Empirical Methods

Topic 2a:

Specification Errors and Large Sample Theory

From Good News to Bad

Latex Color = “LightGrey”

From Good News to Bad I

- A brief review of what we've covered so far:
  1. Introduction
     - ★ (Top00) Introduction to (my views on) Econometrics
  2. The Classical Linear Regression Model (CLRM)
     - ★ (Top01a) Probabilistic and Statistical Foundations
     - ★ (Top01b) CLRM Assumptions and Properties
     - ★ (Top01c) CLRM Interpretation
     - ★ (Top01d) CLRM Goodness-of-fit and Hypothesis Testing
- Nicht schlecht!

From Good News to Bad II

- This is the “good part” of the class
  - Describing all the good things that can happen when all goes well
- But we must be practical and acknowledge an uncomfortable truth:
  - The CLRM Assumptions may well *not* hold
    - ★ (Which we covered briefly in Top01b, Slides 79-94)

From Good News to Bad III

- For the rest of the course we will focus on the most important of these potential failures
  - Assumption 2: Mean-zero error
- When Assumption 2 is violated, we say that (at least one of the) "x_{ik} is endogenous"
  - When Assumption 2 is violated, $\hat{\beta}$ is *biased*
    - ★ i.e. $E(\hat{\beta}) \neq \beta$
  - We also say it's *inconsistent*
    - ★ The large-sample analog to bias which I will introduce shortly
    - ★ (This is worse, actually)

Specification Errors

Latex Color = “LightGrey”

Specification Errors Intro I

- It is normal to introduce the possibility of bias by talking more generally about "specification errors" of one of two types:
  1. The inclusion of irrelevant variables
     - ★ i.e. what happens if you include a covariate that really *doesn't* belong?
     - ★ (For whom it really is that $\beta_k = 0$)
  2. The omission of relevant variables
     - ★ i.e. what happens if you forget to include a covariate that really *does* belong?

Specification Errors Intro II

- We'll evaluate the consequences of each of these mistakes by looking at the two standard properties of an estimator that we care about
  - Bias
  - Efficiency

Including an Irrelevant Variable I

- Let’s consider first the consequences of including an irrelevant variable
  - For convenience, I’ll show this in the simple regression model with a single irrelevant variable
    - ...but the conclusions apply more generally to multiple regression with multiple irrelevant variables
    - (The math is just harder and not worth the extra effort)

Including an Irrelevant Variable II

- Suppose the world is such that:

$$\begin{aligned}
\text{Truth:} \quad y_i &= \beta_1 + \beta_2 x_{2i} + \epsilon_i \\
&= x_i' \beta + \epsilon_i \\
y &= X\beta + \epsilon \\
\\
\text{You estimate:} \quad y_i &= \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \\
&= \tilde{x}_i' \beta + \epsilon_i \\
y &= \tilde{X}\beta + \epsilon
\end{aligned}$$

  - X is the $N \times 2$ matrix given by $X = [\iota \;\; X_2]$
  - $\tilde{X}$ is the $N \times 3$ matrix given by $\tilde{X} = [\iota \;\; X_2 \;\; X_3]$
  - $\tilde{X}_{-2}$ is the $N \times 2$ matrix given by $\tilde{X}_{-2} = [\iota \;\; X_3]$
    - i.e. it is the $\tilde{X}$ matrix omitting X_2
    - (Which we will need momentarily for $\tilde{M}_{-2}$)

Including an Irrelevant Variable III

$$\begin{aligned} \text{Truth:} \quad y_i &= \beta_1 + \beta_2 x_{2i} + \epsilon_i \\ &= x_i' \beta + \epsilon_i \\ y &= X\beta + \epsilon \end{aligned}$$

$$\begin{aligned} \text{You estimate:} \quad y_i &= \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \\ &= \tilde{x}_i' \beta + \epsilon_i \\ y &= \tilde{X}\beta + \epsilon \end{aligned}$$

- The key insight: the *true* $\beta_3 = 0$
  - i.e., x_{3i} is *irrelevant*

Including an Irrelevant Variable IV

- Recall from our earlier slides (Top01c) our formula for $\hat{\beta}_k$ using partitioned regression:

  $$\hat{\beta}_k \quad = \quad (X_k' M_{-k} X_k)^{-1} X_k' M_{-k} y$$

  where $M_{-k} = I_N - X_{-k}(X_{-k}' X_{-k})^{-1} X_{-k}'$ is the "residual maker" for X_{-k}

- We apply this here to ask what about the properties of $\tilde{\beta}_2$
  - The key parameter in the *true* model...
    - ...*estimated* on the model that includes an irrelevant variable, $y = \tilde{X}\beta + \epsilon$
    - i.e. $\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$

Including an Irrelevant Variable V

- For our model with an irrelevant variable

$$\begin{aligned}\tilde{\beta}_2 &= (X_2'\tilde{M}_{-2}X_2)^{-1}X_2'\tilde{M}_{-2}y \\ \text{where} \quad \tilde{M}_{-2} &= I_N - \tilde{X}_{-2}(\tilde{X}'_{-2}\tilde{X}_{-2})^{-1}\tilde{X}'_{-2}\end{aligned}$$

  - i.e. $\tilde{M}_{-2}$ is the residual maker for the matrix $\tilde{X}_{-2} = [\iota \quad X_3]$

- And so we can finally evaluate our two questions:
  - Is $\tilde{\beta}_2$ biased?
  - Is $\tilde{\beta}_2$ efficient?

