

---

# API-202B Empirical Methods II

## Session #5: Omitted variables

[miguel\\_santos@hks.harvard.edu](mailto:miguel_santos@hks.harvard.edu)  
@miguelsantos12

# Omitted variable bias (OVB): An example to motivate the discussion

---

## **UVA Study: Surgical Patients on Medicaid Are 13% More Likely to Die Than Those Without Insurance**

By Avik Roy, National Review Online, July 17, 2010

“...Medicaid is broken.... a large study by the University of Virginia [found] that surgical patients on Medicaid are 13% more likely to die than those with no insurance at all...”

“[The authors] divided the patients up by the type of insurance... and adjusted the database in order to control for age, gender, income, geographic region, operation, and comorbid conditions.”

# Today's class: Omitted variable bias (OVB)

---

- Defining omitted variable bias
- An example
- Quantifying omitted variable bias
- Signing the omitted variable bias
- Two more examples of OVB
- Takeaways
- Vocabulary
- Optional: Omitted variable bias formula derivation

# What is Omitted Variable Bias (OVB)?

---

- Multivariate regression controls for factors possibly influencing the outcome other than the factor of interest: It is often stronger than bivariate regression when we are doing causal inference
- We will formalize the concept of **omitted variable**: a factor we don't observe but might explain some of the association between our X and Y of interest
- **Omitted variable bias (OVB)** occurs when two conditions are true:
  1. The OV is correlated with our regressor of interest (i.e. treatment, STR)
  2. Correlated with the outcome we are measuring (i.e. test scores)
- Such omitted variables can make **OLS estimates biased**, so that they do not accurately measure causal impacts
- We will learn how to assess the magnitude and sign of the **bias**

## Omitted Variable Bias (OVB): An example

---

- We want to estimate the causal effect of health insurance on health. Does having insurance make you healthier?
- Assume we have data that allow us to run this bivariate PRF:

$$healthy_i = \beta_0 + \beta_1 insured_i + \varepsilon_i$$

Y            Healthy = self-reported health            (on a scale of 1-10)

X<sub>1</sub>        Insured = 1 if insured                            (= 0 otherwise)

- Would this study have internal validity? Does  $\hat{\beta}_1$  provide a causal estimate of the effect of insurance? Explain.

# What is Omitted Variable Bias (OVB)?

---

- These alternate explanations cause omitted variable bias (OVB)
- Why not correct this problem by adding the relevant variables to the model, using multivariate regression to eliminate OVB?
  - We are looking at someone else's analysis
  - We don't have the omitted variable in our data set
  - We don't know what to control for
- It is thus important to understand how OVB can potentially affect our estimates because we might not always be able to control for it/eliminate it

## Quantifying Omitted Variable Bias (OVB)

---

- Assume we are trying to estimate the impact of  $X_1$  on  $Y$  but another factor  $X_2$  may also be important (and potentially correlated with  $X_1$  and  $Y$ )
- The **true PRF** is thus given by the following *long* regression:

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

- If  $X_2$  is not observable/not in our data we can only run this *short* regression:

$$Y = \alpha_0 + \alpha_1 X_1 + v$$

- The bias is defined as the difference between the true impact  $\beta_1$  (hypothetically measured by the long regression) and the measured impact  $\alpha_1$  (derived from our short regression)

$$\text{Bias} = \alpha_1 - \beta_1$$

## Quantifying Omitted Variable Bias (OVB)

---

- OVB can make your estimate of the causal impact to be:
  - **Overstated (overestimated):** Omitted variable makes the size of the estimate larger than the true one (further away from zero).
    - Occurs if true impact  $\beta_1$  and bias have same signs.
  - **Understated (underestimated):** Omitted variable makes the magnitude of the estimate smaller than the truth (closer to zero).
    - Occurs if true impact  $\beta_1$  and bias have opposite signs, **and** absolute value of bias is smaller than true impact  $\beta_1$ .
    - [In extreme cases where the bias is larger than the true impact  $\beta_1$ , the sign of the coefficient can be the opposite of the truth]



## Omitted Variable Bias (OVB): An example

---

- Let us return to the health insurance variable and focus on education as an OV

$X_2$        $HSgrad = 1$  if high-school graduate (= 0 otherwise)

- Here are the estimates of the *short* regression:

$$healthy_i = \hat{\alpha}_0 + \hat{\alpha}_1 insured_i + \hat{v}_i$$

```
. regress healthy insured, robust noheader
```

healthy	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
insured	.1773881	.0553319	3.21	0.001	.0689246	.2858515
_cons	7.001143	.0502654	139.28	0.000	6.902611	7.099675

- Interpret  $\hat{\alpha}_1$ , the insured coefficient. Is it statistically significant?

