Empirical Methods

Topic 1b:

The CLRM: Foundations

The Classical Linear Regression Model Foundations

CLRM Intro I

- OK, let's get to it.
- In what follows, I'm going to present the regression framework commonly known as the Classical Linear Regression Model
- In a nutshell, the CLRM is two things:
  1. Estimation by Ordinary Least Squares (OLS)
  2. A set of assumptions under which OLS has good properties
- Let's do each in turn...

CLRM Intro II

- In what follows, I'm going to derive the properties of OLS under the assumptions of the CLRM using matrix notation
- We have a large audience with diverse backgrounds, so if I jump to the middle and say the OLS estimator in simple regression...
  - $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$
    - ★ (i.e., with only one element in x_i)...
- ...is written and calculated as:

$$\boxed{\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}} \qquad \boxed{\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \qquad \boxed{V(\hat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$$

  - How many of you have seen these before?
  - How many of you have *derived* these before?

CLRM Intro III

- And that the OLS estimator for multiple regression...
  - $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \epsilon_i$
    - ⋆ (i.e., with K elements in x_i)...
- ...is written and calculated as:

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y} \qquad \boxed{V(\hat{\beta}) = \sigma^2(X'X)^{-1}}$$

  - How many of you have seen these before?
  - How many of you have *derived* these before?

CLRM Intro IV

- Even if you have seen these before...
  - ...and even if you have derived them before...
  - ...I'm going to present them again (well, the latter anyway)...
  - ...and - critically - give you some intuition about them
- As well as all the results in the CLRM
  - So I very much expect to show/tell you things that you *haven't* seen before

The CLRM in Matrix Notation

Latex Color = "LightGrey"

Notation

- A challenge in learning econometrics is the lack of standard notation across sources,
  - Particularly when you get to IV estimation
- Understanding notation is critical, however, when you code an estimation routine on a computer
- In the exercise sessions last week, Matteo showed you the notation we would use in this course

The CLRM in Two Notations I

- As shown in the exercise sessions, we can write the CLRM in one of two ways
- First we can write the CLRM enumerating each observation, i:

$$\boxed{y_i = x_i'\beta + \epsilon_i, \qquad i = 1, \ldots, N}$$

  - where
    - y_i is a scalar (i.e. is 1×1)
    - x_i' is a $1 \times K$ row vector,
    - β is a $K \times 1$ column vector, and
    - ϵ_i is a scalar
  - This is the basis for what is called "summation notation"
    - (The reasons for which will be clear later)

The CLRM in Two Notations II

- Alternatively we can write the CLRM by stacking all the observations into vectors/matrices:

$$\boxed{y = X\beta + \epsilon}$$

  - where
    - y is a $N \times 1$ column vector
    - X is a $N \times K$ matrix,
    - β is a $K \times 1$ column vector, and
    - ϵ is a $N \times 1$ column vector
  - This is the basis for what is called "matrix notation"

The CLRM in Matrix Notation III

- When one is using matrix notation, you must make sure your matrices are *conformable*
  - For matrix addition: the matrices must have the same number of rows and columns
  - For matrix multiplication: the # columns in the left matrix must equal the # rows in the right
- Conformability is *sneaky important*:
  - It both (a) ensures you don't make a mistake and (b) helps you understand what's going on in each matrix equation
  - With time and practice you will build an intuition for how relationships between random variables map into your math (and vice versa)

Derivation of the OLS formula

- In the exercise session, Matteo derived the OLS formula using matrix notation, yielding:

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y}$$

Notation for OLS Matrices I

$$\boxed{y = X\beta + \epsilon} \qquad \boxed{\hat{\beta} = (X'X)^{-1}X'y}$$

- Let's dig into the notation of the OLS formula a bit
  - X is $N \times K$, so X' is $K \times N$
  - Write X' as a row vector with typical element x_i

$$X' = \begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_N \end{bmatrix}$$

  - Write X as a column vector with typical element x_i'

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_i' \\ \vdots \\ x_N' \end{bmatrix}$$

Notation for OLS Matrices II

- Then $(X'X) = \begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_N \end{bmatrix} \begin{bmatrix} x_1' \\ \vdots \\ x_i' \\ \vdots \\ x_N' \end{bmatrix} = \sum_{i=1}^{N} x_i x_i'$

  where each $x_i x_i'$ is a *
  - ⋆ And $(X'X)^{-1} = (\sum x_i x_i')^{-1}$

- Similarly $(X'y) = \begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} = \sum_{i=1}^{N} x_i y_i$

  where each $x_i y_i$ is a *

Notation for OLS Matrices III

- The point here: Be comfortable with:

$$\boxed{X'X = \sum x_i x_i'} \qquad \text{and} \qquad \boxed{X'y = \sum x_i y_i}$$

- For the equation in each box...
  - I call the left-hand-side representation "matrix notation"
  - I call the right-hand-side representation "summation notation"

Notation for OLS Matrices IV

- And thus:

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y} \quad \Leftrightarrow \quad \boxed{\hat{\beta} = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i}$$

  - These are *equivalent representations*
  - They are merely two different ways to write the same thing
    - ★ (We often use the first for parsimony)
    - ★ (I often use the second for intuition)

Intuition for the OLS Formula

Latex Color = "LightGrey"

Intuition for the OLS formula I

- Indeed, let's use the second representation now to give a little intuition for the OLS formula
- To make things easier, let's assume a simple regression ($K = 2$)
  - i.e. there is only one (non-constant) element in x_i:
    - $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$

Intuition for the OLS formula II

- Recall the formula for the slope coefficient in a simple regression:

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\frac{1}{N-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N-1}\sum(x_i - \bar{x})^2}\end{aligned}$$

  - Does either the numerator or denominator (or both) remind you of any formulas you know?*
    - ★ _____