Including an Irrelevant Variable VI

- Is $\tilde{\beta}_2$ biased?

$$\begin{aligned} E(\tilde{\beta}_2) &= E[(X_2'\tilde{M}_{-2}X_2)^{-1}X_2'\tilde{M}_{-2}y] \\ &= E[(X_2'\tilde{M}_{-2}X_2)^{-1}X_2'\tilde{M}_{-2}(X\beta+\epsilon)] \end{aligned}$$

Plug in true model here

$$\text{Note} \quad \tilde{M}_{-2}X = \tilde{M}_{-2}[\iota \quad X_2] = [0 \quad \tilde{M}_{-2}X_2]$$

Why?

$$\Rightarrow \quad \tilde{M}_{-2}X\beta = \tilde{M}_{-2}X_2\beta_2$$

$$\begin{aligned} \Rightarrow \quad E(\tilde{\beta}_2) &= \beta_2 + E[(X_2'\tilde{M}_{-2}X_2)^{-1}X_2'\tilde{M}_{-2}\epsilon] \\ \Rightarrow \quad &= \beta_2 \end{aligned}$$

  - where the last equality follows because $(X_2'\tilde{M}_{-2}X_2)^{-1}X_2'\tilde{M}_{-2}$ is a function of the elements of $\tilde{X}$...
    - And we still have that our Mean-Zero Error assumptions holds
    - (Which here is $E(\epsilon_i|\tilde{x}_i) = 0$)

- **Bottom line:** Including an irrelevant variable does *not* cause bias
  - Good!

Including an Irrelevant Variable VII

- Is $\tilde{\beta}_2$ efficient?

$$\begin{aligned} \text{In the true model:} \quad V(\hat{\beta}_2) &= \sigma^2(X_2' M_{-2} X_2)^{-1} \\ \text{With an irrelevant variable:} \quad V(\tilde{\beta}_2) &= \sigma^2(X_2' \tilde{M}_{-2} X_2)^{-1} \end{aligned}$$

where $M_{-2} = [\iota]$ and $\tilde{M}_{-2} = [\iota \quad X_3]$

- $\Rightarrow V(\hat{\beta}_2) = V(\tilde{\beta}_2)$ if and only if $M_{-2}X_2 = \tilde{M}_{-2}X_2$
  - Which happens when???*
    - ★ ______________________________
  - Is this likely???*
    - ★ ______________________________

Including an Irrelevant Variable VIII

- Is $\tilde{\beta}_2$ efficient, cont.

$$\begin{aligned} \text{In the true model:} \quad V(\hat{\beta}_2) &= \sigma^2(X_2' M_{-2} X_2)^{-1} \\ \text{With an irrelevant variable:} \quad V(\tilde{\beta}_2) &= \sigma^2(X_2' \tilde{M}_{-2} X_2)^{-1} \end{aligned}$$

- Let's assume (as would be normal) that $M_{-2}X_2 \neq \tilde{M}_{-2}X_2$.
- Then which variance is bigger, $V(\hat{\beta}_2)$ or $V(\tilde{\beta}_2)$???*
  - ____
- Why?*
  - ____
- This might remind you of the discussion we had regarding*
  - ____

Specification Errors Intermission

- OK, we’re halfway:

Table: Consequences of Specification Errors

	Inclusion of Irrelevant Variables	Omission of Relevant Variables
Bias	Unbiased	(???)
Efficiency	Inefficient	(???)

- What of omitting a relevant variable?

Omitting a Relevant Variable I

- Suppose (instead) the world is such that:

$$\text{Truth:} \quad y = X\beta + \gamma q + \epsilon$$

$$\text{You estimate:} \quad y = X\beta + \epsilon$$

where

- The normal CLRM assumptions hold
  - ★ Including (A2: Mean-zero error)
  - ★ (Implying $E(\epsilon|X, q) = 0$))
- X is a $N \times K$ matrix
- q is a $N \times 1$ (omitted relevant) variable

Omitting a Relevant Variable II

$$\begin{aligned} \text{Truth:} \quad y &= X\beta + \gamma q + \epsilon \\ \text{You estimate:} \quad y &= X\beta + \epsilon \end{aligned}$$

- The key insight:
  - q is mistakenly omitted
- Let's evaluate our two questions:
  - Is $\hat{\beta}$ biased?
  - Is $\hat{\beta}$ efficient?

