HANDOUT 5 – OMITTED VARIABLES

AGENDA

- Defining omitted variable bias (OVB)
- An example
- Quantifying omitted variable bias (OVB)
- Signing the omitted variable bias (OVB)
- Three more Examples of OVB
- Takeaways
- Vocabulary
- Optional (bonus): Omitted variable bias formula derivation

BIBLIOGRAPHY FOR TODAY´S CLASS

- Roy, Avik. National Review Online, July 17, 2010. UVa Study: Surgical Patients on Medicaid Are 13% More Likely to Die Than Those Without Insurance. (*)
- Stock and Watson (2007), 6.1, 9.2 (pp. 316-318) (**)

WHAT IS OBV?

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 omitted variable (OV) is correlated with our regressor of interest (treatment, STR)
2. The omitted variable (OV) is correlated with the outcome we are measuring (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**

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 allows us to run this bivariate PRF:

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

Y	Healthy=self-reported health	(on scale of 1-10)
X_1	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.

These alternate explanations cause omitted variable bias (OVB).

Why not correct this problem by adding the relevant variable 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 β_1 (hypothetically measured by the long regression) and the measured impact α_1 (derived from our shot regression).

$$Bias = \alpha_1 - \beta_1$$

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 of the true impact and bias have same signs.
- **Understated (underestimated):** omitting variable makes the magnitude of the estimate smaller than the truth (closer to zero).
    - Occurs if true impact and bias have opposite signs, **and** absolute value of bias is smaller than true impact.
    - [In extreme cases where the bias is larger than the true impact, the sign of the coefficient can be the opposite of the truth]

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?

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"?

SIGNING THE OMITTED VARIABLE BIAS (OVB) (when we lack the OV)

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

$$Bias = \alpha_1 - \beta_1 = \beta_2\gamma_1$$

β_2: relationship of X_2 to Y, from the *long* regression.

γ_1: relationship of X_2 to X_1, from a bivariate 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 can't compute β_2 and γ_1

We can -however- figure out the sign of OVB using by observing that:

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

This formula doesn't help much with the magnitude of the bias because we often can't compute β_2 and γ_1

We can often figure out the sign of the bias by making educated guesses about the signs of β_2 and γ_1

Fill in the following chart with + or - symbols representing a positive or negative bias in each of these cases:

	$Corr(X_1,X_2) > 0$	$Corr(X_1,X_2) < 0$
$Corr(Y,X_2) > 0$		
$Corr(Y,X_2) < 0$		

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

EXAMPLEs OF OVB

For each of the following examples:

- Write down the short and long regressions.
- Determine the (expected) sign 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".

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.

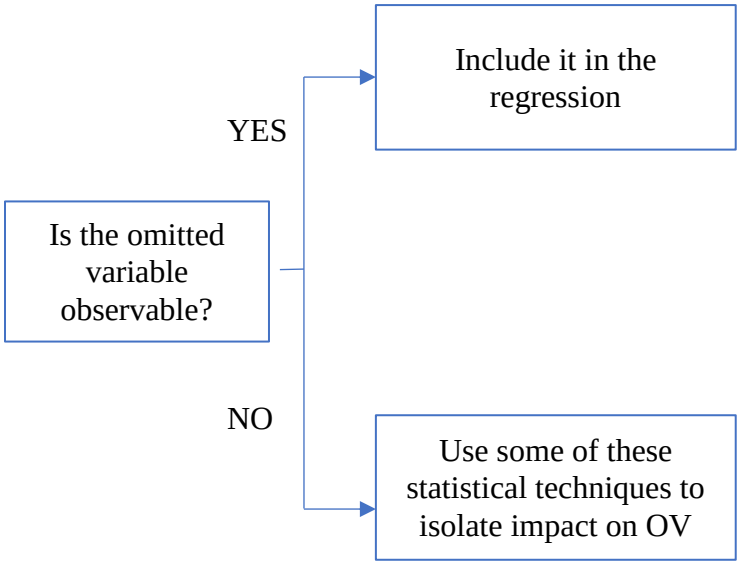
Example 2:

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.

Example 3:

Students who received subsidizes lunch in schools score 10 points lower on tests than students who pay full price. Describe the bias from omitting family income as an explanatory variable.

How to address omitted variable bias. (OVB)?



Costs & benefits of adding new variable:

- Benefits: might help in addressing OVB.
- Costs: including a variable when it does not belong (when the population regression coefficient is zero) reduces the precision of the estimators of other regression coefficients.

- Randomized control experiments
- Instrumental variables
- Difference in differences (if un-observables are time-invariant)

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 rates 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).
- Is the OV causing the bias is not observable, there are other statistical techniques that might help in isolating its impacts
    - Randomized control experiments
    - Instrumental variables
    - Difference-in-differences (if un-observables are time invariant)

VOCABULARY

- Omitted variable: any variable not included as an independent variable in the regression that might influence the regressor(s) and dependent variable of interest
- Omitted variable bias (conditions for): arises because a variable that is a determinant of Y and is correlated with a regressor has been omitted from the regression.
- OVB can make your estimate of the causal impact to be (in terms of β_1 – coefficient accompanying the variable of interest in the long regression; α_1 – coefficient accompanying the variable of interest in the short) regression; and Bias $= \alpha_1 - \beta_1$)
    - **Overstate (overestimated):** Omitting variable makes the magnitude of the estimate larger than the truth (farther from zero).
        - Occurs of the true impact and bias have same signs.
    - **Understated (underestimated):** omitting variable makes the magnitude of the estimate smaller than the truth (closer to zero).
        - Occurs if true impact and bias have opposite signs, **and** absolute value of bias is smaller than true impact.

- Sign of OVB Bias: (Y: dependent variable, X_1: included independent variable, X_2 excluded independent variable).
    - Positive:
        - Negative correlation X_1 and X_2, negative correlation Y and X_2; or
        - Positive correlation X_1 and X_2, positive correlation Y and X_2.
    - Negative:
        - Positive correlation X_1 and X_2, negative correlation Y and X_2; or
        - Negative correlation X_1 and X_2, positive correlation Y and X_2.

OPTIONAL (BONUS): OMITTED VARIABLE BIAS 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 β_1 but instead can only estimate a different quantity α_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.
- 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:

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

γ_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 β_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.

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

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 α_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 β_2 and γ_1. This yields two conditions for OVB. The slope coefficient in the short regression will be biased only if:
  - X_2 is a true determinant of Y $(\beta_2 \neq 0)$

and

X_2 is correlated with X_1 $(\gamma_1 \neq 0)$

- 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 β_1.