Intuition for the OLS formula III

- OK, halfway there. The OLS coefficient somehow relates to sample variances and covariances
- Let's plug in s_x^2 for the sample variance of x_i and s_{xy} for the sample covariance of x_i and y_i and do some more manipulations

$$\begin{aligned}
\hat{\beta}_2 &= \frac{\frac{1}{N-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N-1}\sum(x_i - \bar{x})^2} \\
&= \frac{s_{xy}}{s_x^2} \\
&= \frac{s_{xy}}{s_x s_x} \\
&= \frac{s_{xy} s_y}{s_x s_y s_x}
\end{aligned}$$

$$\boxed{\hat{\beta}_2 = r_{xy} \frac{s_y}{s_x}}$$

where r_{xy} is the sample correlation between x_i and y_i

Intuition for the OLS formula IV

$$\boxed{\hat{\beta}_2 = r_{xy}\frac{s_y}{s_x}}$$

- Now *this* is a (*more*) intuitive formula:
  - It says that the OLS slope coefficient, $\hat{\beta}_2$...
  - ...measures the sample correlation between x_i and y_i, r_{xy}...
  - ...measured in units-of-y-per-units-of-x, $\frac{s_y}{s_x}$.
    - ★ See Stata Example in class

Intuition for the OLS formula V

- OK, that was the intuition for the OLS formula when $K = 1$
  - What about for larger K?
  - What is the intuition then?
- I'll show this in the next deck as I have to introduce some new tools before I (easily) can
  - (Tho the short answer: it's a very similar intuition)
  - ($\hat{\beta}_k$ measures the *partial* correlation between x_{ik} and y_i)

Correlation versus Causation Redux

- Remember that statistics - including OLS - can only tell you *correlations* in the data
  - The OLS formula says that - in essence - each of the elements in $\hat{\beta}$ measures the (partial) correlation between each of the elements in x_i and y_i
- Important: This doesn't mean that we can yet make causal statements about "what will happen to y_i if I change x_{ik}"
  - We'll need "our CLRM assumptions" in order to do that
  - (They are coming up next!)

CLRM Assumptions

Latex Color = "LightGrey"

CLRM Assumptions Intro I

- As mentioned in the last deck, one important factor when choosing estimators is to look at their properties...
  - ...and to choose one that has nice properties
- As mentioned earlier in this deck, the CLRM is two things:
  1. Estimation by OLS
  2. A set of assumptions under which OLS has good properties

CLRM Assumptions Intro II

- Why should you care about assumptions???
- Three reasons:
  1. They let us evaluate the properties of the OLS estimator
     - ⋆ This is a **a very big deal**.
  2. They tell what to worry about
     - ⋆ i.e., which of the assumptions is/are most important for an estimator's nice properties
  3. They (may) tell us how to fix things
     - ⋆ If a particular assumption is violated, it focuses our attention on looking for weaker assumptions that might not be violated but would still allow an estimator with decent properties

Assumption 1: Linearity I

- There are Five assumptions underlying the CLRM
  - We introduced the first one in the last deck
- Assumption 1: Linearity
  - We assume that the population regression function takes a linear form:

$$\begin{aligned} y_i &= E(y_i|x_i) + \epsilon_i, \qquad i = 1, ..., N \\ &= \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \epsilon_i \end{aligned}$$

$$\boxed{y_i = x_i'\beta + \epsilon_i, \quad i = 1, ..., N}$$

- Or, in matrix notation:

$$\boxed{y = X\beta + \epsilon}$$

Assumption 1: Linearity II

A quick further detail:

- When we assume linearity here, we are assuming *linearity in* β
  - And *not* (necessarily) linearity in x_{ik}
- Thus a nonlinear-in-x_{ik} model...

$$y_i = \beta_1 + \beta_2 x_{i2}^2 + \beta_3 e^{x_{i3}} + \beta_4 \log(x_{i4}) + \epsilon_i$$

- ...can be rewritten as as a linear-in-x_{ik}^* model...

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2}^* + \beta_3 x_{i3}^* + \beta_4 x_{i4}^* + \epsilon_i \\ &= x_i^{*\prime} \beta + \epsilon_i \end{aligned}$$

- ...by simply relabeling our original variables:

$$x_{i2}^* = x_{i2}^2 \qquad x_{i3}^* = e^{x_{i3}} \qquad x_{i4}^* = \log(x_{i4})$$

Assumption 2: Mean-Zero Error I

- Our remaining assumptions are assumptions on the error term, ϵ_i
- Assumption 2: (Conditional-)Mean-Zero Error
  - We assume that the mean of the error term, ϵ_i, given the explanatory variables, x_i, is zero:

$$\boxed{E(\epsilon_i|x_i) = 0, \quad i = 1, ..., N}$$

  - Or, in matrix notation:

$$\boxed{E(\epsilon|X) = 0}$$

Assumption 2: Mean-Zero Error II

$$\boxed{E(\epsilon_i|x_i) = 0}$$

- Some intuition please?

See Figure in Class

Assumption 2: Mean-Zero Error III

$$\boxed{E(\epsilon_i|x_i) = 0}$$

- Two important consequences of our Mean-Zero Error assumption:
  1. $E(\epsilon_i|x_i) = 0 \Rightarrow E(\epsilon_i) = 0, \quad i = 1, ..., N,$ or $E(\epsilon) = 0$:
     - ★ By the *Law of Iterated Expectations*, the unconditional expectation of ϵ_i is given by the expectation with respect to x_i of the conditional expectation of ϵ_i given x_i:

$$\begin{aligned} E(\epsilon_i) &= E_{x_i} E[\epsilon_i|x_i] \\ &= E_{x_i} 0 \\ &= 0 \end{aligned}$$

     - ★ This is intuitive: if $E(\epsilon_i|x_i) = 0$ *for any given* x_i, it is no surprise it's still zero when we "average" across x_i

Assumption 2: Mean-Zero Error III

$$\boxed{E(\epsilon_i|x_i) = 0}$$

- Two important consequences, cont:
  2. $E(\epsilon_i|x_i) = 0 \Rightarrow Cov(x_i, \epsilon_i) = 0$:

$$\begin{aligned} Cov(x_i, \epsilon_i) &= E(x_i\epsilon_i) - E(x_i)E(\epsilon_i) \\ &= E_{x_i}E[x_i\epsilon_i|x_i] - \mu_x 0 \\ &= E_{x_i}\left(x_i E[\epsilon_i|x_i]\right) \\ &= E_{x_i}\left(x_i 0\right) \\ &= 0 \end{aligned}$$