Omitting a Relevant Variable III

- Is $\hat{\beta}$ biased?

$$\begin{aligned}
E(\hat{\beta}) &= E[(X'X)^{-1}X'y] \\
&= E[(X'X)^{-1}X'(X\beta + \gamma q + \epsilon)] && \text{Plug in true model here} \\
&= \beta + E[\gamma(X'X)^{-1}X'q + (X'X)^{-1}X'\epsilon] \\
&= \beta + \gamma(X'X)^{-1}X'q && \text{Under (A2: Mean-zero error)} \\
&= \beta + \gamma\hat{\beta}_{q\_on\_X}
\end{aligned}$$

where $\hat{\beta}_{q\_on\_X} = (X'X)^{-1}X'q$ is...

  - ...the OLS coefficient (hence $\hat{\beta}$) from...
  - ... the regression of q on X (hence $_{q\_on\_X}$)

- **Bottom line:** Omitting a relevant variable causes *bias*
  - Uh oh!

Omitting a Relevant Variable IV

$$
\boxed{
\begin{array}{rrcl}
\text{Truth:} & y & = & X\beta + \gamma q + \epsilon \\
\text{You estimate:} & y & = & X\beta + \epsilon \\
\Rightarrow & E(\hat{\beta}) & = & \beta + \gamma (X'X)^{-1}X'q \\
 & & = & \beta + \gamma \hat{\beta}_{q\_on\_X}
\end{array}
}
$$

- Each of the last two lines is called the *formula for omitted variable bias*
  - You sometimes see it as the second-to-last line
  - You sometimes see it as the last line

Omitting a Relevant Variable V

$$\boxed{\begin{aligned} E(\hat{\beta}) &= \beta + \gamma(X'X)^{-1}X'q \\ &= \beta + \gamma\hat{\beta}_{q\_on\_X} \end{aligned}}$$

Two factors determine the presence and sign of any bias:

1. γ
   - Measuring the impact of the omitted variable (q) on the dependent variable (y)
2. $\hat{\beta}_{q\_on\_X} = (X'X)^{-1}X'q$
   - Measuring the correlation between each x_{ik} and q_i
     - ★ (Recall the intuition for $\hat{\beta}$ from Top01b, Slides 20-21)

Omitting a Relevant Variable VI

$$\boxed{\begin{aligned} E(\hat{\beta}) &= \beta + \gamma(X'X)^{-1}X'q \\ &= \beta + \gamma\hat{\beta}_{q\_on\_X} \end{aligned}}$$

Implications of the omitted variable bias formula:

- There is no omitted variable bias if one of two conditions holds:
  - $\gamma = 0$
    - ★ Uninteresting: in this case, there is no omitted variable!
  - $X'q = 0$
    - ★ Unlikely: that *each* of the elements of X is uncorrelated with q

Omitting a Relevant Variable VII

Implications, cont:

- Suppose *only one* element of X is correlated with q, e.g.
  - $X_K'q = \sigma_{Kq}, \;\; X_k'q = 0, \forall k = 1, ..., K-1$
- Quick Quiz: *True, False, or Uncertain: Only* $\hat{\beta}_k$ is biased. The other elements of $\hat{\beta}_k$ are unbiased.* ▸ See answer and derivation in class

Omitting a Relevant Variable VIII

Implications, cont:

- Can we *sign* the bias *on an individual coefficient*?
  - Recall (again!) our formula for $\hat{\beta}_k$ using partitioned regression:

$$\begin{aligned}
\hat{\beta}_k &= (X_k' M_{-k} X_k)^{-1} X_k' M_{-k} y \\
E(\hat{\beta}_k) &= E[(X_k' M_{-k} X_k)^{-1} X_k' M_{-k}(X\beta + \gamma q + \epsilon)] && \text{Plug in true model here} \\
\text{Note} \quad M_{-k}X &= M_{-k}[X_{-k} \;\; X_k] = [0 \;\; M_{-k}X_k] && \text{(Just like a few slides ago)} \\
\Rightarrow \quad E(\hat{\beta}_k) &= E[(X_k' M_{-k} X_k)^{-1} X_k' M_{-k}(X_k\beta_k + \gamma q + \epsilon)] \\
&= \beta_k + \gamma (X_k' M_{-k} X_k)^{-1} X_k' M_{-k} q
\end{aligned}$$

Omitting a Relevant Variable IX

$$\boxed{E(\hat{\beta}_k) \quad = \quad \beta_k + \gamma(X_k' M_{-k} X_k)^{-1} X_k' M_{-k} q}$$

