OMITTED VARIABLE BIAS

Fundamentals of

PROGRAM EVALUATION

JESSE LECY

A NOTE ON TERMS IN THIS SECTION:

$$TestScore = \beta_0 + \beta_1 ClassSize + \beta_2 SES + \beta_3 TeachQuality + \varepsilon$$

"Full Model", i.e. the "truth". **The slopes will be correct** because we have all of the variables included, therefore we use Greek letters.

$$TestScore = b_0 + b_1 ClassSize + \cancel{SES} \quad b_2 TeachQuality + e$$

"Naive Model" - We are missing variables and therefore we **do NOT know if the slopes are correct**. They represent our best guess. They may contain bias. We use Latin characters to denote this.

You might be used to thinking in terms of population statistics and sample. In regressions, you can have the entire population in your sample, but if you are missing variables in your regression then your slopes will be wrong. To map concepts, when I say "full model" think population statistic (the truth), and when I say "naïve model" think sample statistic (the best guess).

THE MAIN QUESTION TO ASK YOURSELF:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$ (full model)

$$Y = b_0 + b_1 X1$$ (naïve model)

Does omitting a variable introduce bias into our estimate of program impact?

$$\beta_1 = b_1 \quad ???$$

If we have an omitted variable, will our estimate of the program impact (b1) sufficiently represent the true program impact (β1)?

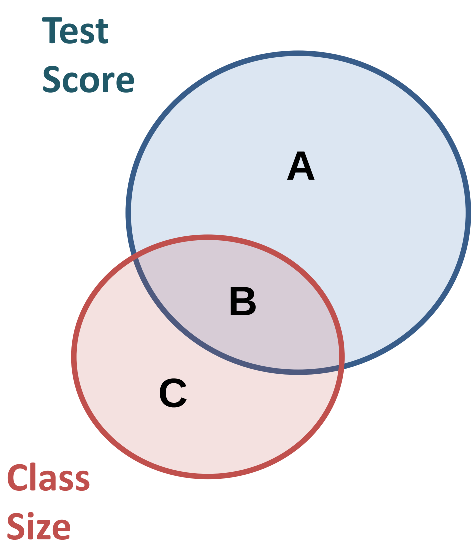
NOTE!

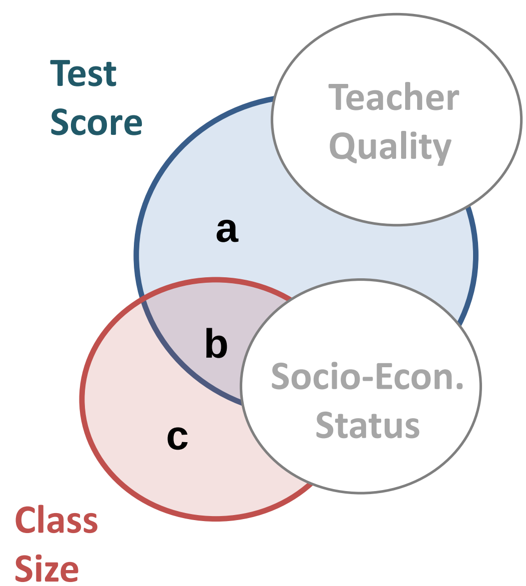
We will ALWAYS have omitted variables in observational studies because we either can't measure the variable we care about, or else it's just not available in the data we have available.

The real question is not whether it is there, but how much will it affect our estimates?

WHAT WE KNOW SO FAR:

We think about control variables as variables that remove variance from our model so we can focus on the policy variable.





$$TS = b_0 + b_1 CS$$

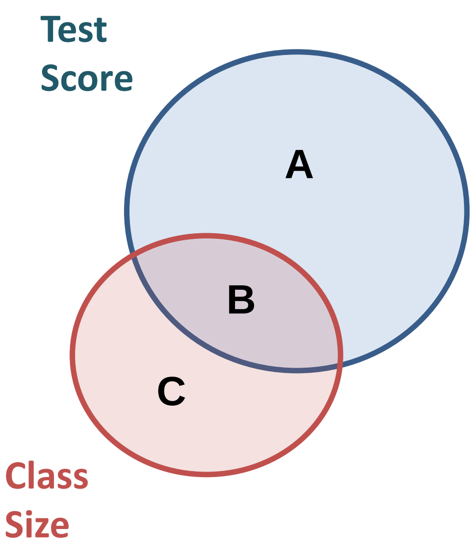
$$TS = \beta_0 + \beta_1 CS + \beta_2 SES + \beta_2 TQ$$

