Section 4: Omitted Variable Bias

Valentine Gilbert

September 30, 2021

Overview

Two Ways of Looking at a Linear Equation

Consider the linear equation: $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $X_i \in \{0, 1\}$

- There are two ways of interpreting this equation:
  1. As a **regression equation:** $\beta_1 = cov(Y_i, X_i) / var(X_i) = E[Y_i | X_i = 1] - E[Y_i | X_i = 0]$
     - $E[u_i] = 0$ and $cov(u_i, X_i) = E[u_i X_i] = 0$ *by construction*
  2. As a **causal equation:** $\beta_1 = E[Y_i(1) - Y_i(0)]$ is the average causal effect of X_i
     - Then u_i may be correlated with X_i – The errors “have a life of their own”
     - $u_i \equiv Y_i(0) - E[Y_i(0)]$
- This raises the question: When does a regression estimate a causal model?
- Answer: When the errors from the causal model have the properties of regression residuals

Mean Independence

Suppose we're interested in the causal effect of a binary treatment $X_i \in \{0, 1\}$

- The OLS coefficient on X_i is the difference in means:

$$\begin{aligned}
\beta_1^{OLS} &= E[Y_i|X_i = 1] - E[Y_i|X_i = 0] \\
&= E[\beta_0 + \beta_1 X_i + u_i|X_i = 1] - E[\beta_0 + \beta_1 X_i + u_i|X_i = 0] \\
&= \beta_1 + E[u_i|X_i = 1] - E[u_i|X_i = 0]
\end{aligned}$$

- Therefore, $\beta_1^{OLS} = \beta_1$ iff u_i is mean independent of X_i (Note that u_i is the error from the *causal* linear equation)
- Notice that u_i is mean independent of X_i if $E[u_i|X_i = 1] - E[u_i|X_i = 0] = 0$, and this is equivalent to $E[Y_i(0)|X_i = 1] - E[Y_i(0)|X_i = 0] = 0$
- Regression therefore estimates a causal effect iff potential outcomes are mean independent of the treatment variable

Conditional Mean Independence, Causal and Control Variables

- Often u_i is not mean independent of X_i, but u_i is mean independent of X_i *conditional on* controls, $\boldsymbol{w_i}$
- That is, $E[u_i|X_i = 1, \boldsymbol{w_i}] = E[u_i|X_i = 0, \boldsymbol{w_i}]$
- Then the OLS regression of Y_i on X_i and $\boldsymbol{w_i}$ recovers β_1, the causal effect of X_i
    - Example: In the Tennessee STAR Experiment, students were randomized to classrooms *within* schools. Do you expect the errors were unconditionally mean independent of the treatment?
- Notice the conceptual distinction between X_i and the covariates in $\boldsymbol{w_i}$
- Can we say that the errors are mean independent of $\boldsymbol{w_i}$? Do we care?

OVB 1/4: Motivation

The most important formula in economics.
-Bruich, Chamberlain, and Feldstein

- “Careful reasoning about OVB is an essential part of the ’metrics game” and “The OVB formula is the Prime Directive of applied econometrics” - Angrist and Pischke (2015)
- Questions where we care about OVB:
  - What’s the effect of an additional year of education on earnings?
  - How much does additional health care spending improve health?
  - How do different institutions affect economic development?
- Potential point of confusion: Omitted variable bias would more appropriately be called omitted variable inconsistency
  - The reason we care so much about OVB is that if our estimates suffer from it, they’re inconsistent estimates of our target parameter

OVB 2/4: Intuition

- Consider three regression equations:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i && \textit{(Long regression)} \\
Y_i &= \alpha_0 + \alpha_1 X_i + u_i && \textit{(Short regression)} \\
Z_i &= \gamma_0 + \gamma_1 X_i + v_i && \textit{(Auxiliary regression)}
\end{aligned}
$$

- The OVB formula is: $\alpha_1 = \beta_1 + \beta_2 \gamma_1$
- What's the intuition?
  - If we don't hold Z_i constant (i.e. omit it from our regression), an increase in X_i is accompanied by a γ_1-unit increase in Z_i
  - Every one-unit increase in Z_i is associated with a β_2-unit increase in Y_i (conditional on X_i)
  - Increasing X_i without conditioning on Z_i therefore has a direct effect on Y_i of β_1 and an indirect effect (through Z_i) of $\beta_2 \gamma_1$

OVB 3/4: An Alternative Derivation

- In class we derived the OVB formula using the regression coefficient equations
- An alternative derivation transforms the long regression into the short regression by rewriting the omitted variable using the auxiliary regression:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i + \beta_2 [\gamma_0 + \gamma_1 X_i + v_i] + \varepsilon_i \\ &= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X_i + (\varepsilon_i + \beta_2 v_i) \end{aligned}$$

- Is $\varepsilon_i + \beta_2 v_i$ mean 0?
- Is $\varepsilon_i + \beta_2 v_i$ correlated with X_i?
- $\alpha_0 = \beta_0 + \beta_2 \gamma_0$, $\alpha_1 = \beta_1 + \beta_2 \gamma_1$, $u_i = \varepsilon_i + \beta_2 v_i$

OVB 4/4: Application

- Let's apply this to a problem set question: Does the Holt data on adopted children and their families suggest that **maternal education** has a causal effect on **children's education**?
  - What's the short regression for this question?
  - What would you like to include in the long regression?
  - Can you sign the bias?

