EEP/IAS 118 - Introductory Applied Econometrics, Section 4

Pierre Biscaye and Jed Silver

September 2021

Agenda for today

1. Multicollinearity and Omitted Variable Bias (Lecture 7)
2. Confidence Intervals and Hypothesis Testing (Lectures 8 and 9)

Review: Recall we have 5 important assumptions

For multiple linear regression:

1. Linearity in parameters
2. Random sampling
3. No perfect collinearity
4. Zero conditional mean errors
5. Homoskedasticity

We'll review assumptions 3 and 4 today (please refer to the notes for further more detail on each assumption)

Review: Multicollinearity

- **Definition**: Two (or more) variables are said to be perfectly collinear if one variable is a linear combination of the other variable(s) (e.g., $x_2 = ax_1 + b$)
- **Intuition**: In the MLR framework we want to "hold all else constant" to interpret the effect of one variable, say x_1. We can't do this if x_2 is a linear function of x_1, because then moving x_1 also moves x_2!
- **Example**: x_1 a dummy for being in class today, x_2 a dummy for not being in class today; $x_2 = 1 - x_1$.
- **Note**: some correlation between X variables is normal - we only have a problem when there is a *perfect* or near perfect (very high) correlation (collinearity) between X variables
    - Problem with *near* multicollinearity is that the variance of our estimator $\hat{\beta}$ increases greatly (see section notes).

Omitted Variable Bias

- **Assumption MLR4**: $\mathbb{E}[u|x_1 \cdots x_k] = 0$
  - This is necessary to obtain an unbiased estimate of β
  - $\mathbb{E}[\hat{\beta}_1] = \beta_1$
- **This assumption can fail**:
  - If we fail to include a relevant variable (i.e., that explains variation in y) that is also correlated with the included x.
- **Consequence**: Biased estimates
  - $\mathbb{E}[\hat{\beta}_1] \neq \beta_1$
  - Commonly referred to as Omitted Variable Bias (OVB)

Omitted Variable Bias Example

Why does omitting an important variable introduce bias? Let's think about an example of a model of car thefts as a function of financing for job training programs (assume more financing for job training programs in reality leads to fewer car thefts):

$$cartheft = \beta_0 + \beta_1 jobtrainingfinance + u$$

- What is missing from this model?
- Many things, but let's focus on one variable: the level of gang presence in an area

Omitted Variable Bias Example

Why does leaving out gang violence lead to a biased estimate of β_1?

$$cartheft = \beta_0 + \beta_1 jobtrainingfinance + u$$

- Financing for job training programs and the level of gang violence are themselves likely to be correlated. In areas with lots of gangs, the government may allocate more money to job training programs $\Rightarrow cov(jobtraining, gangs) > 0$
- At the same time gangs also lead to more car thefts in an area $\Rightarrow cov(cartheft, gangs) > 0$
- Therefore, if we don't account for gang activity, it might seem like more financing for job training actually causes *more* car thefts. But really we are just picking up the effect of gang activity!
- This implies that our estimator $\hat{\beta}_1$ will be **upward** biased.

Omitted Variable Bias Direction

This chart shows the resulting bias on our included variable x when we omit the variable x_{ov} depending on the covariance between x_{ov} and y and x_{ov} and the included x:

	$Cov(x, x_{ov}) > 0$	$Cov(x, x_{ov}) < 0$
$Cov(y, x_{ov}) > 0$	Upward bias	Downward bias
$Cov(y, x_{ov}) < 0$	Downward bias	Upward bias

Summary: same sign correlations $\rightarrow$ upward bias; different sign correlations $\rightarrow$ downward bias

Omitted Variable Bias

- If you work through the math in your notes you will see that omitting a relevant variable that is correlated with x leads to the following expression:

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\rho$$

  which you can use to sign the bias as well

- $\tilde{\beta}_1$ is the coefficient on X from the "biased" regression
- $\beta_1 =$ coefficient on X from the "unbiased" regression

$\rightarrow$ true relationship between X and Y

- $\beta_2 =$ coefficient on $X_{omitted}$ from the "unbiased" regression

$\rightarrow$ true relationship between $X_{omitted}$ and Y

- ρ is the correlation between X and $X_{omitted}$
- Use the signs of ρ and β_2 to sign the bias

OVB Example - Question Type 1

We ran the following two regressions:

$$\widehat{\ln(wage)} = 1.19 + 0.101educ + 0.011exp$$

$$\widehat{\ln(wage)} = 1.06 + 0.117educ + 0.011exp - 0.25female$$

1. Interpret the coefficients on $educ$ in both regressions
2. In what direction was the coefficient on $educ$ biased due to the exclusion of $female$ from the regression?
3. Discuss the coefficient on $female$
4. Based on 1) the direction of bias and 2) the coefficient on $female$, what does this imply about the covariance between $female$ and $educ$?