  - ★ We will later rely on the zero-covariance implication of our Mean-Zero Error assumption, so be sure to remember this one!
  - ★ (FYI: I often say $Cov(x_i, \epsilon_i) = 0$ as a shorthand for $E(\epsilon_i|x_i) = 0$)

Assumption 3: Homoskedasticity I

- Assumption 3: Homoskedasticity
  - We assume that the variance of the error term, ϵ_i, is constant across observations:

$$\boxed{V(\epsilon_i) = \sigma^2, \quad i = 1, ..., N}$$

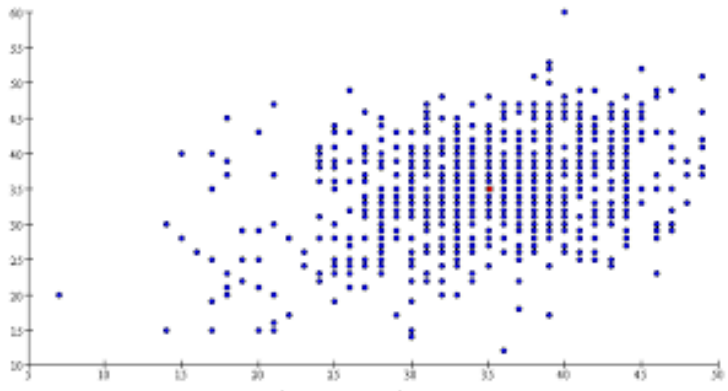
  - Homoskedasticity means the variance of the error term doesn't change across different values of i
    - ★ In particular: that it's not related to x_i
  - The opposite of Homoskedasticity is Heteroskedasticity
    - ★ Where $V(\epsilon_i)$ *does* vary with x_i
    - ★ (Often increasing with x_i)

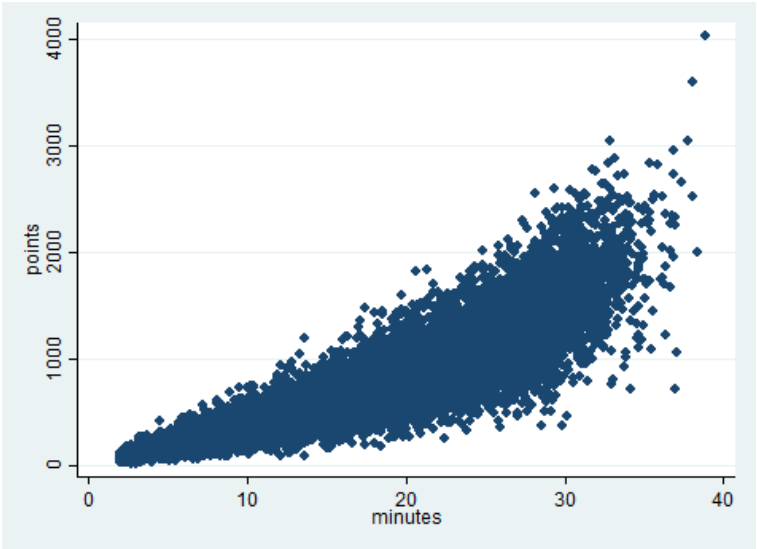
Assumption 3: Homoskedasticity II

- Some intuition please?
  - There are formal tests of all these assumptions, including this one
  - :-)
- We often look for Homo/Heteroskedasticity in two ways:
  1. In plots of y_i against x_i
  2. In plots of e_i against x_i
     - ★ Where e_i is the residual from the regression of y_i on x_i

Assumption 3: Homoskedasticity III

- Examples of plots of y_i versus x_i
  - With homoskedasticity on the left and heteroskedasticity on the right

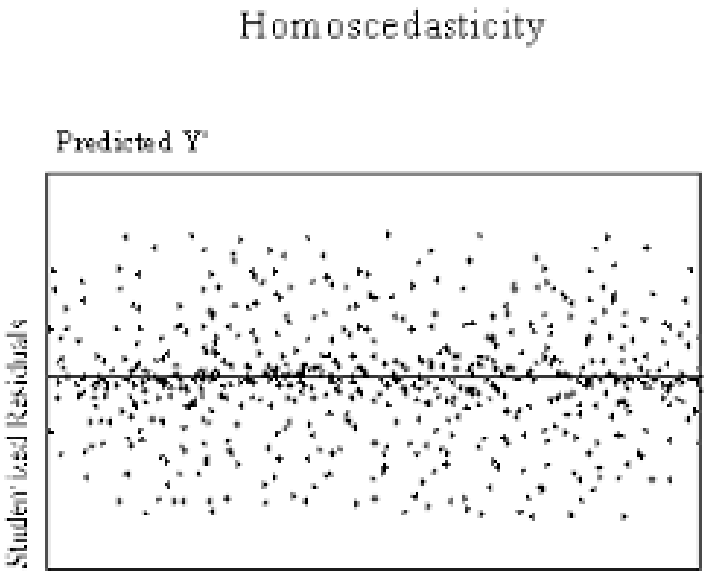


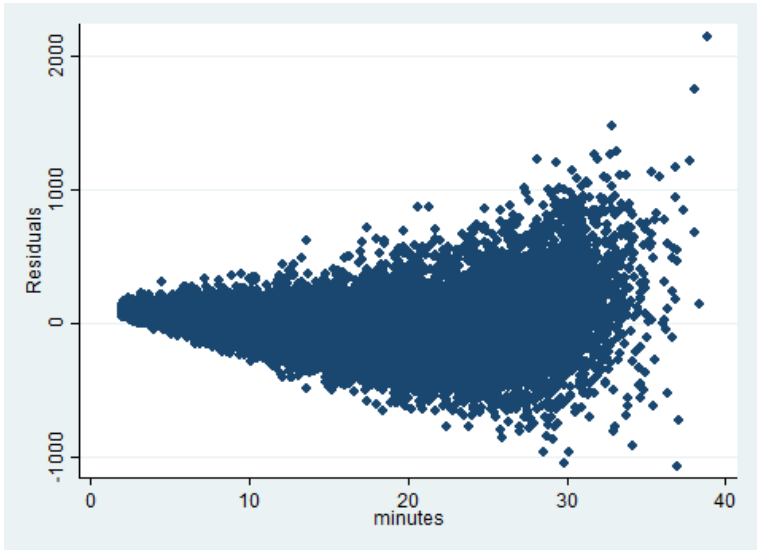


  - What are we looking for? *

Assumption 3: Homoskedasticity IV

- Examples of plots of e_i versus x_i
  - With homoskedasticity on the left and heteroskedasticity on the right