- As earlier, two things determine the sign of the bias
  1. γ, the impact of the omitted variable on y
  2. $(X_k' M_{-k} X_k)^{-1} X_k' M_{-k} q = (X_k^{*'} X_k^{*})^{-1} X_k^{*'} q^{*}$
     - ★ (where X_k^* and q^* are the residuals from the regression of each variable on X_{-k})
- What is this second term, exactly?*
  - ▶ ____
- Can you easily sign it?*
  - ▶ ____
  - ▶ ____

Omitting a Relevant Variable X

$$\boxed{E(\hat{\beta}_k) \quad = \quad \beta_k + \gamma(X_k'M_{-k}X_k)^{-1}X_k'M_{-k}q}$$

- Suppose we feel that we *can* sign both γ and the conditional correlation between X_k and q, $X_k'M_{-k}q$
- Then we can get a sense of the sign of the bias on a particular parameter of interest:

Table: Sign of Bias due to Omitted Variables

		$X_k'M_{-k}q$	
		Positive	Negative
γ	Positive	Positive	Negative
	Negative	Negative	Positive

Omitting a Relevant Variable XI

$$\boxed{E(\hat{\beta}_k) \quad = \quad \beta_k + \gamma(X_k' M_{-k} X_k)^{-1} X_k' M_{-k} q}$$

- Understanding the formula for omitted variable bias...
  - ...and trying to sign its two key components...
  - ...is an *essential skill* in the (intelligent) application of econometrics

- Don't Forget It!

Omitting a Relevant Variable XII

- So we've just established that with an omitted variable, $\hat{\beta}$ is biased.
- What of our other condition? Is $\hat{\beta}$ efficient?
  - Ummm....
- Who cares???
- When it comes to the "flaws" of an estimator...
  - ...I have lexicographic preferences:

If an estimator is biased, I don't care about efficiency!

Omitting a Relevant Variable XIII

- A quibble with the conclusion: "If an estimator is biased, I don't care about efficiency"
  - A more accurate sentence: "If an estimator is *inconsistent*, I don't care about efficiency"
    - (I do care about the efficiency of consistent but biased estimators)
    - (As there are quite a few of these)
    - (Including the Instrumental Variables estimator)
  - But!
    - Omitted variables cause inconsistency as well as bias
    - I haven't yet introduced consistency, so for now we use the "looser" statement on the last slide

Specification Errors Conclusion

- OK, we're all the way there:
- What are the consequences of the two types of specification errors

	Inclusion of Irrelevant Variables	Omission of Relevant Variables
Bias	Unbiased	Biased
Efficiency	Inefficient	(Who Cares?)

- So...
  - ...when in doubt:

(Far) Better to include an irrelevant variable than omit a relevant one!

Large-sample Theory

Latex Color = “LightGrey”

Large-Sample Properties I

- Everything summarized to this point emphasized the “small-sample” properties of OLS
- This is a useful way to teach introductory econometrics, but...
  - ...many econometric estimators - including the Instrumental Variables estimator we’re going to cover next - rely more on “large-sample” properties
    - $\equiv$ the properties of the estimator as the sample gets large
    - (Usually written as $N \to \infty$)

Large-Sample Properties II

- The two primary large-sample properties we rely on are:
  1. Consistency of the OLS estimator, $\hat{\beta}$
     - ★ Analogous to unbiasedness of $\hat{\beta}$
  2. Asymptotic normality of $\hat{\beta}$
     - ★ Analogous to normality of $\hat{\beta}$
- Before we do, however, we introduce consistency and asymptotic normality for a general estimator of β in the population regression function

$$y \quad = X\beta + \epsilon$$

  - (Which we'll call $\tilde{\beta}$)

Consistency I

- $\tilde{\beta}$ is a *consistent* estimator of β if, as $N \to \infty$, the sampling distribution of $\tilde{\beta}$ concentrates around β.
    - Equivalently, we say that the "*probability limit* of $\tilde{\beta}$ is β" and write "plim $\tilde{\beta} = \beta$" (or $\tilde{\beta} \xrightarrow{p} \beta$)
    - The statistical theorem that forms the basis for consistency is called a *Law of Large Numbers*
    - A sufficient (but not necessary) set of conditions for consistency is:

$$\lim_{N\to\infty} E(\tilde{\beta}) = \beta \quad \text{and} \quad \lim_{N\to\infty} V(\tilde{\beta}) = 0,$$

        - ★ i.e., $\tilde{\beta}$ is (a) asymptotically unbiased and (b) its variance shrinks to zero as $N \to \infty$.