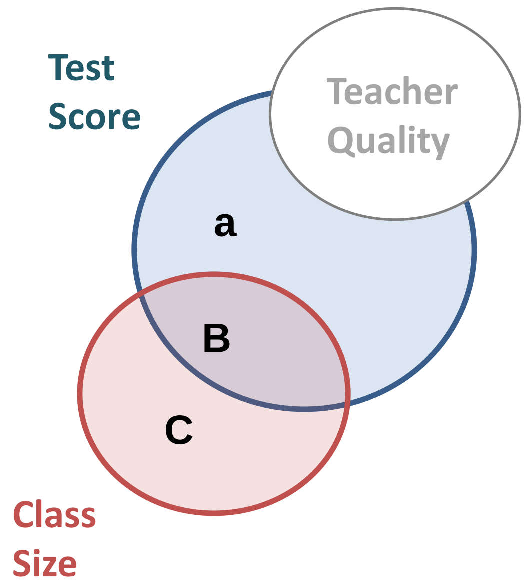
WHAT WE KNOW SO FAR:

$$b_1 = \frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)}$$

$$SE_{b1} = \frac{residual}{sample\,size \cdot \mathrm{var}(x_1)}$$

WHAT WE KNOW SO FAR:



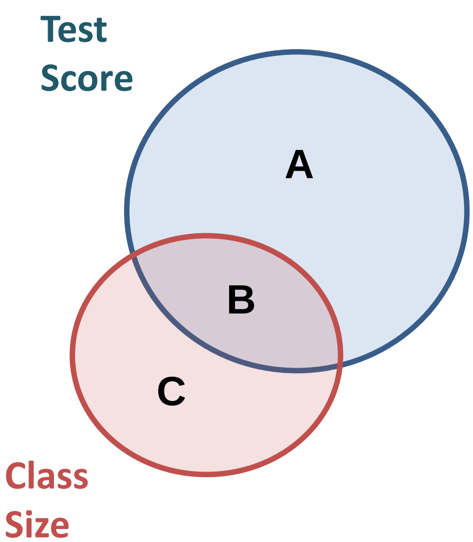


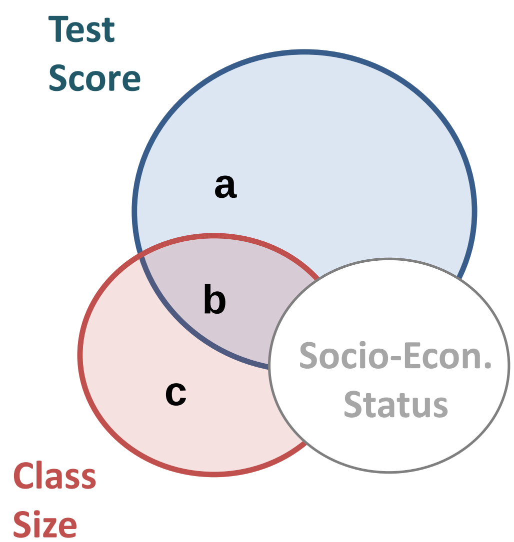
$$slope: \frac{B}{B+C} \rightarrow \frac{B}{B+C}$$

$$SE_{b1}: \frac{A}{B+C} \rightarrow \frac{a}{B+C}$$

When we add a control that is uncorrelated with the policy variable, it explains extra variance of Y but does not affect the policy slope.