Assumption 4: No Correlation I

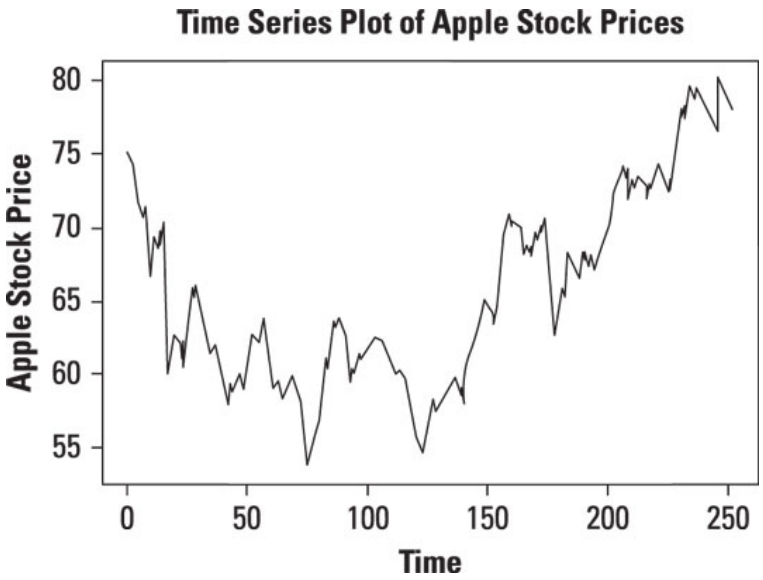
- Assumption 4: No (Serial) Correlation
  - We assume that the covariance between any two error term, ϵ_i and ϵ_j, is zero:

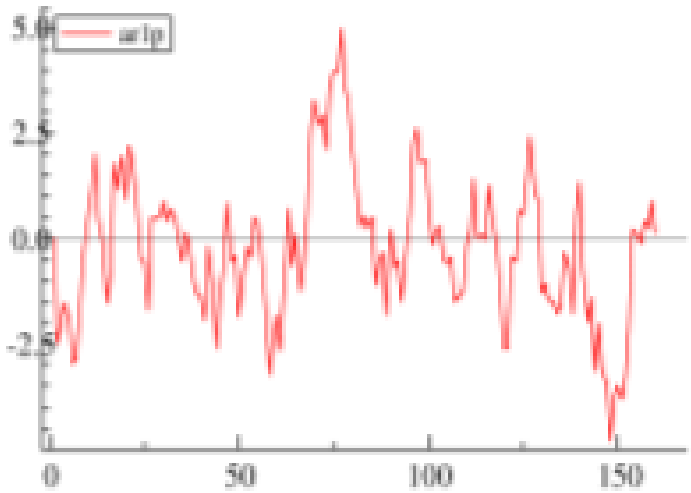
$$\boxed{Cov(\epsilon_i, \epsilon_j) = 0, \quad i, j = 1, ..., N, \quad i \neq j}$$

  - Serial correlation is most common in *time-series* data
    - ★ Especially positive serial correlation
    - ★ E.g., if GDP / stock prices / interest rates are above their mean in period $t-1$, they are often still above their mean in period t

Assumption 4: No Correlation II

- In time-series data, we again look for autocorrelation with plots of y_t versus t (left) and/or e_t versus t (right)





  - What are we looking for? *

Assumption 4: No Correlation III

- In cross-section data, correlation between ϵ_i and ϵ_j can occur due to *clustering*
  - When different i share important unobservable characteristics
    - ★ (Captured in ϵ_i)
  - Examples:
    - ★ A sample of workers, groups of whom come from the same firm
    - ★ A sample of cities, groups of whom come from the same state
- Just to be clear:
  - This is the stuff we're *assuming away*!
  - (For now)

Assumptions 3 and 4 in Matrix Notation I

- I haven't yet shown you how to write either Assumption 3 (Homoskedasticity) or Assumption 4 (No Correlation) in matrix notation
  - This is because it's easiest to do both at the same time
- We define the mean and variance of vectors of random variables in ways that are analogous to how we defined mean, variance, and covariance for individual pairs of random variables:
  - The mean of a random vector, ϵ, is:

$$E(\epsilon) = E\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_N) \end{bmatrix} = 0 \quad \text{under (A2, Mean-Zero Error)}$$

Assumptions 3 and 4 in Matrix Notation II

- The variance-covariance *matrix* of a random vector, ϵ, is:

$$
\begin{aligned}
V(\epsilon) &\equiv E(\epsilon\epsilon') - E(\epsilon)E(\epsilon') \\
&= E\left(\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_N \end{bmatrix}\right) - \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_N) \end{bmatrix} \begin{bmatrix} E(\epsilon_1) & E(\epsilon_2) & \cdots & E(\epsilon_N) \end{bmatrix} \\
&= \begin{bmatrix} E(\epsilon_1^2) - [E(\epsilon_1)^2] & E(\epsilon_1\epsilon_2) - [E(\epsilon_1)E(\epsilon_2)] & \cdots & E(\epsilon_1\epsilon_N) - [E(\epsilon_1)E(\epsilon_N)] \\ E(\epsilon_2\epsilon_1) - [E(\epsilon_2)E(\epsilon_1)] & E(\epsilon_2^2) - [E(\epsilon_2)^2] & \cdots & E(\epsilon_2\epsilon_N) - [E(\epsilon_2)E(\epsilon_N)] \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_N\epsilon_1) - [E(\epsilon_N)E(\epsilon_1)] & E(\epsilon_N\epsilon_2) - [E(\epsilon_N)E(\epsilon_2)] & \cdots & E(\epsilon_N^2) - [E(\epsilon_N)^2] \end{bmatrix}
\end{aligned}
$$