Consistency II

- What is the intuition for these conditions?
- The first, $\lim_{N\to\infty} E(\tilde{\beta}) = \beta$ is intuitive:
  - As $N \to \infty$, the expected value of $\tilde{\beta}$ should be β
    - ★ ($E(Estimator)$ often won't be a function of N)
    - ★ (Or will be a simple function of N, e.g. $E(s^2_{MLE}) = \frac{N-K}{N}\sigma^2$)
    - ★ (So this condition - for $\tilde{\beta}$ - simplifies to $E(\tilde{\beta}) = \beta$)
    - ★ (i.e. $\tilde{\beta}$ is unbiased)
    - ★ (You can now see why I call consistency the large-sample analog to unbiasedness)

Consistency III

- What of the second condition, $\lim_{N\to\infty} V(\tilde{\beta}) = 0$?
- Why would the variance of an estimator converge to zero?
  - Hmmm.....
  - Does this sound familiar?
  - Didn't we talk about this once?
  - What did we calculate the variance of and show it converged to zero?*
    - ★ ____

Consistency IV

- Let's show this for the OLS estimator
- Recall we showed that under Assumptions 1-4, we could write the variance of $\hat{\beta}$ as

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

  - (Topic 01b, Slide 53)
  - (Actually I used $N-1$ but we can also use N alone)
- And I further described how

$$\frac{1}{N}X'X \rightarrow \text{[a constant]}$$

Consistency V

- Let's formalize the latter with a new assumption, Assumption 6: Regular X's:

$$\lim_{N\to\infty} \frac{1}{N} X'X = E(X'X) \equiv \Sigma_{xx}$$

  - where Σ_{xx} is a finite, nonsingular $K \times K$ matrix
    - ★ (This the large-sample analog of the "No Perfect Multicollinearity" condition)
    - ★ (That we introduced when discussing multicollinearity)

Consistency VI

- Recall that limits of functions are equal to the functions of their limits, e.g.

$$\lim_{N\to\infty} A(N) \times B(N) = \lim_{N\to\infty} A(N) \times \lim_{N\to\infty} B(N)$$

- Let's use that property to see what $V(\hat{\beta})$ converges to as $N \to \infty$

$$\begin{aligned} V(\hat{\beta}) &= \tfrac{1}{N}\sigma^2(\tfrac{1}{N}X'X)^{-1} \\ \Rightarrow \lim_{N\to\infty} V(\hat{\beta}) &= \lim_{N\to\infty} \tfrac{1}{N} \times \lim_{N\to\infty} \sigma^2 \times \lim_{N\to\infty} (\tfrac{1}{N}X'X)^{-1} \\ &= \lim_{N\to\infty} \tfrac{1}{N} \times \sigma^2 \times \Sigma_{xx}^{-1} \\ &= 0 \end{aligned}$$

Consistency VII

- Job Done! Since
  - $E(\hat{\beta}) = \beta$, and
  - $V(\hat{\beta}) \xrightarrow{p} 0$
- ...we've satisfied the sufficient conditions for consistency for the OLS estimator, $\hat{\beta}$:

$$\boxed{\hat{\beta} \xrightarrow{p} \beta}$$

Consistency VIII*

- I briefly mention the mathematical underpinnings of consistency
  - A full treatment would come in a PhD Econometrics sequence
- Let X_i be a sequence of random variables and let $\bar{X}_N$ be the sample mean, $\frac{1}{N}\sum_{i=1}^{N} X_i$
- There are two Laws of Large Numbers that we rely on in econometrics to show consistency:
  1. When the X_i are independent and identically distributed (iid) with $E(X_i) = \mu$, the **Kolmogorov LLN** proves $\bar{X}_N \xrightarrow{p} \mu$
  2. When the X_i are independent but non-identically distributed (inid) with $E(X_i) = \mu_i$ and finite variance, the **Markov LLN** proves $\bar{X}_N \xrightarrow{p} E(\bar{X}_N)$

Consistency IX

- A very useful property of probability limits comes from **Slutsky's Theorem**:
  - Let $\{x_N : N = 1, 2, ...\}$ be a sequence of $K \times 1$ random vectors such that $x_N \xrightarrow{p} c$.
  - Then if $g(\cdot)$ is a function continuous at c, $g(x_N) \xrightarrow{p} g(c)$
    - ⋆ In essence: probability limits pass through nonlinear functions
- This is useful because expectations *don't*:
  - e.g. if plim $\hat{\beta} = \beta$ and $E(\hat{\beta}) = \beta$,

$$\text{plim}\left(\frac{f(\hat{\beta})}{g(\hat{\beta})}\right) = \frac{f(\beta)}{g(\beta)}, \qquad \text{but} \qquad E\left(\frac{f(\hat{\beta})}{g(\hat{\beta})}\right) \neq \frac{f(\beta)}{g(\beta)}$$

Consistency Assumptions I

- When we showed unbiasedness for $\hat{\beta}$, I first introduced our assumptions and then showed the property
  - Here I've put the cart before the horse:
    - ★ Described the property without being careful to share what assumptions are necessary for it
- As mentioned, a proper treatment of large-sample properties comes in a PhD course...
  - ...but I like to still give you a flavor of what's needed