WHAT WE KNOW SO FAR:





$$slope: \frac{B}{B+C} \rightarrow \frac{b}{b+c}$$

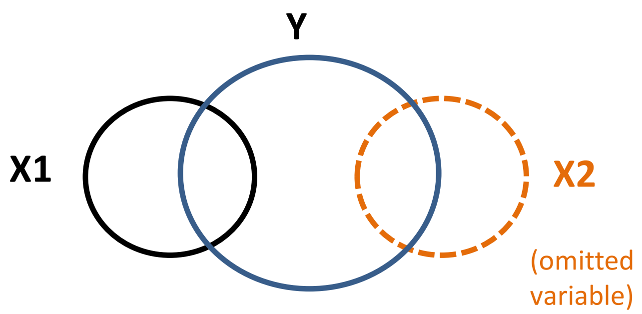
$$SE_{b1}: \frac{A}{B+C} \rightarrow \frac{a}{b+c}$$

When we add a control variable that IS correlated with the policy variable it affects both the slope and the standard error.

OMITTED VARIABLE BIAS:

All that we are doing with omitted variable bias is asking, what happens when we leave the control variable out of the model?

CASE #1



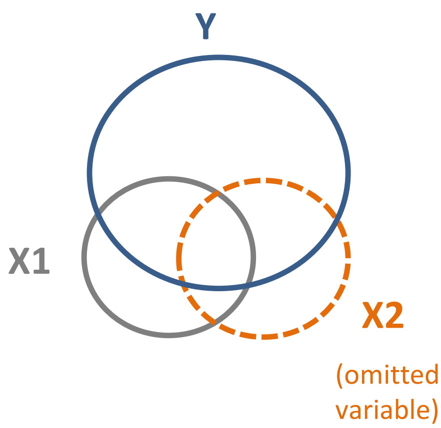
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = b_0 + b_1 X1$$

$$\beta_1 = b_1$$

Since the omitted variable X2 is uncorrelated with the policy variable X1, then leaving it out does not change the slope b1. There is no bias.

CASE #2



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = b_0 + b_1 X1$$

$$\beta_1 \neq b_1$$

In this case, omitting X2 from the model will change the slope b1 because X1 and X2 have shared covariance. Our naïve estimate WILL be biased.

HOW DO OMITTED VARIABLE IMPACT REGRESSION RESULTS?

DV: Test Scores	Model 1 (SES & TQ omitted)	Model 2 (SES omitted)	Model 3 (TQ omitted)	Model 4	Model 5 (Full Model)
Constant	353.192***	38.542***	148.015***	271.628**	-39.322**
	(5.723)	(1.589)	(4.778)	(89.272)	(12.092)
Class Size	-4.376***	-4.549***		-2.642	-2.893***
	(0.205)	(0.028)		(1.905)	(0.256)
Quality of Instruction		64.014***			64.012***
		(0.281)			(0.275)
Socio-Economic Status			45.791***	18.274	17.448***
			(2.153)	(19.960)	(2.687)
R-squared	0.313	0.987	0.312	0.313	0.988
N	1000	1000	1000	1000	1000

Bias is the difference between the “truth” (Model 5 in this case) and what we would get if we ran a naïve regression (Model 1 here).

Note that the bias can be quite large.

$$b_1 = -4.376$$
$$\beta_1 = -2.893$$
$$bias = b_1 - \beta_1 = -1.483$$

$$size\ of\ bias \approx \frac{-1.483}{-2.893} = 51\%$$

We overestimate the impact of our program by 51% !

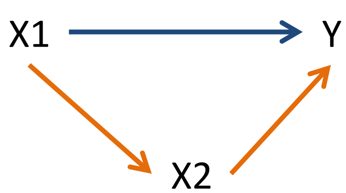
CALCULATING OMITTED VARIABLE BIAS:

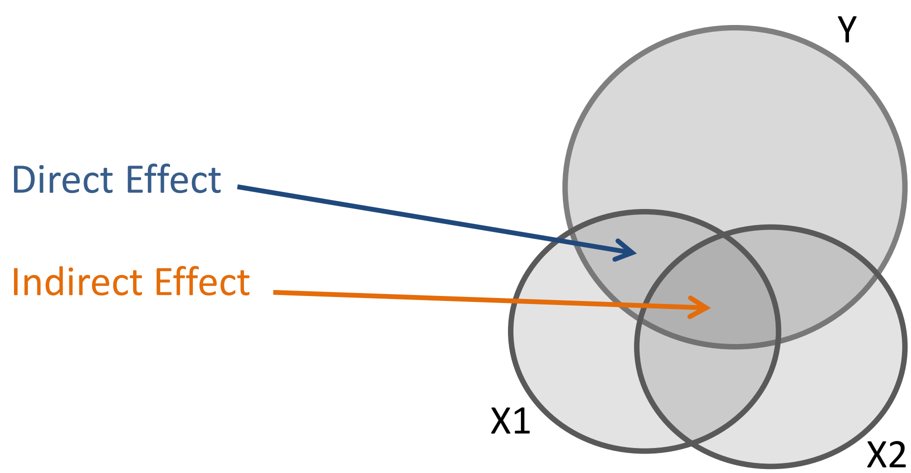
The definition of bias is the difference between the true slope and our best guess of the slope:

$$b_1 = -4.376$$
$$\beta_1 = -2.893$$
$$bias = b_1 - \beta_1 = -1.483$$

Note that this is not very useful in practice because if you know the true slope β1 you will not need to calculate bias!

WHERE BIAS COMES FROM:



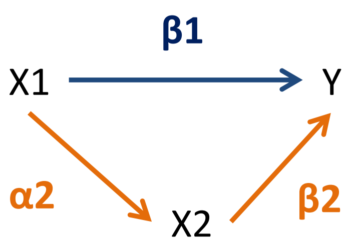


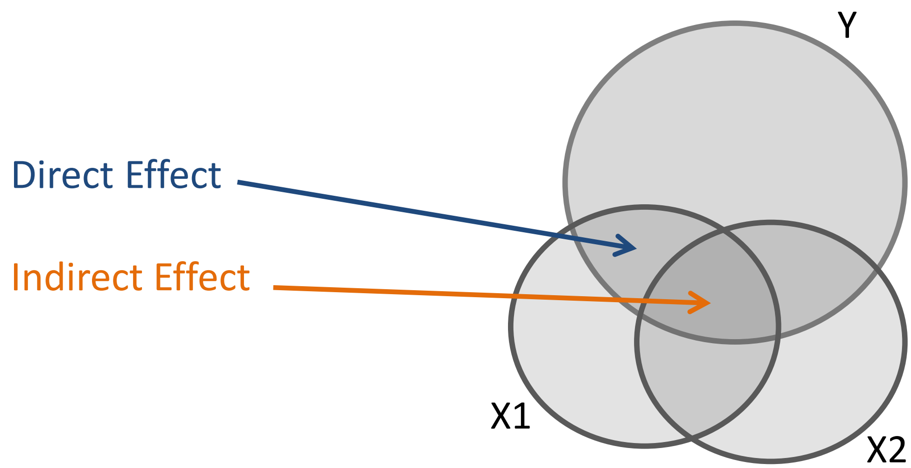
b1 = Direct Effect + Indirect Effect

β1 = Direct Effect

bias = b1 – β1 = Indirect Effect

THE MATH:





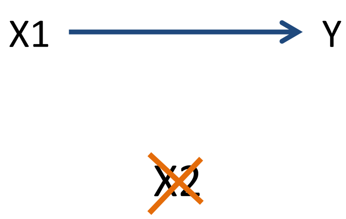
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_1$$ (full regression)

$$X_2 = \alpha_0 + \alpha_1 X_1 + \varepsilon_2$$ (auxiliary regression)

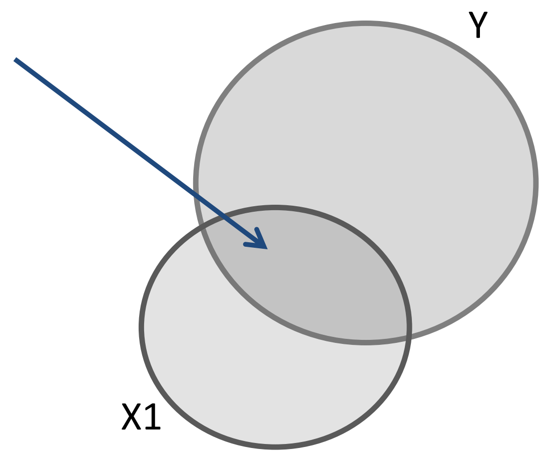
$$bias = \beta_2 \alpha_1$$ (path diagram for X1→X2→Y)

THE MATH:

True Slope plus Bias



$$Y = b_0 + b_1 X1$$

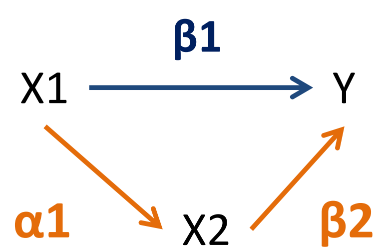


$$b_1 = \beta_1 + bias$$

If we run a naïve model and exclude X2 then the slope b1 will include both the direct and indirect effects.