OVB Example - Question Type 1

$$\widehat{\ln(wage)} = 1.19 + 0.101educ + 0.011exp$$
$$\widehat{\ln(wage)} = 1.06 + 0.117educ + 0.011exp - 0.25female$$

1. Interpret the coefficients on $educ$ in both regressions
   A one year increase in $educ$ leads to a predicted 10.1% (11.7%) increase in wages
2. In what direction was the coefficient on $educ$ biased?
   0.101 - 0.117 = -0.016, so we have downward bias
3. What does this imply about the covariance between $female$ and $educ$?
   - $cov(female, wage) < 0$
   - Downward bias from excluding $female$
   - $(-) = cov(fem, educ) * (-) \Rightarrow cov(fem, educ) > 0$

OVB Example - Question Type 2

Anderson (2008) examines whether state "primary" seat belt laws (e.g., cops can pull you over just for not wearing your seat belt) reduce traffic fatalities. Suppose we run this regression of fatalities on population and the presence of the law:

$$\widehat{fatalities} = \hat{\beta}_0 + \hat{\beta}_1 pop + \hat{\beta}_2 primary$$

$$\widehat{fatalities} = 156.002 + 0.1232 pop + 17.258 primary$$

1. If we were naive (i.e. weren't concerned about OVB), how would we interpret this regression?
2. Identify a possible important omitted variable
3. Sign the bias this omission would cause on $\hat{\beta}_{primary}$

OVB Example - Question Type 2

$$\widehat{fatalities} = 156.002 + 0.1232pop + 17.258primary$$

1. If we were naive (i.e. weren't concerned about OVB), how would we interpret this regression?
   Having a primary seatbelt law acutally **increases** *traffic fatalities! This surprising result should tip us off that OVB is a possible problem*
2. Identify a possible important omitted variable
   State speed limit is an example. States with high speed limits are more likely to pass a primary seatbelt law.
3. Sign the bias this omission would cause on $\hat{\beta}_{primary}$
   We know:
   - $cov(speed, primary) > 0$
   - $cov(speed, fatalities) > 0$
   - $\Rightarrow$ upward bias

Switching gears: Confidence Intervals & Hypothesis Testing!

- We are going to be transitioning from point estimates (and bias!) to hypothesis testing
- Why? Because now we want to use statistics to tell us how much confidence we have that our results aren't just random noise.
- A point estimate from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. So while OLS gives us the best possible fit of our model to our sample data, we would like to know how close the estimate is likely to be to the true population parameter.
- Today, we discuss the notion of confidence intervals.

Statistics Reminders

We have a random sample $X_1 \cdots X_n$ for a variable X

Population parameters	μ	$\sum_{j=1}^{k} x_j f(x_j)$
	σ_X^2	$E[(X - E(X))^2]$
	σ_X	$\sqrt{E[(X - E(X))^2]}$
Sample estimators	$\bar{X}$	$\frac{1}{n}\sum_i X_i$
	s_X^2	$\frac{1}{n-1}\sum_i (X_i - \bar{X})^2$
	s_X	$\sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}$
Estimator parameters	$E(\bar{X})$	μ
	$Var(\bar{X})$	$\frac{\sigma_X^2}{n}$
	$Sd(\bar{X})$	$\frac{\sigma_X}{\sqrt{n}}$
SE of estimator	$Se(\bar{X})$	$\frac{s_X}{\sqrt{n}}$

Important Theorem in Statistics

The central limit theorem (CLT) states that the average from a random sample for any population (with finite variance), when standardized, has an asymptotic standard normal **distribution**.

Consider a random sample $X_1, \cdots, X_n$ from a population with mean μ and variance σ^2, then

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma_X / \sqrt{n}} \xrightarrow{d} N(0,1)$$

$\Rightarrow$ If we take many samples and calculate the sample means ($\bar{X}_n$), these will be normally distributed. If we then subtract the true population mean and divide by the estimator standard deviation, the distribution of this new random variable Z_n has PDF that is a standard normal (mean zero, standard deviation one)

Confidence Intervals

Why is this useful?

- Remember that our sample estimator $\bar{X}$ for the true population mean μ is a random variable. Therefore, it would be useful to be able quantify the uncertainty around $\bar{X}$ to help us say something meaningful about what the possible true value of μ might be. To do so we need to know something about how $\bar{X}$ is distributed
- Use our statistics knowledge!
  - The CLT says $\frac{\bar{X}-\mu}{\sigma_X/\sqrt{n}}$ is distributed standard normal
  - We know that for any standard normal variable v, $Pr(-1.96 < v < 1.96) = 95\%$

Confidence Intervals

We can take these two facts to write:

$$Pr(-1.96 < \frac{\bar{X} - \mu}{\sigma_X / \sqrt{n}} < 1.96) = 0.95$$

Rearranging:

$$Pr(\bar{X} - 1.96\frac{\sigma_X}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma_X}{\sqrt{n}}) = 0.95$$