Assumptions 3 and 4 in Matrix Notation II

- For the CLRM, this equals:

$$
\begin{aligned}
V(\epsilon) &= \begin{bmatrix} E(\epsilon_1^2) - [E(\epsilon_1)^2] & E(\epsilon_1\epsilon_2) - [E(\epsilon_1)E(\epsilon_2)] & \cdots & E(\epsilon_1\epsilon_N) - [E(\epsilon_1)E(\epsilon_N)] \\ E(\epsilon_2\epsilon_1) - [E(\epsilon_2)E(\epsilon_1)] & E(\epsilon_2^2) - [E(\epsilon_2)^2] & \cdots & E(\epsilon_2\epsilon_N) - [E(\epsilon_2)E(\epsilon_N)] \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_N\epsilon_1) - [E(\epsilon_N)E(\epsilon_1)] & E(\epsilon_N\epsilon_2) - [E(\epsilon_N)E(\epsilon_2)] & \cdots & E(\epsilon_N^2) - [E(\epsilon_N)^2] \end{bmatrix} \\
&= \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_N) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_N\epsilon_1) & E(\epsilon_N\epsilon_2) & \cdots & E(\epsilon_N^2) \end{bmatrix} \\
&= \begin{bmatrix} V(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_N) \\ Cov(\epsilon_2, \epsilon_1) & V(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\epsilon_N, \epsilon_1) & Cov(\epsilon_N, \epsilon_2) & \cdots & V(\epsilon_N) \end{bmatrix} && \text{by definition of V/Cov when } E(\epsilon_i) = 0 \\
&= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} && \text{under (A3, Homoskedasticity) and (A4, No Correlation)} \\
&= \sigma^2 I_N && \text{where } I_N \text{ is an } N \times N \text{ identity matrix}
\end{aligned}
$$

- In sum: $\boxed{V(\epsilon) = \sigma^2 I_N}$ (Matteo will go over these slides more slowly in the exercise session!)

CLRM Assumptions Pre-Summary

- Assumptions 2, 3, and 4...
  - ...(Mean-zero Error, Homoskedasticity, and No Correlation)...
  - ...are often summarized as follows:

$$\boxed{\epsilon_i | x_i \sim (0, \sigma^2) \ , \quad Cov(\epsilon_i, \epsilon_j) = 0, \quad i, j = 1, ..., N, \quad i \neq j}$$

    - ★ (Where the conditioning of ϵ_i on x_i is often not made explicit)
    - ★ (i.e. you might only see $\epsilon_i \sim (0, \sigma^2)$ instead of $\epsilon_i | x_i \sim (0, \sigma^2)$)

- Or, in matrix notation (which is much easier):

$$\boxed{\epsilon | X \sim (0, \sigma^2 I_N)}$$

Assumption 5: Normality

- Assumption 5: Normality
  - We assume that ϵ_i is Normally distributed
  - Relative to the last slide, we add only one letter:

$$\epsilon_i | x_i \sim N(0, \sigma^2) \ , \quad Cov(\epsilon_i, \epsilon_j) = 0, \quad i, j = 1, ..., N, \quad i \neq j$$

- Or, in matrix notation,

$$\epsilon | X \sim N(0, \sigma^2 I_N)$$

- But it's an *important* letter...
  - ...as Normality is the assumption that let's us (easily) do hypothesis testing in the CLRM

CLRM Assumptions Summary I

- Using Summation Notation, we can summarize our five CLRM assumptions as follows:
  - A1, Linearity:

$$\boxed{y_i = x_i'\beta + \epsilon_i, \quad i = 1, ..., N}$$

  - A2-A5, Mean-zero Error, Homoskedasticity, No Correlation, & Normality:

$$\boxed{\epsilon_i | x_i \sim N(0, \sigma^2), \quad Cov(\epsilon_i, \epsilon_j) = 0, \quad i, j = 1, ..., N, \quad i \neq j}$$

CLRM Assumptions Summary II

- Using Matrix Notation, we can summarize our five CLRM assumptions as follows:
  - A1, Linearity:

$$\boxed{y = X\beta + \epsilon}$$

  - A2-A5, Mean-zero Error, Homoskedasticity, No Correlation, & Normality:

$$\boxed{\epsilon | X \sim N(0, \sigma^2 I_N)}$$

Properties of OLS

Latex Color = "LightGrey"

What's Next?

- Now that we've introduced the five CLRM assumptions, let me do three different things:
  - Use them to derive the nice properties of OLS
  - Describe at a high level why they are valuable
  - Briefly ask the three questions you should ask for each
    1. How likely violated?
    2. What consequences if violated?
    3. How to fix if violated?

OLS Properties: Intro

- We will focus on three properties of the OLS estimator
  1. Unbiasedness
     - ⋆ i.e. Is $E(\hat{\beta}) = \beta$?
  2. Efficiency
     - ⋆ i.e. What is $V(\hat{\beta})$ and how does it compare with $V(\tilde{\beta})$ for another estimator, $\tilde{\beta}$?
  3. Its sampling distribution
     - ⋆ i.e. what is the distribution of $\hat{\beta}$
     - ⋆ (Needed to do hypothesis testing)

OLS Properties: Expected Value I

- We first establish a *very* useful relationship between $\hat{\beta}$, β, and ϵ:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\beta} &= (X'X)^{-1}X'(X\beta + \epsilon) \quad \text{by (A1, Linearity)} \end{aligned}$$

$$\boxed{\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon}$$

- Or, sometimes also like this:

$$\hat{\beta} - \beta = (X'X)^{-1}X'\epsilon$$

OLS Properties: Expected Value II

- Now we take the expectation of $\hat{\beta}$ to evaluate biasedness
  - Note: When we take expectations in this course, they will always be *conditional expectations*
    - ⋆ i.e., when I write $E(\hat{\beta})$, this is implicitly *always* $E(\hat{\beta}|X)$
- Let's do it:*
- **Bottom Line:** the OLS estimator, $\hat{\beta}$, is unbiased

OLS Properties: Variance I

- We turn next to the variance of $\hat{\beta}$:*

OLS Properties: Variance II

$$\boxed{V(\hat{\beta}) = \sigma^2(X'X)^{-1}}$$

- What about some intuition???
- It turns out there are three factors that influence $V(\hat{\beta})$
  - To see them, it is useful to re-write slightly the formula above:

$$\begin{aligned} V(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\ &= \frac{1}{N-1}\sigma^2\left(\frac{1}{N-1}X'X\right)^{-1} \end{aligned}$$

where $\frac{1}{N-1}X'X$ is the "sample variance-covariance matrix" of X

Aside: The sample variance-covariance matrix?