Consistency Assumptions II

- For consistency, we only relied on our first two (key) assumptions:
  - (A1, Linearity) and
  - (A2, Mean-zero error)
- Even the latter is a bit stronger than we need
  - Strictly speaking, we don't need

$$\begin{aligned} E(\epsilon_i|x_i) &= 0 && \text{in summation notation} \\ E(\epsilon|X) &= 0 && \text{in matrix notation} \end{aligned}$$

  - We only need

$$\begin{aligned} Cov(x_i,\epsilon_i) = E(x_i\epsilon_i) &= 0 && \text{in summation notation} \\ E(X'\epsilon) &= 0 && \text{in matrix notation} \end{aligned}$$

Consistency Assumptions III

$$\begin{aligned}\text{Don't need:} \quad E(\epsilon|X) &= 0 \\ \text{Only need:} \quad E(X'\epsilon) &= 0\end{aligned}$$

- In words:
  - We don't need the conditional mean of $\epsilon_i|x_i$ to be zero
  - We only need that the covariance (correlation) between ϵ_i and (each element of) x_i is zero
- I showed when I first introduced the Mean-zero Error assumption, that the first condition implies the second
  - So the second is weaker and we always like to rely on weaker assumptions

Consistency Assumptions IV

$$\begin{aligned} \text{Don't need:} \quad E(\epsilon|X) &= 0 \\ \text{Only need:} \quad E(X'\epsilon) &= 0 \end{aligned}$$

- In practice, I don't make a big deal of the difference between them:
    - The first assumption rules out that the possibility that nonlinear functions of x_i could be useful at predicting the distribution of ϵ_i...
        - ...whereas the second one does not
    - This is sufficiently unlikely that I don't bother about it

Asymptotic normality I

- We just said that a sufficient condition for consistency is that the variance of the estimator converges to zero
- Thus its asymptotic distribution is *degenerate*
  - i.e., it converges to a mass point at β
    - ★ (because $V(\hat{\beta}) \to 0$ as $N \to \infty$ and $E(\hat{\beta}) = \beta$)
- That's not so useful if we want to do hypothesis testing!
  - So what do we do???

Asymptotic normality II

- Recall for consistency that the variance of the estimator converges to $\frac{1}{N}$ times a constant (matrix)
  - e.g., for $V(\hat{\beta})$, we showed that

$$\lim_{N\to\infty} V(\hat{\beta}) \quad = \quad \lim_{N\to\infty} \frac{1}{N} \times \left(\sigma^2 \Sigma_{xx}^{-1}\right)$$

- Easy then: to ensure the distribution of $\hat{\beta}$ converges to a proper distribution...
  - ...just multiply our estimator by $\sqrt{N}$! Why?
    - ★ When we take the variance, we will square it...
    - ★ Thus its variance will be $\frac{N}{N}$ times a constant (matrix)...
    - ★ ...which is just the constant (matrix)!

Asymptotic normality III

- The asymptotic distribution of a consistent estimator $\tilde{\beta}$ is the distribution of $\sqrt{N}(\tilde{\beta} - \beta)$ as $N \rightarrow \infty$
  - Equivalently, we say that the "$\sqrt{N}(\tilde{\beta} - \beta)$ *converges in distribution* to <something>" and write "$\sqrt{N}(\tilde{\beta} - \beta) \xrightarrow{d}$ <something>"
    - In econometrics, estimators typically converge to Normal distributions and we write $\sqrt{N}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \cdot)$
    - (Where we have to define the appropriate variance-covariance matrix depending on the estimator)
  - And that "$\tilde{\beta}$ *is approximately distributed as* <something>" and write $\tilde{\beta} \overset{a}{\sim}$ <something>.
    - In econometrics, we'd typically write $\tilde{\beta} \overset{a}{\sim} N(\beta, \cdot)$

Asymptotic Normality Underpinnings I*

- The statistical theorem that forms the basis for asymptotic normality is called a *Central Limit Theorem*
- As before, let X_i be a sequence of random variables with $E(X_i) = \mu$ and let $\bar{X}_N$ be the sample mean
  - Let $V[\bar{X}_N] = \frac{1}{N^2}\sum_{i=1}^{N}(X_i - \mu)^2$ be the variance of the sample mean and define

$$Z_N = \frac{\bar{X}_N - E[\bar{X}_N]}{\sqrt{V[\bar{X}_N]}}$$

Asymptotic Normality Underpinnings II*

There are two Central Limit Theorems that we rely on in econometrics to show asymptotic normality:

1. When the X_i are iid:
   - **Lindeberg-Levy CLT:** Let $\{X_i\}$ be iid with $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ Then:
   $$Z_N \xrightarrow{d} N[0,1]$$
2. When the X_i are inid:
   - **Liapounov CLT:** Let $\{X_i\}$ be independent with $E[X_i] = \mu_i$ and $V[X_i] = \sigma_i^2$. If $\lim(\sum_{i=1}^N (E[|X_i - \mu_i|^{2+\delta}])/(\sum_{i=1}^N \sigma_i^2)^{(2+\delta)/2} = 0$ for some $\delta > 0$, then Then:
   $$Z_N \xrightarrow{d} N[0,1]$$
     - ★ We usually set $\delta = 1$, thus we require the existence of third-order moments.

