Webinar Series in Applied Quantitative Analysis

Date	Topic
	Session One
February 29	Potential Outcomes and Omitted Variable Bias I (Theory)
March 7	Potential Outcomes and Omitted Variable Bias II (Application)
	Session Two
March 21	Difference-in-differences I (Theory)
March 28	Difference-in-differences II (Application)
	Session Three
April 25	Power analysis, clustering and sample size calculations I (Theory)
May 2	Power analysis, clustering and sample size calculations II (Application)
	Session Four
May 23	Fixed-effects I (Theory)
May 30	Fixed-effects II (Application)
	Session Five
June 20	Instrumental variables I (Theory)
June 27	Instrumental Variables II (Application)
	Session Six
July 25	Lagged dependent variables and the Arellano-Bond Estimator I (Theory)
August 1	Lagged dependent variables and the Arellano-Bond Estimator II (Application)

AIR®
Advancing Evidence.
Improving Lives.

https://www.airp3-africa.org/applied-quantitative-analysis-webinar-series

AERC
CREA

Potential outcomes and selection bias

Ashu Handa

Institute Fellow – AIR

Kenan Eminent Professor of Public Policy – UNC-CH

March 7, 2024

Recap of Session 1 on this topic

- Examined the implications of omitting an important variable from the OLS regression
  - Derived the omitted variable bias formula or 'bias formula'
  - The most important formula of your PhD career
  - Used this formula to understand selection bias

- The Bias Formula
  - $Y_i = \beta_0 + \beta_1 P + \beta_2 Z + \mu_i$ $\quad$ *True relationship*
  - $Y_i = \alpha_0 + \alpha_1 P \qquad\qquad + \varepsilon_i$ $\quad$ *Estimated relationship (Z omitted)*

  - $\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}$

$$\widehat{\alpha_1} = \underbrace{\beta_1}_{\text{True 'impact' of P on Y}} + \underbrace{\beta_2 * \widehat{\gamma_1}}_{\text{Selection bias}}$$

Selection bias has two elements

β_2 = the relationship between the omitted variable (Z) and the outcome (Y)

γ1 = the relationship between the omitted variable (Z) and included variable (P)

Question

What if the omitted variable (Z) is unrelated to the included variable (P)?

(So γ_1 in the Bias formula is zero)

Do we have selection bias?

Do orphans have worse nutritional status? OLS regression results using Lesotho DHS
Dependent variable is height for age z-score of children aged 0-60 months
Variable of interest is orphan (1 if orphan, 0 otherwise): orphans have mean z-score 0.214 less than non-orphans,
Controlling for sex, education of head and poorest wealth quintile

We forgot to control for age!! What is the bias in the 'orphan' coefficient?

Use the bias formula!

```
. reg haz female orphan hhed poorest

      Source |       SS           df       MS      Number of obs   =     2,119
-------------+----------------------------------   F(4, 2114)      =     12.94
       Model |  84.1278346         4  21.0319586   Prob > F        =    0.0000
    Residual |  3436.78063     2,114  1.62572404   R-squared       =    0.0239
-------------+----------------------------------   Adj R-squared   =    0.0220
       Total |  3520.90846     2,118  1.66237416   Root MSE        =     1.275

------------------------------------------------------------------------------
         haz | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   .1801014   .0554511     3.25   0.001      .071357    .2888459
      orphan |  -.2141423   .0785791    -2.73   0.006    -.3682427   -.0600418
        hhed |   .0332078   .0074088     4.48   0.000     .0186786     .047737
     poorest |  -.1300109   .0640165    -2.03   0.042    -.2555528    -.004469
       _cons |  -1.586969   .0594029   -26.72   0.000    -1.703463   -1.470474
------------------------------------------------------------------------------
```

$$\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}$$

Estimated coefficient = true coefficient + {COV(omitted, outcome)*COV(omitted, included)}

-0.214 = β_1 + {COV(age, nutritional status)*COV(age, orphan)}

-0.214 = β_1 + { (?) * (?) }

```
. reg haz female orphan hhed poorest

      Source |       SS           df       MS      Number of obs   =     2,119
-------------+----------------------------------   F(4, 2114)      =     12.94
       Model |  84.1278346         4  21.0319586   Prob > F        =    0.0000
    Residual |  3436.78063     2,114  1.62572404   R-squared       =    0.0239
-------------+----------------------------------   Adj R-squared   =    0.0220
       Total |  3520.90846     2,118  1.66237416   Root MSE        =     1.275

------------------------------------------------------------------------------
         haz | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   .1801014   .0554511     3.25   0.001      .071357    .2888459
      orphan |  -.2141423   .0785791    -2.73   0.006    -.3682427   -.0600418
        hhed |   .0332078   .0074088     4.48   0.000     .0186786     .047737
     poorest |  -.1300109   .0640165    -2.03   0.042    -.2555528   -.004469
       _cons |  -1.586969   .0594029   -26.72   0.000    -1.703463   -1.470474
------------------------------------------------------------------------------
```