NOTE:

To run the auxiliary regression, just think about the effects of X1 working through X2, so make sure X2 is on the left hand side of the auxiliary regression.



OMITTED VARIABLE BIAS DERIVED

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

(Don't need to know for the test)

$$\textit{Decomposed}$$

$$Y = b_0 + b_1 X_1 + \varepsilon_1 \qquad (1)$$

$$X_2 = \alpha_0 + \alpha_1 X_1 + \varepsilon_2$$

$$\textit{Substitute for } X_2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \left(\alpha_0 + \alpha_1 X_1 + \varepsilon_2\right) + \varepsilon \qquad (2)$$

$$Y = \beta_0 + \beta_2 \alpha_0 + \beta_1 X_1 + \beta_2 \alpha_1 X_1 + \varepsilon + \beta_2 \varepsilon_2$$

$$Y = \beta_0 + \beta_2 \alpha_0 + \left(\beta_1 + \beta_2 \alpha_1\right) X_1 + \varepsilon + \beta_2 \varepsilon_2$$

$$\textit{bc of the Equivalence of } (1) \textit{ and } (2):$$

$$b_1 X_1 = \left(\beta_1 + \beta_2 \alpha_1\right) X_1$$

$$\beta_1 = b_1 - \beta_2 \alpha_1$$

$$\beta_1 = b_1 - bias \qquad OR \qquad bias = b_1 - \beta_1$$

EXAMPLE OF CALCULATIONS:

$$Y = b_0 + b_1 X_1 + \varepsilon_2$$

(naïve regression)

```
lm(formula = TestScores ~ ClassSize)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 141.67517    4.57143   30.99   <2e-16 ***
ClassSize    -0.43285    0.02059  -21.02   <2e-16 ***
```

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_1$$

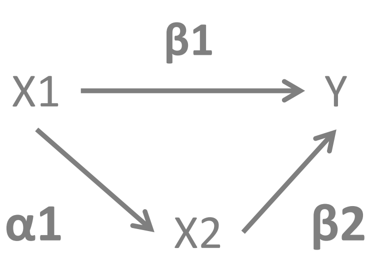
(full regression)

```
lm(formula = TestScores ~ ClassSize + SES)

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 141.5526    4.5959  30.800  <2e-16 ***
ClassSize    -0.3770    0.2071  -1.820   0.069 .
SES           5.6495   20.8534   0.271   0.787
```

$$X_2 = \alpha_0 + \alpha_1 X_1 + \varepsilon_2$$

(auxiliary regression)

```
lm(formula = SES ~ ClassSize)

Coefficients:
              Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  0.02168831 0.00694244    3.124  0.00184 **
ClassSize   -0.00988298 0.00003127 -316.098  < 2e-16 ***
```



$$\beta_1 = b_1 - \beta_2 \alpha_1$$

where

$$\beta_2 \alpha_1 = bias$$

or

$$bias = b_1 - \beta_1$$

$$\beta_2 \alpha_1 = 5.65 \cdot -0.0099 = -.056$$

$$b_1 - \beta_1 = -0.433 - (-0.377) = -0.056$$

THE TAKE-AWAY:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_1$$
$$Y = b_0 + b_1 X1 + e$$
$$X_2 = \alpha_0 + \alpha_1 X_1 + \varepsilon_2$$

(1) $\Rightarrow bias = \beta_2 \alpha_1$

Bias is the product of two slopes: X1→X2 & X2 →Y

(2) $\Rightarrow b_1 = \beta_1 + bias$

The naïve slope is the actual slope plus bias

(3) $slope = \dfrac{\text{cov}(x, y)}{\text{var}(x)}$

The sign of a slope is always determined by the sign of the covariance, i.e. the correlation

WHY DOES THIS MATTER?

