MRU Mastering Econometrics **M. Joshway**
Spring 2020

The Long and Short of OVB

1 The OVB formula

1.1 One vs. Two

You'd like to regress log wages (Y_i) on schooling (S_i), controlling for ability (A_i):

$$Y_i = \alpha + \rho S_i + \gamma A_i + \epsilon_i \tag{1}$$

Your desired regression arises, perhaps, from a linear model for the CEF (linearity is not obligatory; see next note for details). Alas, you don't observe ability, so you make do with the short regression on schooling alone:

$$Y_i = \alpha^* + \rho^* S_i + \upsilon_i$$

- Substituting (1) into the formula for bivariate regression slope, we learn that

$$\rho^* = \frac{C(Y_i, S_i)}{V(S_i)} = \rho + \gamma\delta_{AS}$$

  where δ_{AS} is the regression of A_i on S_i. We say:

  Short equals long plus {the effect of omitted in long times the regression of omitted on included}

- This *omitted variables bias* (OVB) formula is regression's golden rule
- In a wage equation, where the omitted variable is ability, OVB is called *ability bias*
  - Too big or too small, that is the question!

1.2 Two vs. Four

Suppose your long regression has four regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i \qquad E[\epsilon_i X_{ji}] = 0;\ j = 1,\ 2,\ 3,\ 4 \tag{2}$$

- *Regression anatomy*: each β_j can be obtained from the bivariate regression of Y_i on $\tilde{x}_{ji}$ where $\tilde{x}_{ji}$ is the residual from a regression of X_{ji} on the other three regressors
- You'd like to estimate the parameters of equation (1), the long regression of your dreams. Alas, you're missing data on X_{3i} and X_{4i}. So, you settle . . . for the short regression:

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \nu_i \qquad E[\nu_i X_{ji}] = 0;\ j = 1,\ 2 \tag{3}$$

- What's the relationship between β_1^* and β_1? Between β_2^* and β_2? The regression anatomy formula formula for β_1^* gives

$$\beta_1^* = \frac{Cov(Y_i, \tilde{x}_{1i})}{V(\tilde{x}_{1i})}, \tag{4}$$

  where $X_{1i} = \gamma_{10} + \gamma_{11} X_{2i} + \tilde{x}_{1i}$ is used to partial out (remove) the influence of X_{2i} on X_{1i}. Now, substitute the long reg for Y_i in (4):

$$\begin{aligned}\beta_1^* &= \frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i,\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\ &= \frac{Cov(\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\ &= \beta_1 \frac{Cov(X_{1i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} + \frac{Cov(\beta_3 X_{3i} + \beta_4 X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\ &= \beta_1 + \frac{\beta_3 Cov(X_{3i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} + \frac{\beta_4 Cov(X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})}\end{aligned}$$

Write this as

$$\beta_1^* = \beta_1 + \beta_3 \delta_{31} + \beta_4 \delta_{41}$$

where

$$\delta_{31} = \frac{Cov(X_{3i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \qquad \text{(Regression of } X_3 \text{ on } X_1 \text{ in a model that includes } X_2\text{)}$$

$$\delta_{41} = \frac{Cov(X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \qquad \text{(Regression of } X_4 \text{ on } X_1 \text{ in a model that includes } X_2\text{)}$$

Likewise,

$$\beta_2^* = \beta_2 + \beta_3 \delta_{32} + \beta_4 \delta_{42}$$

where

$$\delta_{32} = \frac{Cov(X_{3i},\, \tilde{x}_{2i})}{V(\tilde{x}_{2i})} \qquad \text{(Regression of } X_3 \text{ on } X_2 \text{ in a model that includes } X_1\text{)}$$

$$\delta_{42} = \frac{Cov(X_{4i},\, \tilde{x}_{2i})}{V(\tilde{x}_{2i})} \qquad \text{(Regression of } X_4 \text{ on } X_2 \text{ in a model that includes } X_1\text{)}$$

- OVB, same as it ever was:

  Short equals long plus {the effect(s) of omitted times the regression(s) of omitted on included}, all computed in a models maintaining the set of controls included in both short and long

1.3 Sample Short and Long

- OVB formulas hold in the sample as well as in the population. Let $\hat{\beta}_1^*$ be the OLS estimate of β_1^* in (3) and let $\hat{\beta}_2^*$ be the corresponding estimate of β_2^*. Then, we have

  $$\hat{\beta}_1^* = \frac{\sum Y_i \tilde{x}_{1i}}{\sum \tilde{x}_{1i}^2} = \hat{\beta}_1 + \hat{\beta}_3 \hat{\delta}_{31} + \hat{\beta}_4 \hat{\delta}_{41},$$

  where hats denote estimates and $\tilde{x}_{1i}$ is the residual from a regression of X_1 on X_2 in the sample, and

  $$\hat{\beta}_2^* = \frac{\sum Y_i \tilde{x}_{2i}}{\sum \tilde{x}_{2i}^2} = \hat{\beta}_2 + \hat{\beta}_3 \hat{\delta}_{32} + \hat{\beta}_4 \hat{\delta}_{42},$$

  where $\tilde{x}_{2i}$ is the residual from a regression of X_2 on X_1 in the sample.