## Omitted Variable Bias (OVB): An example

---

- Here are the estimates of the *long* regression:

$$healthy_i = \hat{\beta}_0 + \hat{\beta}_1 insured_i + \hat{\beta}_2 hsgrad_i + \hat{u}_i$$

```
. regress healthy insured hsgrad, robust noheader
```

healthy	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
insured	.0909924	.0564715	1.61	0.107	-.019705	.2016898
hsgrad	.3668858	.0662785	5.54	0.000	.2369645	.4968071
_cons	6.765707	.0693001	97.63	0.000	6.629863	6.901551

- Interpret  $\hat{\beta}_1$ , the *insured coefficient*. Is it statistically significant?
- By ignoring *hsgrad*, what is the size of the bias in the coefficient of “insured”?

## Figuring out sign of Omitted Variable Bias (OVB) when we lack the OV

---

- A little algebra (see bonus slide) reveals that bias can be expressed as:

$$\text{Bias} = \alpha_1 - \beta_1 = \beta_2 \gamma_1$$

- $\beta_2$ : relationship of  $X_2$  to  $Y$ , from the *long* regression
- $\gamma_1$ : relationship of  $X_2$  to  $X_1$ , from an *auxiliary* regression of  $X_2$  on  $X_1$
- The coefficient in the short regression is thus biased only if:
  - $X_2$  is a true determinant of  $Y$  ( $\beta_2 \neq 0$ ) **AND**
  - $X_2$  is correlated with  $X_1$  ( $\gamma_1 \neq 0$ )
- Hard to compute the sign of OVB bias if we can't compute  $\beta_2$  and  $\gamma_1$
- We can -however- figure out the sign of OVB using by observing that:

$$\text{Sign of bias} = \text{Sign of corr}(Y, X_2) * \text{Sign of corr}(X_1, X_2)$$

$$\text{Sign of bias} = \text{Sign of corr}(Y, X_2) * \text{Sign of corr}(X_1, X_2)$$

---

- This formula doesn't help much with the magnitude of the bias because we often can't compute  $\beta_2$  and  $\gamma_1$
- We can often figure out the sign of the bias by making educated guesses about the signs of  $\beta_2$  and  $\gamma_1$
- Fill in the following chart with + or - symbols representing a positive or negative bias in each of these cases:

	$\text{Corr}(X_1, X_2) > 0$	$\text{Corr}(X_1, X_2) < 0$
$\text{Corr}(Y, X_2) > 0$	+	-
$\text{Corr}(Y, X_2) < 0$	-	+

## The sign of Omitted Variable Bias (OVB)

---

- In the health insurance example, we found a positive bias. Can you relate this to the signs of  $\hat{\beta}_2$  and  $\hat{\gamma}_1$ ? Explain.

## Omitted Variable Bias (OVB) Example

---

- For each of the following examples:
  - Write down the short and long regressions
  - Determine the (expected) signs of  $\hat{\beta}_2$ ,  $\hat{\gamma}_1$  and the bias
  - Explain the sign of the bias in words
  - Determine whether  $\hat{\alpha}_1$  is an under- or over-estimate of the true impact of the “treatment”

## Omitted Variable Bias (OVB) Example #1

---

- The Tennessee STAR experiment showed small classes had test scores 13.7 points higher than regular classes. Describe the bias from omitting family income as an explanatory variable.