$bias = \alpha_1 \beta_2$

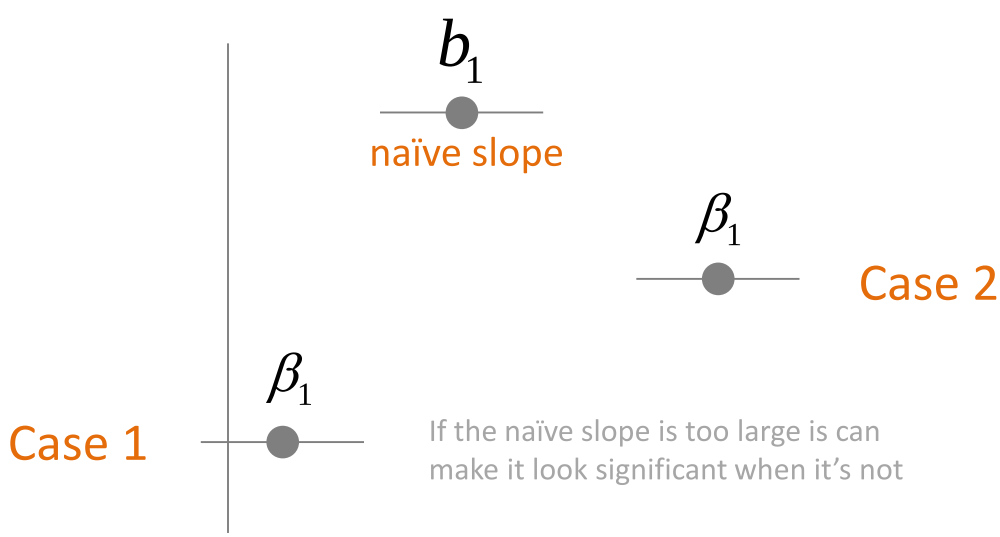
$Where \alpha_1 {\sim} cor(x_1, x_2)$
$And \beta_2 {\sim} cor(x_2, y)$

$b_1 = \beta_1 + bias$

$\therefore$

$If \;\; bias \;\; (+) \;\; then \;\; b_1 > \beta_1$ Case 1: Naïve slope is too large

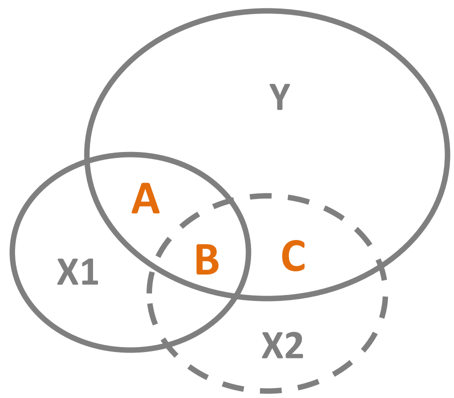
$If \;\; bias \;\; (-) \;\; then \;\; b_1 < \beta_1$ Case 2: Naïve slope is too small



WHEN DOES O.V.B. OCCUR?

CASE 1: OMITTED VARIABLE CORRELATED WITH POLICY VARIABLE

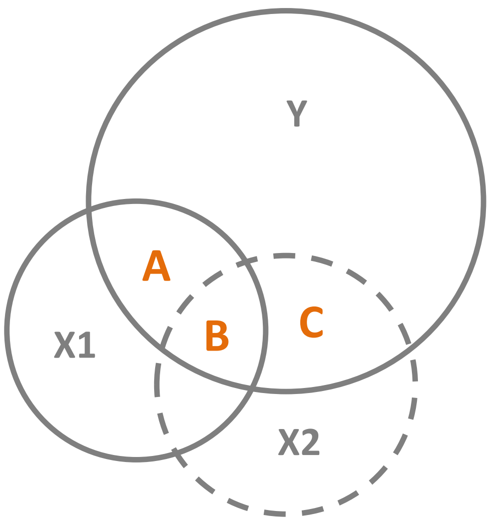
In this case, the omitted variable X2 is correlated with the policy variable X1. There is shared co-variance, represented by the region B. This is the region that is discarded as part of the regression procedure



$$b_1 \approx A + B$$
$$\beta_1 \approx A$$
$$bias \approx B$$
$$\beta_1 = b_1 - bias \approx (A+B) - B$$

The naïve slope, b_1, and the full-model slope, B_1, will now be different because of the exclusion of the region B. The naïve model will be biased as a result of omitting X2.

