Econometrics Discussion Section 2: Omitted Variable Bias

John Green

Spring 2025

Mis-specified model

- Suppose the true model is:
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$
- But we estimate the mis-specified model:
$$Y = \beta_0 + \beta_1 X + u$$
where $u = \beta_2 Z + \epsilon$
- We already know that our estimator $\hat{\beta}_1^{OLS}$ for β_1 will be biased if:

Mis-specified model

- Suppose the true model is:
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$
- But we estimate the mis-specified model:
$$Y = \beta_0 + \beta_1 X + u$$
where $u = \beta_2 Z + \epsilon$
- We already know that our estimator $\hat{\beta}_1^{OLS}$ for β_1 will be biased if X *and* Z *are correlated*: $cov(X, Z) \neq 0$.

Mis-specified model

- This is called *omitted variable bias* (OVB): by failing to include a variable in our model, we get a biased (and inconsistent) estimate of β_1
- The corollary is that we do not have OVB if X and Z are Uncorrelated
  - Important, since there are *always* going to be omitted variables in the error

Analyzing the bias

- We can actually work out the bias in $\hat{\beta}_1^{OLS}$ when we omit Z:

$$\hat{\beta}_1^{OLS} \rightarrow_p \beta_1 + \frac{\sigma_u}{\sigma_X} \times \rho_{X,u}$$

- With a careful analysis, we can start to understand the magnitude and direction of the bias
- Focus on first term: $\dfrac{\sigma_u}{\sigma_X}$
- How does bias change with σ_u and σ_X?

Analyzing the bias

- We can actually work out the bias in $\hat{\beta}_1^{OLS}$ when we omit Z:

$$\hat{\beta}_1^{OLS} \rightarrow_p \beta_1 + \frac{\sigma_u}{\sigma_X} \times \rho_{X,u}$$

- How does bias change with σ_u and σ_X?
    - If σ_u is "large" relative to σ_X, then the bias is large
    - Or, if σ_X is "small" relative to σ_u, then the bias is large (same thing)
- important to keep in mind what u is: **not** just Z, but $\beta_2 Z$.
- What about $\rho_{X,u}$?

Analyzing the bias

- We can actually work out the bias in $\hat{\beta}_1^{OLS}$ when we omit Z:

$$\hat{\beta}_1^{OLS} \rightarrow_p \beta_1 + \frac{\sigma_u}{\sigma_X} \times \rho_{X,u}$$

- important to keep in mind what u is: **not** just Z, but $\beta_2 Z$.
- What about $\rho_{X,u}$?
- Will depend on two things:
  - Relationship between Z and Y: β_2
  - Relationship between Z and X: $cov(X, Z)$
- If they move in different directions, then $\rho_{X,u} < 0$ and the bias is negative; same direction, then $\rho_{X,u} > 0$ and the bias is positive

Example #1

- Regress wages on years of education:

$$w_i = \beta_0 + \beta_1 Educ_i + u_i$$

- What might be omitted?

Example #1

- Regress wages on years of education:

$$w_i = \beta_0 + \beta_1 Educ_i + u_i$$

- Omitted variable is field of study: think about becoming a doctor vs. becoming a nurse, or paralegal vs. lawyer
- So "true" model is $w_i = \beta_0 + \beta_1 Educ_i + \beta_2 major_i + e_i$
- What direction is bias?

Example #1

- Regress wages on years of education:

$$w_i = \beta_0 + \beta_1 Educ_i + u_i$$

- Omitted variable is average time to degree in type of degree: think about becoming a doctor vs. becoming a nurse, or paralegal vs. lawyer
- So "true" model is $w_i = \beta_0 + \beta_1 Educ_i + \beta_2 degree_i + e_i$
- $\beta_2 > 0$ (longer time to degree $\rightarrow$ degree brings higher wage)
- $cov(Educ_i, degree_i) > 0$ (longer time to degree $\rightarrow$ more years of education)
- So bias should be...