-0.214	=	β_1	+ {COV(age, nutritional status)*COV(age, orphan)}
-0.214	=	β_1	+ { (-) * (+) }
-0.214	=	β_1	+ { - }
-0.214	<	β_1	

```
. reg haz female orphan hhed poorest age

      Source |       SS           df       MS      Number of obs   =     2,119
-------------+----------------------------------   F(5, 2113)      =     32.05
       Model |  248.223136         5  49.6446273   Prob > F        =    0.0000
    Residual |  3272.68533     2,113  1.54883357   R-squared       =    0.0705
-------------+----------------------------------   Adj R-squared   =    0.0683
       Total |  3520.90846     2,118  1.66237416   Root MSE        =    1.2445

------------------------------------------------------------------------------
         haz | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   .1508001   .0541987     2.78   0.005     .0445116    .2570885
      orphan |  -.1443102   .0769978    -1.87   0.061    -.2953097    .0066892
        hhed |   .0316495    .007233     4.38   0.000      .017465    .0458341
     poorest |  -.1309022   .0624844    -2.09   0.036    -.2534394   -.0083649
         age |  -.0163933   .0015926   -10.29   0.000    -.0195166   -.0132699
       _cons |  -1.079907   .0760828   -14.19   0.000    -1.229112   -.9307018
------------------------------------------------------------------------------
```

Recap: ‘Effect’ of orphanhood on nutritional status in the presence of omitted variable bias

- When we fail to control for age, we over-estimate the negative ‘effect’ of orphanhood on nutritional status (-0.214 vs -0.144)

- Why?
  - There is a positive correlation between age and orphanhood (older children are more likely to be orphaned)
  - And there is a negative correlation between age and nutritional status (older children are more likely to be malnourished)
  - The selection bias term is negative, so our estimate is too low (in absolute value)

- Rule: If selection bias is negative, our estimate is too low (in absolute value); if selection bias is positive, our estimate is too high (in absolute value)

- Example: "What is the economic return to completing tertiary education"
- Y=log(earnings), 'tertiary'=1 if went to university, 0 if not
- Coefficient of tertiary=0.613 — earnings are 61% higher for those who went to university relative to those who did not, controlling for age, sex, race, region of residence and marital status (U.S. Current Population Survey)

Issue: 'ability' – those with higher innate ability more likely to go to university, and would also earn more even without university. How does this bias our estimated coefficient? Is 61% an overestimate or underestimate?

```
. reg log_inc tertiary age age2 male white south marry
```

Source	SS	df	MS
Model	2439.82254	7	348.546077
Residual	4721.12634	4,530	1.04219124
Total	7160.94888	4,537	1.57834447

Number of obs	=	4,538
F(7, 4530)	=	334.44
Prob > F	=	0.0000
R-squared	=	0.3407
Adj R-squared	=	0.3397
Root MSE	=	1.0209

log_income	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
tertiary	.6129801	.035577	17.23	0.000	.5432318	.6827284
age	.2021754	.0061616	32.81	0.000	.1900957	.2142551
age2	-.0021902	.0000715	-30.65	0.000	-.0023303	-.0020501
male	.5183789	.0305843	16.95	0.000	.4584187	.5783391
white	.0633043	.0442809	1.43	0.153	-.0235078	.1501165
south	.0047449	.0336276	0.14	0.888	-.0611816	.0706714
marry	.086426	.0342922	2.52	0.012	.0191965	.1536555
_cons	5.060835	.1255051	40.32	0.000	4.814784	5.306886

Outcome = income
Included variable (variable of interest) = tertiary schooling
Omitted variable = ability (unobserved)
Estimated coefficient (with omitted variable bias) = 0.61
Bias formula: $\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}$

$\widehat{\alpha_1} = \beta_1 + \beta_2 * \widehat{\gamma_1}$
Estimated coefficient = true coefficient + {COV(omitted, outcome)* COV(omitted, included)}
0.61 = β_1 + {COV(ability, income) * COV(ability, tertiary)}

0.61 = β_1 + {COV(ability, income)*COV(ability, tertiary)}
0.61 = β_1 + { (?) * (?) }
0.61 = β_1 + { (+) * (+) }

0.61 = β_1 + { (positive number) }
0.61 > β_1
<u>Our estimate is too high; we have overestimated the return to tertiary due to ability bias</u>

Today's session: Potential outcomes framework to understand selection bias

- Popular framework in the study of causal inference
  - Rubin, Donald (2005). "Causal Inference Using Potential Outcomes". *J. Amer. Statist. Assoc.* **100** (469): 322–331.
  - https://www.tandfonline.com/doi/abs/10.1198/016214504000001880

- Typically used when studying treatment effects in impact studies
  - Did program P have an impact on outcome Y?
  - Impact implies causality
  - To measure impact, we must know the counterfactual—what would have happened in the absence of the program