$$\boxed{\frac{1}{N-1}X'X}$$

- Q: What is a sample variance-covariance *matrix*?
  - A: the matrix analog to s_x^2 and s_{xy} ...
    - ★ ...for *each* element/pair of the K elements of X:

$$\frac{1}{N-1}X'X = \begin{bmatrix} \frac{1}{N-1}\sum x_{i1}^2 & \frac{1}{N-1}\sum x_{i1}x_{i2} & \cdots & \frac{1}{N-1}\sum x_{i1}x_{iK} \\ \frac{1}{N-1}\sum x_{i1}x_{i2} & \frac{1}{N-1}\sum x_{i2}^2 & \cdots & \frac{1}{N-1}\sum x_{i2}x_{iK} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N-1}\sum x_{i1}x_{iK} & \frac{1}{N-1}\sum x_{i2}x_{iK} & \cdots & \frac{1}{N-1}\sum x_{iK}^2 \end{bmatrix}$$

(Ask Matteo to show this to you in the exercise sessions if it's not clear)

$$= \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1K} \\ s_{12} & s_2^2 & \cdots & s_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1K} & s_{2K} & \cdots & s_K^2 \end{bmatrix}$$

OLS Properties: Variance III

$$V(\hat{\beta}) = \frac{1}{N-1}\sigma^2\left(\frac{1}{N-1}X'X\right)^{-1}$$

We can now see the three factors that influence $V(\hat{\beta})$:

1. Sample Size, $N-1$
   - As $N-1 \uparrow$, $V(\hat{\beta}) \downarrow$
     - ★ (More data is good!)
2. The variance of ϵ_i, σ^2
   - As $\sigma^2 \downarrow$, $V(\hat{\beta}) \downarrow$
     - ★ (The smaller is σ^2, the closer y_i is to $x_i'\beta$ and...
     - ★ ...the closer y_i is to $x_i'\beta$, the easier it is to identify β)

OLS Properties: Variance IV

$$V(\hat{\beta}) = \frac{1}{N-1}\sigma^2\left(\frac{1}{N-1}X'X\right)^{-1}$$

The three factors influencing $V(\hat{\beta})$, cont:

3. The sample variance-covariance matrix, $\frac{1}{N-1}(X'X)$
   - As $\frac{1}{N-1}(X'X) \uparrow$, $V(\hat{\beta}) \downarrow$
     - ★ Q: Why so?

OLS Properties: Variance V

- A: If there isn't much "spread" in the distribution of x_i...
  - i.e. $\frac{1}{N-1}X'X$ is small
- ...it can be hard to know what exactly is the "right" regression line
  - i.e. what is the "right" $\hat{\beta}$ in the left figure below?

- But if $\frac{1}{N-1}X'X$ *is* large (as on the right), we can more easily pin down $\hat{\beta}$
  - i.e. $V(\hat{\beta}) \downarrow$

The role of $\frac{1}{N-1}$ in $\frac{1}{N-1}X'X$ I

- Above I split out the formula for $V(\hat{\beta})$ as follows:

$$V(\hat{\beta}) = \frac{1}{N-1}\sigma^2(\frac{1}{N-1}X'X)^{-1}$$

- I then talked about how:
  - $\frac{1}{N-1}$, *by itself*, implies that as the sample size increases ($N \uparrow$), we get tighter estimates of $\hat{\beta}$ ($V(\hat{\beta}) \downarrow$)
    - ★ That's right: more data reduces the variance of an estimator that relies on averaging
  - But this just raises a question:
    - ★ Why doesn't this also happen with $\frac{1}{N-1}X'X$?

The role of $\frac{1}{N-1}$ in $\frac{1}{N-1}X'X$ II

- The short version is that:
  - $\frac{1}{N-1} \times$ [a constant] $\rightarrow$ 0
    - ★ This is what's happening for $\frac{1}{N-1}$ *by itself*
  - $\frac{1}{N-1} \times$ [something that grows with N] $\overset{(often)}{\rightarrow}$ [a constant]
    - ★ This is what's happening $\frac{1}{N-1}X'X$

- Matteo can show you a simple example of this in the exercise sessions if it's still unclear

The role of $\frac{1}{N-1}$ in $\frac{1}{N-1}X'X$ III

- When we get to Instrumental Variables estimation, we'll rely on large-sample results...
  - ...which will require us to make a number of assumptions like this
- Indeed Assumption 6: Regular X's will be:

$$\frac{1}{N}X'X \rightarrow \Sigma_{xx}$$

  - where Σ_{xx} is a finite, nonsingular $K \times K$ matrix
- The important thing for you:
  - The implications of dividing by N (or, nearly equivalently, dividing by $N-1$) differs substantially if it is
    - ★ By itself versus
    - ★ Attached to a sum

Properties of OLS under the Five Assumptions

OLS Properties Introduction

- OK, we've calculated $E(\hat{\beta})$ and $V(\hat{\beta})$. *Finally:*
  - What are these long-awaited properties of OLS?
- There are two:
  1. The OLS estimator, $\hat{\beta}$, is unbiased and efficient*
  2. The OLS estimator, $\hat{\beta}$, is normally distributed
     - ★ (These correspond to our two primary uses of statistics:...
     - ★ ...estimation and hypothesis testing)

OLS Properties: Gauss-Markov Theorem I

- Under assumptions (A1)-(A4) the **Gauss-Markov Theorem** establishes that the OLS estimator, $\hat{\beta}$, is B L U E
  - (Which is what, exactly?)*
- BLUE stands for
  - .
  - .
  - .
  - .
- What does each of these mean?

OLS Properties: Gauss-Markov Theorem II

What does each of these mean?