## Omitted Variable Bias (OVB) Example #2

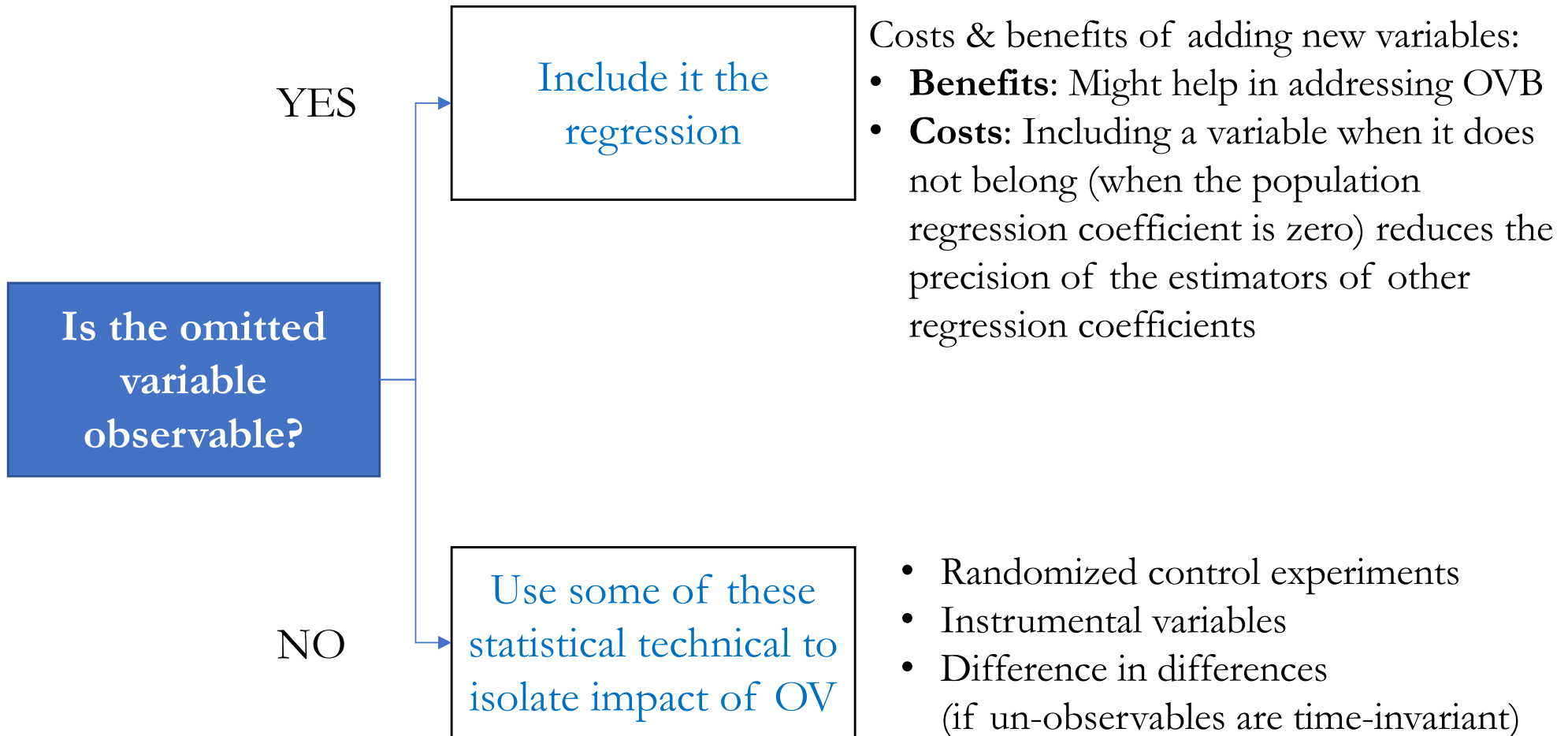
---

- For 65 year-olds, each additional daily dose of aspirin is associated with 0.2 fewer heart attacks per lifetime. Describe the bias from omitting the number of heart attacks prior to age 65 as an explanatory variable.



