining causal effect

- RQ: What is the effect of federal financial aid on students' access to college (Cellini, 2008)?
  - Is this a causal inference research question?
  - What is the relevant experiment?
  - Are students randomly assigned to receive federal financial aid?
- Experimental data vs. Observational data
  - Experimental data: people randomly assigned to "treatment" vs. "control" group
- Cellini (2008): Multivariate regression can be used to deal with the problem that the "treatment" (in this case receiving federal financial aid) wasn't randomized in order to assess the effect of X on Y (receiving financial aid on college access).
- We do our best to recreate experimental conditions for observational data by minimizing omitted variable bias (more on this later)!

OLS Assumption 1 (mathematically)

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= years of schooling (12+ = attended college), X= 0/1 received financial aid, u_i = all other variables that affect Y but were not included in the model
- OLS Assumption 1 (in words)
  - the independent variable X_i is unrelated to the "other variables" not included in the model, u_i
- OLS Assumption 1 (mathematically)
  - $E(u_i|X_i) = 0$; the expected value of u_i, given any value of X_i, equals zero
  - In other words:
    - Pretend that u_i consists of only one variable
    - OLS assumption 1 states that the mean value of the omitted variable is equal to zero no matter the value of variable X_i

OLS Assumption 1

- OLS Assumption 1: $E(u_i|X_i) = 0$
- **Cellini 2008**: Assumption is *always* satisfied in random assignment experiment
  - Effect of financial aid (X) on college access
  - X=0 (did not receive financial aid); X=1 (received financial aid)
  - We randomly assign students to receive versus not receive financial aid
  - Other factors u_i (e.g., academic achievement, socioeconomic status) are *by construction* unrelated to values of X because we randomly assigned students to X=1 or X=0
- **Cellini 2008**: in observational studies (like analyzing what is the effect of financial aid on student access to college), this assumption is usually violated!
  - For example, Receiving financial aid (X) is likely correlated with omitted variables, u_i that have an effect on Y (e.g., academic achievement, socioeconomic status)

OLS Assumption 1 in practice

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= years of schooling (12+ = attended college), X= 0/1 received financial aid, u_i = all other variables that affect Y but were not included in the model
- OLS Assumption 1 (in words)
  - the independent variable X_i is unrelated to the "other variables" not included in the model, u_i
- In Practice:
  - Are there any variables that are not in your model that...
  - **(1) Affect Y and (2) have a relationship (e.g. correlated) with X**?
  - If so, OLS Assumption 1 is violated
- Can you think of another omitted variable that violates OLS Assumption 1 for this RQ?

"No relationship" vs "No correlation"

- Note: "no relationship" vs "no correlation"
  - $E(u_i|X_i) = 0$ implies that u_i and X_i have no relationship (includes linear and non-linear)
  - $Corr(u_i|X_i) = 0$ implies that u_i and X_i have no *linear relationship*
    - e.g., Pearson's R correlation coefficient
- So correlation might be zero, but $E(u_i|X_i) \neq 0$ due to the existence of a non-linear relationship

Omitted Variable Bias

Introduction to Omitted Variable Bias

- $Y_i = \beta_0 + \beta_1 X_i + u_i$; Y=test score; X=class size
  - We want to know the *causal effect* of X on Y
- Bias (general): when $\hat{\beta}_1$ consistently underestimates β_1 or overestimates β_1
- **Omitted Variable Bias**: bias in $\hat{\beta}_1$ due to variables being omitted from the model
- **For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions**:
  - (1) Z affects value of Y (i.e. Z is part of u_i)
  - (2) *and* Z has a relationship with X

Omitted Variable Bias Example

- $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= average class reading test score
  - X= class size
  - Z= % of ELL students (omitted from model)
- For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:
  - (1) Z affects value of Y (i.e. Z is part of u_i);
  - (2) **and** Z has a relationship with X
- How does % of ELL students satisfy criteria of omitted variable bias?
  - (1) % of ELL affects value of average reading test scores (ELL students are likley to score at lower reading levels than native english speakers);
  - (2) **and** % of ELL students has a relationship with class size (policy: greater proportion of ELL students require smaller class sizes)
- Would omitting Z = "time of test administered" result in omitted variable bias?
- Would omitting Z = "teacher's years of experience" result in omitted variable bias?

How to check for omitted variable bias

- Does Z affect Y?
  - Ask yourself if it is plausible that omitted variable Z affects Y
- Does Z have a relationship with X?
  - Ask yourself if it is plausible that omitted variable Z has some relationship with X
  - Logical argument or diagnostic tests (e.g., `df %>% summarise(cor(X, Z, use = "complete.obs"`)))
- In practice, diagnostic tests not used as much as logical arguments/literature review
  - Correlation only picks up linear relationships, omitted variable bias includes non-linear relationships
  - Relationship between X and Z is about "conditional relationship," after controlling for other covariates
- Sometimes you don't have a good measure of omitted variable Z

Group Exercise

For each research question below, identify two "hypothetical" variables that would result in a violation of OLS Assumption 1 (i.e., they meet the two conditions of Omitted Variable Bias)

- Be ready to explain how each variable meets the two conditions!

(1) Group 1: What is the effect of participating in a fraternity/sorority (X) on GPA (Y)?

(2) Group 2: What is the effect of participating in Head Start (X) on long-term academic achievement (Y)?