Asymptotic Normality Underpinnings III

- We also rely on an important property analogous to Slutsky's Theorem for consistency
  - Call the *Continuous Mapping Theorem*
- Which says:
  - Let $\{x_N\}$ be a sequence of $K \times 1$ random vectors such that $x_N \xrightarrow{d} x$.
  - Then if $g(\cdot)$ is a continuous function, $g(x_N) \xrightarrow{d} g(x)$
    - ★ In essence: if we can find the limiting distribution of x_N, we can find the limiting distribution of *nonlinear functions* of x_N

Asymptotic Normality Underpinnings IV

- We're going to use the continuous mapping therom because the limiting distribution of the OLS estimator, $\hat{\beta}$, depends on two terms:
    1. $\sqrt{N}(\frac{1}{N}\sum x_i\epsilon_i)$
        - ★ (This is the random vector that we will use our CLT for)
        - ★ (Note $\sqrt{N}\frac{1}{N} = \frac{1}{\sqrt{N}}$)
    2. $(\frac{1}{N}\sum x_i x_i')^{-1}$
        - ★ (This term pre-multiplies the first term in the OLS estimator's asymptotic distribution)

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ I

- Under assumptions (A1, Linearity) and (A2, Mean-zero Error)
  - (And the (weak) additional assumption that $V(x_i\epsilon_i)$ is finite)
    - ★ (So that we can satisfy the conditions of our favorite CLT)
- We can show that:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i\epsilon_i \quad \xrightarrow{d} \quad N(E(x_i\epsilon_i), V(x_i\epsilon_i))$$

where

- $E(x_i\epsilon_i) =$__________
- The form of $V(x_i\epsilon_i) = E(\epsilon_i^2 x_i x_i') = E(\epsilon_i^2)E(x_i x_i')$ depends on further assumptions about the variance-covariance matrix of ϵ
- (There are two common cases)

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ II

- Case 1: If we're willing to make our earlier assumptions (A3, Homoskedasticity) and (A4, No Correlation), then

$$\begin{aligned} V(x_i\epsilon_i) &= E(\epsilon_i)^2 E(x_i x_i') \\ &= \sigma^2 E(x_i x_i') && \text{under (A3) and (A4)} \\ &= \sigma^2 \Sigma_{xx} && \text{under (our new) (A6)} \end{aligned}$$

- Thus (in this case):

$$\tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^{N} x_i\epsilon_i \quad \xrightarrow{d} \quad N(0, \sigma^2\Sigma_{xx})$$

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ III

- In this (first) case, we can re-write the fundamental relationship we showed in Topic 01b, Slide 50 as...

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

where

  - We've shifted between matrix and summation notation and
  - We've just derived the asymptotic distribution of the last term

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ IV

$$\boxed{\begin{array}{rcl} \sqrt{N}(\hat{\beta}-\beta) & = & (\frac{1}{N}X'X)^{-1}\frac{1}{\sqrt{N}}\sum x_i\epsilon_i \\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i\epsilon_i & \xrightarrow{d} & N(0,\sigma^2\Sigma_{xx}) \end{array}}$$

- ... and apply the Continuous Mapping Theorem to show that

$$\begin{array}{rcl} \sqrt{N}(\hat{\beta}-\beta) & \xrightarrow{d} & N(0,\sigma^2\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}) \\ & \xrightarrow{d} & N(0,\sigma^2\Sigma_{xx}^{-1}) \end{array}$$

$$\begin{array}{rrcl} \text{which} \Rightarrow & \sqrt{N}(\hat{\beta}-\beta) & \overset{a}{\sim} & N(0,\sigma^2(\frac{1}{N}X'X)^{-1}) \\ & \hat{\beta}-\beta & \overset{a}{\sim} & N(0,\sigma^2\frac{1}{N}(\frac{1}{N}X'X)^{-1}) \\ & \hat{\beta} & \overset{a}{\sim} & N(\beta,\sigma^2(X'X)^{-1}) \end{array}$$

where

- $\frac{1}{N}X'X \xrightarrow{p} \Sigma_{xx}$ (by Assumption 6, Regular X's)
- If $V(A)=\Omega$, then $V(BA)=B\Omega B'$

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ V

$$\boxed{\begin{aligned} \sqrt{N}(\hat{\beta} - \beta) &\overset{d}{\rightarrow} N(0, \sigma^2 \Sigma_{xx}^{-1}) \\ \Leftrightarrow \hat{\beta} &\overset{a}{\sim} N(\beta, \sigma^2 (X'X)^{-1}) \end{aligned}}$$

- *Critically*, if we can estimate an estimator's asymptotic distribution,
  - Then we can say that estimator is "approximately distributed as" the small-sample approximation to that distribution

Does this look familiar???