Example #1

- Regress wages on years of education:

$$w_i = \beta_0 + \beta_1 Educ_i + u_i$$

- "True" model is $w_i = \beta_0 + \beta_1 Educ_i + \beta_2 degree_i + e_i$
- $\beta_2 > 0$ (longer time to degree $\rightarrow$ degree brings higher wage)
- $cov(Educ_i, degree_i) > 0$ (longer time to degree $\rightarrow$ more years of education)
- So bias should be positive: by ignoring degree type, years of education will appear **more** important than they really are

Example #2

- Regress wages on years of experience:

$$w_i = \beta_0 + \beta_1 exper_i + u_i$$

- What might be omitted?

Example #2

- Regress wages on years of experience:

$$w_i = \beta_0 + \beta_1 exper_i + u_i$$

- Omitted variable is health; poor health leads to lower wages, and also keeps you out of the labor force
- So "true" model is $w_i = \beta_0 + \beta_1 Educ_i + \beta_2 health_i + e_i$
- What direction is bias?

Example #2

- Regress wages on years of experience:

$$w_i = \beta_0 + \beta_1 exper_i + u_i$$

- Omitted variable is health; poor health leads to lower wages, and also keeps you out of the labor force
- "True" model is $w_i = \beta_0 + \beta_1 Educ_i + \beta_2 health_i + e_i$
- $\beta_2 < 0$ (poor health $\rightarrow$ lower wages)
- $cov(exper_i, health_i) < 0$ (poor health $\rightarrow$ less experience)
- So again, bias will be positive; by ignoring health, years of experience will appear **more** important than they really are

Example #3

- Regress wages on GPA:

$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$

- Omitted variables might be...

Example #3

- Regress wages on GPA:
$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$
- Omitted variable could be difficulty of major, d_i
- "True" model is $w_i = \beta_0 + \beta_1 GPA_i + \beta_2 d_i + e_i$
- What direction is bias?

Example #3

- Regress wages on GPA:
$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$
- Omitted variable could be difficulty of major, d_i
- "True" model is $w_i = \beta_0 + \beta_1 GPA_i + \beta_2 d_i + e_i$
- Difficulty leads to lower GPA ($cov(GPA_i, d_i) < 0$), but higher wages ($\beta_2 > 0$)
- So direction of bias is...

Example #3

- Regress wages on GPA:
$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$
- Omitted variable could be difficulty of major, d_i
- "True" model is $w_i = \beta_0 + \beta_1 GPA_i + \beta_2 d_i + e_i$
- Difficulty leads to lower GPA ($cov(GPA_i, d_i) < 0$), but higher wages ($\beta_2 > 0$)
- So direction of bias is negative: GPA will appear *less* important than it really is, since we are ignoring the confounding effect of major.

Example #4

- Regress wages on GPA:

$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$

- Another omitted variable could be use of AI

Example #4

- Regress wages on GPA:
$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$
- Another omitted variable could be use of AI
- "True" model is $w_i = \beta_0 + \beta_1 GPA_i + \beta_2 AI_i + e_i$
- What direction is bias?

Example #4

- Regress wages on GPA:
$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$
- Another omitted variable could be use of AI
- "True" model is $w_i = \beta_0 + \beta_1 GPA_i + \beta_2 AI_i + e_i$
- Using AI leads to higher GPA ($cov(GPA_i, AI_i) > 0$), but lower wages ($\beta_2 < 0$) since not learning the skills
- So direction of bias is...

Example #4

- Regress wages on GPA:
$$w_i = \beta_0 + \beta_1 GPA_i + u_i$$
- Another omitted variable could be use of AI
- "True" model is $w_i = \beta_0 + \beta_1 GPA_i + \beta_2 AI_i + e_i$
- Using AI leads to higher GPA ($cov(GPA_i, AI_i) > 0$), but lower wages ($\beta_2 < 0$) since not learning the skills
- So direction of bias is negative: GPA will appear *less* important than it really is, since we are ignoring the confounding effect of AI.

Solutions

- This seems very troubling; lots of variables out there to omit
- What can we do?

Solutions

- This seems very troubling; lots of variables out there to omit
- What can we do?
- Include it in the regression! (if you can)
- Use some control variable (instrument) which may proxy for the omitted variable
- Research design: diff-in-diff, RCT
- Advanced techniques: fixed effects, matching methods, regression discontinuity or kink; worst case scenario, try to bound the extent of bias