Case 1: Omitted Variable Correlated with Policy Variable



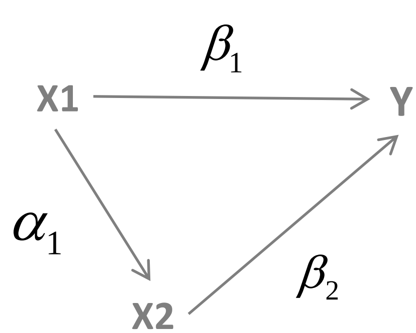
$$b_1 \approx A + B$$

$$\beta_1 \approx A$$

$$bias \approx B$$

$$\beta_1 = b_1 - bias \approx (A + B) - B$$

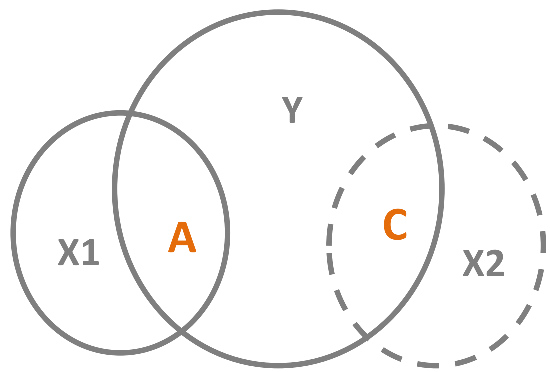
Path Diagram



$$bias = \alpha_1 \beta_2$$

CASE 2: OMITTED VARIABLE UNCORRELATED WITH POLICY VARIABLE

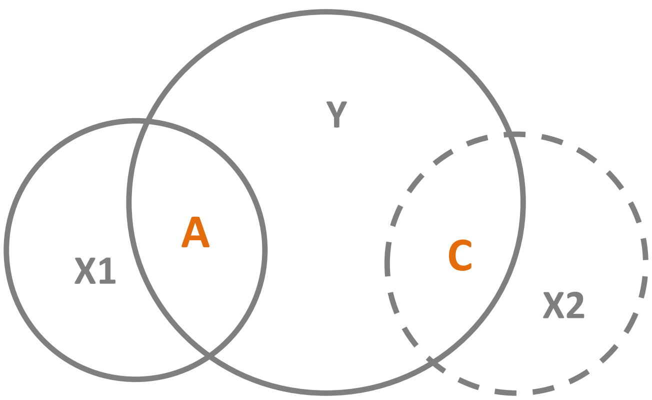
In this case, the omitted variable X2 is uncorrelated with the policy variable X1. There is no overlap in the Venn Diagram.



$$b_1 \approx A$$
$$\beta_1 \approx A$$
$$b_1 = \beta_1 \Rightarrow bias \approx 0$$

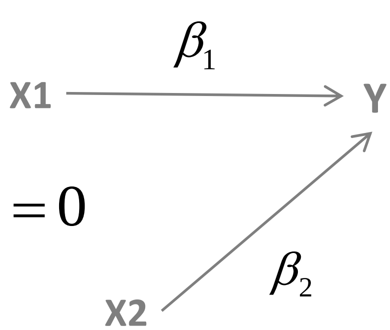
Since the naïve slope, b_1, and the full-model slope, B_1, are the same, there is no bias that results from omitting X2.

CASE 2: OMITTED VARIABLE UNCORRELATED WITH POLICY VARIABLE



$$b_1 \approx A$$
$$\beta_1 \approx A$$
$$b_1 = \beta_1 \Rightarrow bias = 0$$

Path Diagram



$$\alpha_1 = 0$$

$$bias = \alpha_1 \beta_2$$
$$bias = 0 \cdot \beta_2 = 0$$