- "Best" means that $\hat{\beta}$ has the smallest variance within the class of linear unbiased estimators
  - i.e., $V(\tilde{\beta}) - V(\hat{\beta})$ is a positive semidefinite matrix for any LUE $\tilde{\beta}$.
    - ⋆ (positive semidefiniteness is the matrix analog to "$\geq$")

OLS Properties: Gauss-Markov Theorem III

What does each of these mean, cont:

- "Linear" $\equiv$ a linear function of y_i
  - i.e., $\hat{\beta} = \sum w_i y_i$, where
    - ★ $w_i \equiv \left(\sum x_i x_i'\right)^{-1} x_i$ is a* vector of weights
- "U" and "E" are self-evident

OLS Properties: Normality I

- In addition, under assumption (A5), $\hat{\beta}$ is *normally distributed*

$$\boxed{\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})}$$

- How do we know?
  - Well, we've already shown that
    - $E(\hat{\beta}) = \beta$
    - $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- So we only need to show that $\hat{\beta}$ is normally distributed

OLS Properties: Normality II

- But this is easy!
- We've already established that...
  1. Linear combinations of normally distributed random variables are also normally distributed
     - ★ (Matteo covered this in the Probability review)
  2. $\hat{\beta}$ is a linear combination of ϵ_i's
     - ★ Each of which is normally distributed under (A5)

OLS Properties: Normality III

- When did we establish that $\hat{\beta}$ is a linear combination of ϵ_i's?
- Well, you may not have realized it at the time, but we did show that

$$\begin{aligned} \hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon \\ &= \beta + (\sum x_i x_i')^{-1} \sum x_i \epsilon_i \end{aligned}$$

  - (...which is a linear combination of ϵ_i's)

OLS Properties Conclusion

- So that's us done! *Under (A1)-(A5)*:
  - The Gauss-Markov Theorem says $\hat{\beta}$ is BLUE
    - ★ Thus it has nice properties as an estimator
  - $\hat{\beta}$ is Normally distributed
    - ★ Thus we can test hypotheses about its elements

What value the assumptions?

Latex Color = "LightGrey"

What value the assumptions? I

- Pwah, that was a fair bit of heavy lifting
- You now know the five key assumptions underlying the CLRM
  - But have I been clear *why* we bother making these assumptions?
- I've said a few times "OLS has nice properties under the CLRM assumptions"
  - What does this mean???

What value the assumptions? II

- Please don't forget that there are two complementary views of $\hat{\beta}$ from an OLS regression
  1. As a way to calculate (a linear approximation of) $E(y_i|x_i)$
     - ★ A "statistical" interpretation
     - ★ (i.e. Correlation)
  2. As a way to estimate the effect of a change of x_i on y_i
     - ★ An "economic" interpretation (IMO)
     - ★ (i.e. Causation)

What value the assumptions? III

- The first (statistical) interpretation is *always valid*
- *Without any assumptions*, we can write the population regression function as

$$y_i = E(y_i|x_i) + \epsilon_i$$

- And if we're willing to assume linearity, we can write it as

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \epsilon_i \\ &= x_i'\beta + \epsilon_i \end{aligned}$$

What value the assumptions? IV

$$\boxed{y_i = x_i'\beta + \epsilon_i}$$

- We said that when $K = 2$, we can write the OLS slope coefficient as

$$\boxed{\hat{\beta}_2 = r_{xy}\frac{s_y}{s_x}}$$

  - So OLS *always* measures the correlation between x_i and y_i
    - ★ (And for $K > 2$, $\hat{\beta}_k$ measures partial correlations between x_{ik} and y_i)

- *Always*:
  - Statistics allows us to recover correlations in data

What value the assumptions? V

- The CLRM assumptions allow us to interpret those correlations as causal effects
  - In particular Assumption 2, the Mean-zero Error assumption...

$$\boxed{E(\epsilon_i|x_i) = 0}$$

  - Ensures we have an unbiased estimate of β
    - ★ (Which is the causal effect of interest)

What value the assumptions? VI

- In my opinion, (A2, Mean-zero error) is an *economic* assumption
  - Or perhaps better: a statistical assumption with economic foundations
    - ⋆ It is a restriction on the true data generating process that ensures that the regression we estimate permits a causal interpretation of $\hat{\beta}$
- As we'll see, there are *many* DGPs that do *not* permit such an interpretation
  - Which will require us to work much harder to make causal statements using the $\hat{\beta}$ we are able to estimate

What value the assumptions? VII

- As I've said repeatedly, the desire and ability to make causal statements from data is perhaps *the* primary use of econometric methods.
  - And this is only possible using OLS if the CLRM assumptions hold
- [Thus they are pretty important. :-)]
- But...

Assumptions Evaluation

Latex Color = "LightGrey"

See Figures in Class

Assumptions Evaluation: Intro I

- I've said a few times now the order of things with the CLRM assumptions is:
  1. Establish nice properties of OLS under the assumptions
  2. Evaluate the assumptions
- We've now done the first...
  - And it's time to give you an *overview* of the second
    - ★ (Digging into each assumption actually takes lots of time)
    - ★ (i.e. hours, not minutes)
    - ★ (But I want to at least give you an overview here)

Assumptions Evaluation: Intro II

- I have said there are three key questions you should ask for each (any) assumption:
  1. Is it *likely to be violated*?
  2. *What are the consequences* if it is violated?
  3. *How do you fix things* if it is violated?
- I'm now going to give you a flavor of the answer for each of our five assumptions

Assumptions Evaluation: Linearity I

$$\boxed{\text{(A1)} \quad y_i = x_i'\beta + \epsilon_i, \quad i = 1, ..., N}$$

1. Is it likely to be violated?
   - Almost surely - rarely is the world truly linear
2. What are the consequences if it is violated?
   - We will estimate the *approximate* effect of x_{ik} on y_i
     - ★ As a linear function is a "first-order approximation of any nonlinear function"
     - ★ (i.e. if approximating a function with a polynomial of degree n...
     - ★ ...we truncate the Taylor series expansion to this degree)
     - ★ (And a polynomial of degree 1 is a linear function)

Assumptions Evaluation: Linearity II

- Q: How bad is it to approximate a potentially nonlinear function with a linear one?
- A: That depends
  - It depends on how you are using your econometric model
  - Many (most) econometric models seek to understand "big-picture" causal effects
    - ★ e.g., what is the benefit to a student of an extra year of education? What is the effect of a training program on future wages?
  - For such questions, a linear approximation is probably fine
  - For other questions where more detailed answers are needed...
    - ★ ...and the data is rich enough to provide an answer...
    - ★ ...digging deeper into the nonlinear structure of the relationship between y_i and x_{ik} may be appropriate