- What about for whether **single-parent share** affects **upward income mobility** in a census tract?

- What about for the claim that **moderate drinking** (1-2 drinks a day) reduces **mortality**?

Gelbach 1/7: Motivation

- Let's return to the Opportunity Atlas data
- Upward mobility is strongly related to income, but income is correlated with other variables that influence upward mobility

Table: Short and Long Regression Estimates of the Relationship between Upward Mobility and Mean Household Income in 2000

	(1) Short	(2) Long
Mean income ($000s)	0.133	0.013
	(0.001)	(0.001)
Covariates		
College educated share		Yes
Single parent share		Yes
Average test scores		Yes

Standard errors in parentheses

- How much of the economic mobility gradient is attributable to each covariate?

Gelbach 2/7: The Wrong Approach

- One approach to assessing the contribution of different covariates to the mobility gradient is to sequentially add covariates
- With each additional covariate, note the change in the coefficient of interest

Table: Sensitivity Analysis Sequentially Adding Covariates

	(1)	(2)	(3)	(4)
Mean income	0.133	0.104	0.014	0.013
	(0.001)	(0.001)	(0.001)	(0.001)
Covariates				
College educated share		Yes	Yes	Yes
Single parent share			Yes	Yes
Average test scores				Yes

Standard errors in parentheses

Gelbach 3/7: Problem

- What's the problem with this approach?
- It depends on the *order* in which covariates are added!

Table: Sensitivity Analysis Sequentially Adding Covariates

	(1)	(2)	(3)	(4)
Mean income	0.133	0.112	0.055	0.013
	(0.001)	(0.001)	(0.001)	(0.001)
Covariates				
College educated share				Yes
Single parent share			Yes	Yes
Average test scores		Yes	Yes	Yes

Standard errors in parentheses

Gelbach 4/7: Problem

- So the arbitrary choice of covariate order changes our conclusions about how much of the mobility gradient to attribute to each covariate

Table: Amount of Mobility Gradient Explained by Covariates Under Different Sequences

	Change in slope	
	Sequence 1	Sequence 2
Covariates		
College educated share	-0.029	-0.042
Single parent share	-0.09	-0.057
Average test scores	-0.001	-0.021

Gelbach 5/7: Solution

The Gelbach Decomposition uses the omitted variable bias formula to attribute the mobility gap to different covariates

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i && \textit{long regression} \\
Y_i &= \alpha_0 + \alpha_1 X_{1i} + u_i && \textit{short regression} \\
X_{2i} &= \gamma_{02} + \gamma_{12} X_{1i} + v_{2i} && \\
&\vdots && \textit{auxiliary regressions} \\
X_{ki} &= \gamma_{0k} + \gamma_{1k} X_{1i} + v_{ki} &&
\end{aligned}
$$

- The OVB formula gives us: $\alpha_1 = \beta_1 + \beta_2\gamma_{12} + \beta_3\gamma_{13} + \ldots + \beta_k\gamma_{1k}$
- We can therefore calculate how much of the difference in the short and long regression coefficients to attribute to each of the omitted variables

Gelbach 6/7: Scaling the decomposition

There are three ways you can present the results of the decomposition:

1. **In levels:** $\beta_2\gamma_{12}$ is the income earned by low-income children growing up in higher income tracts that is explained by differences in X_{2i}
2. **As a percent of the gap:** $\beta_2\gamma_{12}/(\alpha_1 - \beta_1)$ is the percent of the difference between the long and short regression coefficients that is explained by X_{2i}
3. **As a percent of the gradient:** $\beta_2\gamma_{12}/\alpha_1$ is the percent of the relationship between own income and average neighborhood income that is explained by X_{2i}

The best way to scale the decomposition depends on the context – use your judgment

Gelbach 7/7: Best Practice

I conclude with one final observation: except in special cases [...] there really is no reason to sequentially add X_2 covariates to a base model. Sequential addition can obscure, overstate, or understate the true part of δ that can be attributed to variation in any given set of X_2 variables. The only meaningful way to estimate the sensitivity of β_1 to covariates is to add all the covariates at once and then compare $\hat{\beta}_1^{base}$ and $\hat{\beta}_1^{full}$. Providing tables with subsets of X_2 added sequentially across columns or rows thus makes little sense, and this practice should simply be abandoned.

-Gelbach (2016)