# How to address omitted variable bias (OVB)?

---



# Takeaways

---

- OVB is the single most widespread problem in studies attempting to measure the impacts of public policies or other interventions
- Our goal is to make you a critical consumer of such claims
- Ask yourself: Does another story (i.e. omitted variable) explain some or all of the association that the researchers have found?
- With practice, you will:
  - Develop an intuition for the sign (and maybe magnitude) of the bias
  - Better understand how the reported impact of the policy relates to its true impact
- Is the OV causing the bias observable? Include it in the regression (mind the cost and benefits of adding variables to a multiple regression)
- If the OV causing the bias is not observable, there are other statistical techniques that might help in isolating its impacts

## Be prepared for next class

---

- Read the IMPACT case carefully.
- Pay particularly close attention to Exhibit 1.
- Ask yourself whether you think this regression successfully helps identify which teachers are effective at raising students' test scores.

# Vocabulary

---

- Omitted variable
- Omitted variable bias (conditions for)
- OVB making the causal impact you are measuring
  - Overestimated
  - Underestimated
- Sign of OVB bias – Conditions for OVB to be:
  - Positive
  - Negative

## Bonus: OVB formula derivation

---

- We will now do a bit of math to come up with a formal measurement of OVB, in the context of multiple regression with two explanatory variables. The algebra below may get a bit confusing, but the examples that follow may clarify how we use the algebraic results in practice.
- Assume we are trying estimate the impact of  $X_1$  on  $Y$ , but that another factor  $X_2$  may also affect  $Y$ . The true PRF is thus given by the following "long" regression:

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

- Now assume that  $X_2$  is an omitted variable, perhaps because it's not in our data set. We are thus forced to run the following "short" regression:

$$Y = \alpha_0 + \alpha_1 X_1 + v$$

- In other words, we are trying to estimate the true causal impact  $\beta_1$  but instead can only estimate a different quantity  $\alpha_1$ . The question is: how different are these two estimates?
- This gives us the definition of the bias, the difference between the true impact (measured by the long regression) and the impact we estimate when we omit a variable (in the short regression).
- The mathematical definition of the **bias** is thus:

$$\text{Bias} = \alpha_1 - \beta_1$$

- The practical problem we encounter when doing or reading research is to get a sense of how big this bias is and whether it's positive or negative. We turn now to algebra that will help us figure this out.

## OVB formula derivation

---

To see how we can measure the bias in terms of quantities we might know something about, we can do the following algebraic manipulations:

**Step 1:** Estimate the relationship between  $X_1$  and  $X_2$  by running the regression:

$$X_2 = \gamma_0 + \gamma_1 X_1 + w$$

$\gamma_1$  gives us the relationship between  $X_1$  and  $X_2$ .

**Step 2:** Substitute the above expression for  $X_2$  into the “long” PRF

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(\gamma_0 + \gamma_1 X_1 + w) + u$$

Now the long PRF includes only the variables  $Y$  and  $X_1$ .

**Step 3:** Multiply the  $\beta_2$  across the terms in parentheses.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \gamma_0 + \beta_2 \gamma_1 X_1 + \beta_2 w + u$$

**Step 4:** Group terms by whether constant, containing  $X_1$ , or containing error terms.

$$Y = [\beta_0 + \beta_2 \gamma_0] + [\beta_1 + \beta_2 \gamma_1] X_1 + [\beta_2 w + u]$$

## OVB formula derivation

---

**Step 5:** Now observe that this has the form of our short regression:

$$Y = \alpha_0 + \alpha_1 X_1 + v$$

In particular, the coefficient  $\alpha_1$  can be expressed in other terms:

$$\alpha_1 = \beta_1 + \beta_2 \gamma_1$$

**Step 6:** Rearrange this expression to get a formula for the omitted variable bias!

$$\text{Bias} = \alpha_1 - \beta_1 = \beta_2 \gamma_1$$

- The **magnitude of the bias** thus depends on the magnitudes of  $\beta_2$  and  $\gamma_1$ . This yields two conditions for OVB. The slope coefficient in the short regression will be biased only if:

$$X_2 \text{ is a true determinant of } Y \quad (\beta_2 \neq 0)$$

and

$$X_2 \text{ is correlated with } X_1 \quad (\gamma_1 \neq 0)$$

- If both of these conditions hold, then the estimated slope coefficient for the short regression will be biased. In other words, will be a biased estimator of  $\beta_1$ .