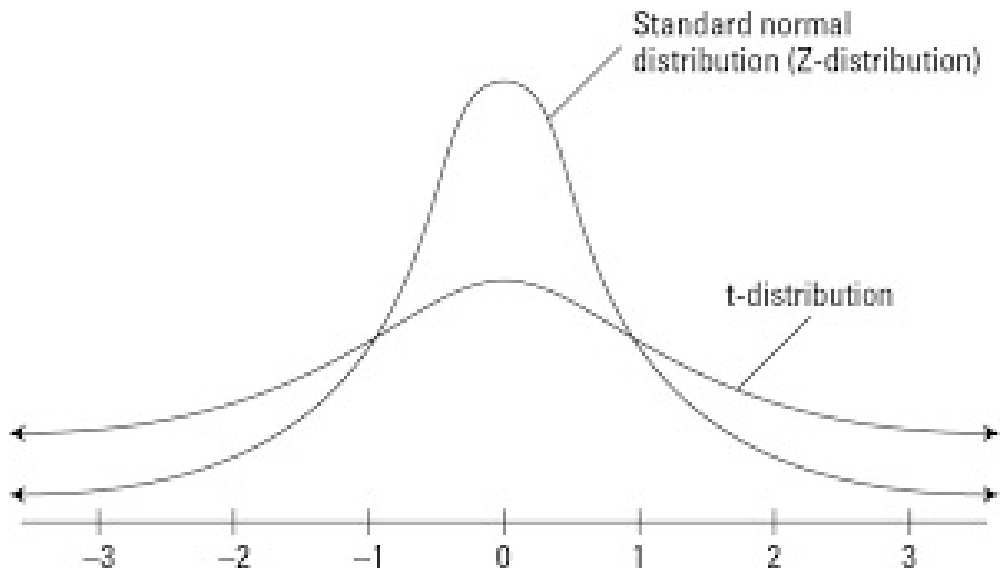
This now is very powerful. This equation indicates that the random range defined by $[\bar{X} - 1.96\frac{\sigma_X}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma_X}{\sqrt{n}}]$ contains the true μ with 95% probability .

- **Note:** the wording here is *very* specific. Look in the notes (or textbook) for a longer discussion.

Confidence Intervals

Remember from last time however that we don't ever observe σ_X. As before, we have to estimate it with s_X.

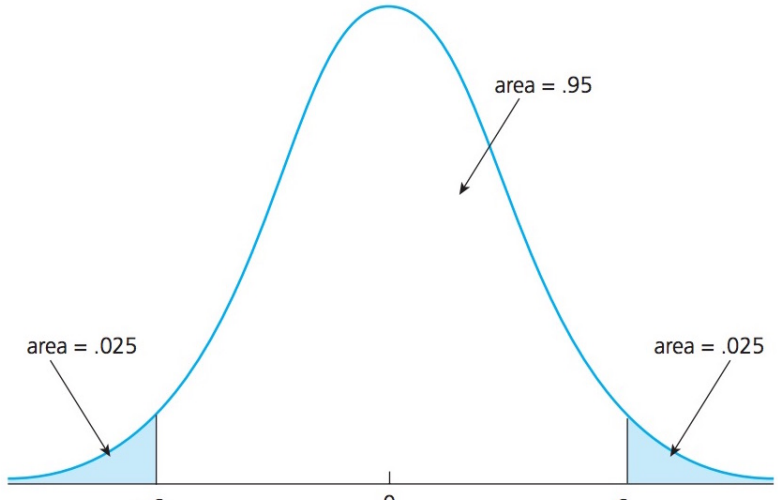
- This costs us something: we lose the normality of the resulting distribution!
- Instead, we have to **use the t-distribution**, which will **widen our confidence interval**:



Constructing Confidence Intervals

We take a random sample of 121 UCB students' heights in inches. Now, to construct a confidence interval for the average height of UCB students:

1. Determine the confidence level - standard is 95%, but 99% and 90% are also used.
2. Compute $\bar{X}$ and s_X from the sample of size n Let's say $\bar{X} = 65$ and $s_X^2 = 4$
3. Find critical value, c, from the t-table. C will depend on the sample size (n) and the confidence level. We look up the value for $n - 1$ degrees of freedom.



t-table

TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

Constructing Confidence Intervals

The formula:

$$CI = \left[\bar{X} - c \cdot \underbrace{\left(\frac{s_x}{\sqrt{n}}\right)}_{se(\bar{X})}, \bar{X} + c\left(\frac{s_x}{\sqrt{n}}\right)\right]$$

From our example:

- c $= 1.98$ (found in t-table for 120 (n-1) degrees of freedom)
- $\bar{X} = 65$
- $s_X = 2$
- $n = 121$

plugging everything in yields:

$$CI = \left[65 - 1.98\left(\frac{2}{\sqrt{121}}\right), 65 + 1.98\left(\frac{2}{\sqrt{121}}\right)\right]$$

Doing the math, the 95% confidence interval is $[64.64, 65.36]$.