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ VI

$$\boxed{\begin{array}{rcl} \sqrt{N}(\hat{\beta}-\beta) & \xrightarrow{d} & N(0,\sigma^2\Sigma_{xx}^{-1}) \\ \Leftrightarrow \hat{\beta} & \overset{a}{\sim} & N(\beta,\sigma^2(X'X)^{-1}) \end{array}}$$

- Formally, $\sigma^2\Sigma_{xx}^{-1}$ is called the "asymptotic variance" of $\sqrt{N}(\hat{\beta}-\beta)$
  - And denoted $Avar(\sqrt{N}(\hat{\beta}-\beta))$
- When we treat $\hat{\beta}$ as "approximately normally distributed", we (rather imprecisely) say that
  - $\sigma^2\Sigma_{xx}^{-1}/N$ is the "asymptotic variance" of $\hat{\beta}$
  - (And estimate it with $\sigma^2(X'X)^{-1}$)

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ VII

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i\epsilon_i \quad \xrightarrow{d} \quad N(0, V(x_i\epsilon_i))$$

- I said the form of $V(x_i\epsilon_i)$ depends on further assumptions about the variance-covariance matrix of ϵ
    - We just covered the (easiest) case with (A3, Homoskedasticity) and (A4, No Correlation)
- While we won't cover in detail the relaxation of either of these assumptions, for completeness let me show you the asymptotic distribution of the OLS estimator in the first of these situations
    - If only to familiarize yourself with the "shape of the formula"
        - ★ Which may prove useful when we derive the IV estimator

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ VIII

- When ϵ_i is *heteroskedastic*, i.e. $E(\epsilon_i^2) = \sigma_i^2$

$$\sqrt{N}(\hat{\beta} - \beta) \quad \overset{d}{\to} \quad N(0, \Sigma_{xx}^{-1} E(\sigma_i^2 x_i x_i') \Sigma_{xx}^{-1})$$

$$\begin{aligned} \hat{\beta} \quad & \overset{a}{\sim} \quad N(\beta, (X'X)^{-1} X' diag(e_i^2) X (X'X)^{-1}) \\ & \overset{a}{\sim} \quad N(\beta, (\sum x_i x_i')^{-1} (\sum e_i^2 x_i x_i') (\sum x_i x_i')^{-1}) \end{aligned}$$

where there is no cancellation of the $\sum x_i x_i'$ terms due to the presence of heteroscedasticity

- This is sometimes called the “sandwich formula” for $Avar(\hat{\beta})$

Asymptotic Normality of the OLS estimator, $\hat{\beta}$ IX

- The sandwich formula isn't too bad as there is no correlation in errors across observations
- When ϵ_i exhibits clustering and/or serial correlation...
  - The two most common forms of correlation in errors across i
- ...we must use matrix notation for everything as we can't easily capture the necessarily correlations using summation notation
  - This I'll leave this to your PhD training!
  - (Or you can look it up in a textbook under "Newey-West standard errors" if you need it)
- For the very dedicated, see Hayashi (PhD-level econometrics textbook):
  - Pages 109-113 (OLS), 209 (GMM), 406-07 (GMM with serial correlation)

Large-Sample Properties Conclusions I

- Let's wrap up what we've shown
- Under (A1, Linearity) and (A2, Mean-zero error), we've shown that
  1. $\hat{\beta}$ is consistent:

     $$\boxed{\hat{\beta} \xrightarrow{p} \beta}$$

  2. $\hat{\beta}$ is asymptotically normally distributed:

     $$\sqrt{N}(\hat{\beta} - \beta) \quad \xrightarrow{d} \quad N(0, .)$$

     ★ If we are willing to add (A3, Homoskedasticity) and (A4, No Correlation), then we can tighten the second condition:

     $$\boxed{\begin{aligned} \sqrt{N}(\hat{\beta} - \beta) \quad &\xrightarrow{d} \quad N(0, \sigma^2 \Sigma_{xx}^{-1}) \\ \Leftrightarrow \hat{\beta} \quad &\overset{a}{\sim} \quad N(\beta, \sigma^2 (X'X)^{-1}) \end{aligned}}$$

Large-Sample Properties Conclusions II

Two important final comments:

1. We *didn't* use (A5, Normality)
   - The reason I call it the “superfluous assumption”
2. When we rely on the asymptotic distribution of $\hat{\beta}$, we *won't* make small-sample adjustments by using the t and F distributions for our hypothesis tests
   - Instead we will use z (Normal) and χ^2 tests
     - ★ This *only* matters for calculating critical values
     - ★ (And not very much there anyway)
     - ★ Conceptually all of our hypothesis tests are the same

Table of Contents