Definitions

$$P = \begin{cases} 1, if\ individual\ participates\ in\ the\ program \\ 0, if\ individual\ does\ not\ participate \end{cases}$$

$$Y_i^1: Outcome\ of\ individual\ i\ with\ the\ program$$

$$Y_i^0: Outcome\ of\ individual\ without\ the\ program$$

Definitions

- Impact of program on person i: $\underbrace{Y_i^1}_{\text{with}} - \underbrace{Y_i^0}_{\text{without}}$

- Impact of program on the entire population of treated: $E[Y^1 - Y^0|P = 1]$

- Average treatment effect on the treated [ATT]

- Issue: In any moment in time, we only see the person's outcome in one state, either with the program or without the program

Average treatment effect on the treated [ATT] consists of two parts

- $ATT = E[(Y^1 - Y^0|P = 1])$

- $ATT = \underbrace{E[Y^1\,|P = 1]}_{\text{Outcome with program of those who receive (observed)}} - \underbrace{E[Y^0\,|P = 1]}_{\text{Outcome without program of those who receive (unobserved)}}$

The evaluation problem is the problem of the missing counterfactual

How do we approximate the missing counterfactual?

Use the outcome without the program (Y^0)of those who did not receive the program ($P=0$)

$$\underbrace{E[Y^0|P=0]}_{\text{Outcome without program of those who did not receive (observed)}} \overset{?}{=} \underbrace{E[Y^0|P=1]}_{\text{Outcome without program of those who did receive (unobserved)}}$$

Are these two the same? Usually not—this is the evaluation problem

How do we approximate the missing counterfactual?

Use the outcome without the program (Y^0)of those who did not receive the program (P=0)

$$\underbrace{E[Y^0|P=0]}_{\text{Control or comparison group (observed)}} \stackrel{?}{=} \underbrace{E[Y^0|P=1]}_{\text{Ideal control group (unobserved)}}$$

Are these two the same? Usually not—this is the evaluation problem

The evaluation problem restated

$$ATT = E[Y^1|P=1] - E[Y^0|P=1]$$

$$+\{E[Y^0|P=0] - E[Y^0|P=0]\}$$

Add and subtract E[Y^0|P=0] then rearrange terms

$$ATT = \{E[Y^1|P=1] - E[Y^0|P=0]\}$$

Observed from treatment and comparison groups

$$+ \{E[Y^0|P=0] - E[Y^0|P=1]\}$$

Selection bias

ATT consists of two terms

$$ATT = \{E[Y^1 \,|P = 1] - E[Y^0|P = 0]\}$$

What we estimate
Difference in Y between the treated and the control or comparison group

$$+ \{\underbrace{E[Y^0 \,|P = 0]}_{\boxed{\text{Control or comparison group (observed)}}} \; - \; \underbrace{E[Y^0|P = 1]}_{\boxed{\text{Ideal (unobserved)}}}\}$$

Selection bias
The difference between our control group and the ideal counterfactual

Example: Impact of health insurance

Person B used as the counterfactual for person A

	Person A [frail]	Person B [healthy]
P (1=treated, 0=untreated)	1	0
Y^0 (outcome w/o P)	3	5
Y^1 (outcome with P)	4	5
Observed outcome	4	5
Unobserved outcome	3	5

Observed effect: $\{[Y^1 \mid P=1] - [Y^0 \mid P=0]\}$	4 - 5 = -1
Selection bias: $\{[Y^0 \mid P=0] - [Y^0 \mid P=1]\}$	5 - 3 = +2
True impact (ATT)	+1

Selection bias practice

Program	Outcome	Selection bias [positive=better off enter program]	Observed effect too high or too low? **$\{[Y^1 \mid P=1] - [Y^0 \mid P=0]\}$**
Completed university	Earnings	Positive	Too high (over-estimate)
Nutrition programs placed purposefully in areas of high malnutrition	Child nutrition	Negative	Too low (under-estimate)
Health spending	Health	Negative	Too low (under-estimate)
Special training in econometrics	Earnings 3 years later	Positive	Too high (over-estimate)

How can we estimate the missing counterfactual?

- Randomized controlled trial (RCT) – strongest approach because we only study units for whom P=1, so we actually observe $E[Y^0|P=1]$

- Quasi-experimental designs
  - Propensity score matching – find similar non-treated units to treated units
  - Discontinuity designs – compare people at the 'border' of being treated
  - Difference-in-differences – exploits similarity in trends

- Use econometrics to eliminate selection bias
  - Instrumental variables
  - Fixed effects

More definitions

- X causes Y if and only if
  - X and Y are correlated
  - X precedes Y (temporal precedence)
  - No third factor explains the correlation between X and Y (confounder)
- Experiment – deliberate manipulation of the cause
- Randomized experiment – cause is manipulated by random assignment
- Quasi-experiment – cause is manipulated, but not randomly