5) Omitted Variables Bias (OVB)

Vitor Kamada

June 2018

Reference

Tables, Graphics, and Figures from

Mastering 'Metrics: The Path from Cause to Effect

Angrist & Pischke (2014): Chapter 2

The College Matching Matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Naive Comparison vs Well-Controlled Comparison

Naive Comparison:

$$E(W|Priv.) = 92K$$
$$E(W|Pub) = 72,5K$$

For Group A:

$$E(W|Smart) = 105K$$
$$E(W|Tall) = 110K$$

For Group B:

$$E(W|Ivy) = 60K$$
$$E(W|Altered) = 30K$$

Well-Controlled Comparison:

$$(\tfrac{3}{5} \times -5K) + (\tfrac{2}{5} \times 30K) = 9K$$

Regression

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i$$

$$\alpha = 40K$$

$$\beta = 10K$$

$$\gamma = 60K$$

College and Beyond (C&B) Dataset

$$lnY_i = \alpha + \beta P_i + \sum_{j=1}^{150} y_j group_{ji} + \delta_1 SAT_i + \delta_2 lnPI_i + e_i$$

Y_i: Earnings

P_i: Private School

PI_i: Parental Income

Barron's rankings: Most Competitive, Highly Competitive, Very Competitive, Competitive, Less Competitive, and Noncompetitive

Private School Effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Private School Effects: Average SAT Score Controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

Omitted Variables Bias (OVB)

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l$$

$$Y_i = \alpha^s + \beta^s P_i + e_i^s$$

$$A_i = \pi_0 + \pi_1 P_i + u_i$$

$$\beta^s = \beta^l + \pi_1 \times \gamma$$

$$\pi_1 \times \gamma = .1667 \times 60K = 10K$$

$$OVB = \beta^s - \beta^l = 20K - 10K$$

Summary of Bias in $\tilde{\beta}_1$ when x_2 is Omitted

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

$$wage = \beta_0 + \tilde{\beta}_1 educ + v$$

$$v = \beta_2 abil + u$$

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	$+$	$-$
$\beta_2 < 0$	$-$	$+$

Summary of Functional Forms Involving Logarithms

Model	Dep. Variable	Ind.Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$log(y)$	$log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Load Data from Blackburn and Neumark (1992)

Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials, Quarterly Journal of Economics 107, 1421-1436.

import pandas as pd

df = pd.read_stata('C:\\Users\\Vitor\\Desktop\\ECO 6100 Introduction to Econometrics (Fall 2018)\\Lectures\\5) Omitted Variables Bias\\WAGE2.dta')

df['const'] = 1

import statsmodels.api as sm

df.head()

	wage	hours	IQ	KWW	educ	exper	tenure	age
0	769	40	93	35	12	11	2	31
1	808	50	119	41	18	11	16	37
2	825	40	108	46	14	11	9	33
3	650	40	96	32	12	13	7	32
4	562	40	74	27	11	14	5	34

	urban	sibs	brthord	meduc	feduc	lwage
0	1	1	2.0	8.0	8.0	6.645091
1	1	1	NaN	14.0	14.0	6.694562
2	1	1	2.0	14.0	14.0	6.715384
3	1	4	3.0	12.0	12.0	6.476973
4	1	10	6.0	6.0	11.0	6.331502

Omitted Ability Bias

```
reg1 = sm.OLS(df['lwage'], df[['const', 'educ']],
                    missing='drop').fit()
print(reg1.summary())
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9731	0.081	73.403	0.000	5.813	6.133
educ	0.0598	0.006	10.035	0.000	0.048	0.072

```
reg2 = sm.OLS(df['lwage'], df[['const', 'educ', 'IQ']],
                    missing='drop').fit()
print(reg2.summary())
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.6583	0.096	58.793	0.000	5.469	5.847
educ	0.0391	0.007	5.721	0.000	0.026	0.053
IQ	0.0059	0.001	5.875	0.000	0.004	0.008

Omitted Ability Bias

```
reg3 = sm.OLS(df['IQ'], df[['const', 'educ']],
                    missing='drop').fit()
print(reg3.summary())
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.6872	2.623	20.468	0.000	48.540	58.835
educ	3.5338	0.192	18.385	0.000	3.157	3.911

$$.0598 \cong .0391 + 3.533(.0059)$$