Assumptions Evaluation: Linearity III

3. How do we fix it if violated?
   - This is straightforward
   - One can estimate econometric models with all kinds of nonlinearities
   - Nonlinearities in x_{ik}
     - ★ $\log(x_{ik})$, polynomials in x_{ik}, interactions between x_{ij} and x_{ik}, etc.*
   - Nonlinearities in β, e.g. $y_i = e^{x_i'\beta} + \epsilon_i$ or $y_i = f(x_i, \beta) + \epsilon_i$
     - ★ Using Nonlinear Least Squares
   - We won't cover nonlinear models in this course
     - ★ But you can always look up such models in a ("fat") textbook

Assumptions Evaluation: Mean-Zero Error I

$$\boxed{\text{(A2)} \quad E(\epsilon_i|x_i) = 0}$$

1. Is it likely to be violated?
   - It certainly *can be*
   - When violated, we say "x_{ik} is *endogenous*"
     - ★ Or "x_i is correlated with the error term"
     - ★ (Recalling $E(\epsilon_i|x_i) \Rightarrow Cov(\epsilon_i, x_i) = 0 \Rightarrow Corr(\epsilon_i, x_i) = 0$)
   - Common reasons for endogeneity:
     1. Simultaneous causality (x_{ik} causes a change in y_i, but y_i also causes a change in x_{ik})
     2. Correlated unobservables (Some third variable, w_i, causes changes in both x_{ik} and y_i)
     3. Measurement error in x_{ik}

Assumptions Evaluation: Mean-Zero Error II

2. What are the consequences if it is violated?
   - Serious! Trouble!
   - $\hat{\beta}_k$ will be biased: it will not estimate the causal effect of x_{ik} on y_i!
     - ★ Instead it will estimate this *plus* other confounding factors
   - As discussed in the very first deck,
     - ★ (IMHO) *This is the most important assumption in econometrics*
   - Furthermore, it's not an assumption that can easily be *tested*
     - ★ Most arguments about violations - and the solutions to them - are theoretical
     - ★ [An incredible amount of time and energy in "doing econometrics" (well) is thinking about and dealing with this assumption]

Assumptions Evaluation: Mean-Zero Error III

3. How do we fix it if violated?
   - Most of the second half of the course will be focused on methods to fix violations in the mean-zero error assumption
   - We will focus on the two most popular methods:
     1. Instrumental Variables
     2. Panel Data Methods
   - So we leave further discussion until then...

Assumps Eval: Homoskedasticity and No Correlation I

$$\boxed{\text{(A3) \& (A4)} \quad V(\epsilon) = \sigma^2 I_N}$$

1. Are they likely to be violated?
   - When violated, we say
     - ★ "ϵ_i is *heteroskedastic*" (for violations of (A3))
     - ★ "ϵ_i is *correlated across observations*" (for violations of (A4))
   - They certainly *can be*
     - ★ Errors can vary with features of the explanatory variables, x_{ik} (e.g. in the impact of income on consumption)
     - ★ Errors can be correlated across observations (e.g. in "clustering")

Assumps Eval: Homoskedasticity and No Correlation II

2. What are the consequences if they is violated?
   - Not that big a deal
     - ★ $\hat{\beta}$ will still be unbiased (Whew!)
     - ★ Tho standard errors will be wrong (i.e. $V(\hat{\beta})$ will be biased)

Assumps Eval: Homoskedasticity and No Correlation III

3. How do we fix them if violated?
   - Two strategies:
     1. (Easy:) Use the right formula for $V(\hat{\beta})$*
     2. (Harder:) Estimate a more efficient model via Generalized Least Squares (GLS)
   - Most people take strategy (1)
   - Since conceptually straightforward with easy fixes, we won't cover these topics this semester
     - ★ Despite their being a "standard" thing to teach
     - ★ (If you need this in future, look in your nearest fat textbook)

Assumptions Evaluation: Normality I

$$\boxed{\text{(A5)} \quad \epsilon_i \sim N(0, \sigma^2)}$$

1. Is it likely to be violated?
   - When the other assumptions continue to hold?
   - Surprisingly, far less frequently than you'd think!
     - ★ The reason: averages of independent random variables (ϵ_i) tend to be normally distributed...
     - ★ ...no matter what their original distribution (!)
     - ★ This powerful result called a Central Limit Theorem (CLT)

Assumptions Evaluation: Normality II

- Indeed, it turns out CLTs allow us to determine that $\hat{\beta}$ is *approximately* normally distributed...
  - ...making only weak assumptions on ϵ_i
  - (Indeed relying on (A1) and (A2)!)
  - (i.e. assumptions we've *already* made!)
- For this reason, I think of (A5) as the "superfluous assumption"
  - We make it to quickly and easily test hypotheses, but it is ultimately redundant

Assumptions Evaluation: Normality III

3. Consequences if it is violated?
4. How do we fix it if violated?
   - Given the power of Central Limit Theorems...
     - ★ And the consequent belief that $\hat{\beta}$ is almost always normally distributed
     - ★ (At least approximately)
   - These topics just don't come up very often

Assumptions Evaluation: Conclusion I

- What's the "Bottom Line" re: our assumptions???
  - i.e. what is the short version of what you should understand?

Assumptions Evaluation: Conclusion II

- I typically sort the assumptions into three groups
  1. The **key assumptions** are (A1, Linearity) and (A2, Mean-zero Error)
     - ⋆ The most important results in econometrics rely on just these two
     - ⋆ (And even then (A1) isn't *so* important)
  2. The **ancillary assumptions** are (A3, Homoskedasticity) and (A4, No Correlation)
     - ⋆ Easily checked, easily fixed if violated
     - ⋆ i.e. No Big Deal
  3. The superfluous assumption is (A5, Normality)
     - ⋆ Actually implied by (A1) and (A2) as long as the i's are not strongly dependent and N is large

What's Next?

- OK: Assumptions done. What's next?
  - We'll turn next to *interpreting* OLS regressions
    - ★ (Super-important!)
  - We'll talk about hypothesis testing with OLS

Table of Contents