  – Show this at home

1.4 When Short Equals Long

1. Omitted variables have coefficients of zero in long
2. Omitted variables are uncorrelated with included variables

2 Empirical OVB

Immigrant and native wages (working men aged 40-49 in the 2016 ACS)

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	67,179	44.59836	2.843473	40	49
wagp	67,179	77017	70468.34	4	665000
wkhp	67,179	44.73532	9.786132	1	99
uhe	67,179	33.90219	27.61486	.0016	201.5789
loguhe	67,179	3.264571	.7410341	-6.437752	5.306181
immig	67,179	.2214829	.4152479	0	1
yearsEd	67,179	13.83362	3.240573	0	21
hsgrad	67,179	.9243365	.264461	0	1
somecol	67,179	.4721565	.4992279	0	1
colgrad	67,179	.3860581	.4868478	0	1
asianpac	67,179	.0833147	.2763594	0	1
white	66,790	.7721216	.4194669	0	1
married	67,179	.7207461	.4486359	0	1

```
58 .
59 . ***short vs long***
60 .
61 . reg loguhe immig
```

Source	SS	df	MS	Number of obs	=	67,179
				F(1, 67177)	=	570.52
Model	310.655619	1	310.655619	Prob > F	=	0.0000
Residual	36578.9048	67,177	.544515307	R-squared	=	0.0084
				Adj R-squared	=	0.0084
Total	36889.5604	67,178	.549131567	Root MSE	=	.73791

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.1637641	.0068562	-23.89	0.000	-.1772022	-.1503259
_cons	3.300842	.0032267	1022.99	0.000	3.294518	3.307166

```
62 .         gen beta_short=_b[immig]

63 . reg loguhe immig yearsEd
```

Source	SS	df	MS	Number of obs	=	67,179
				F(2, 67176)	=	8096.70
Model	7165.30841	2	3582.6542	Prob > F	=	0.0000
Residual	29724.252	67,176	.442483208	R-squared	=	0.1942
				Adj R-squared	=	0.1942
Total	36889.5604	67,178	.549131567	Root MSE	=	.66519

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.0346034	.0062671	-5.52	0.000	-.0468869	-.02232
yearsEd	.0999527	.0008031	124.46	0.000	.0983786	.1015267

```
       _cons |   1.889528   .0117062   161.41   0.000     1.866584    1.912472
------------------------------------------------------------------------------

64 .         gen beta_long=_b[immig]

65 .         gen gamma_long=_b[yearsEd]

66 .
67 . **Regression of omitted on included (aux reg)**
68 .
69 . reg yearsEd immig

      Source |       SS           df       MS      Number of obs   =    67,179
-------------+----------------------------------   F(1, 67177)     =   1893.82
       Model |  19342.5545         1  19342.5545   Prob > F        =    0.0000
    Residual |  686114.857    67,177  10.2135382   R-squared       =    0.0274
-------------+----------------------------------   Adj R-squared   =    0.0274
       Total |  705457.411    67,178  10.5013161   Root MSE        =    3.1959

------------------------------------------------------------------------------
     yearsEd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       immig |  -1.292219   .0296939   -43.52   0.000    -1.350419   -1.234019
       _cons |   14.11983   .0139745  1010.40   0.000     14.09244    14.14722
------------------------------------------------------------------------------

70 .         gen delta=_b[immig]

71 .
72 . **check OVB formula**
73 .
74 .         gen short_chk = beta_long + delta*gamma_long

75 .
76 .         sum short_chk beta_short beta_long gamma delta

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   short_chk |     67,179   -.1637641           0  -.1637641  -.1637641
  beta_short |     67,179   -.1637641           0  -.1637641  -.1637641
   beta_long |     67,179   -.0346034           0  -.0346034  -.0346034
  gamma_long |     67,179    .0999527           0   .0999527   .0999527
       delta |     67,179   -1.292219           0  -1.292219  -1.292219

77 .
78 . ***repeat with maintained controls***
79 .
80 . cap drop delta short_chk beta_short beta_long gamma_long delta

81 .
82 . reg loguhe immig married age

      Source |       SS           df       MS      Number of obs   =    67,179
-------------+----------------------------------   F(3, 67175)     =   1261.06
       Model |  1966.79504         3  655.598348   Prob > F        =    0.0000
    Residual |  34922.7653    67,175  .519877415   R-squared       =    0.0533
-------------+----------------------------------   Adj R-squared   =    0.0533
       Total |  36889.5604    67,178  .549131567   Root MSE        =    .72103

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       immig |  -.1844893   .0067131   -27.48   0.000    -.197647   -.1713317
     married |   .3467611   .0062125    55.82   0.000     .3345845    .3589376
        agep |   .0071519   .0009788     7.31   0.000     .0052334    .0090704
       _cons |   2.736543   .0439402    62.28   0.000      2.65042    2.822666
------------------------------------------------------------------------------

83 .         gen beta_short=_b[immig]

84 . reg loguhe immig yearsEd married age

      Source |       SS           df       MS      Number of obs   =    67,179
-------------+----------------------------------   F(4, 67174)     =   4753.57
       Model |  8138.3322         4  2034.58305   Prob > F        =    0.0000
    Residual |  28751.2282    67,174  .428011257   R-squared       =    0.2206
-------------+----------------------------------   Adj R-squared   =    0.2206
       Total |  36889.5604    67,178  .549131567   Root MSE        =    .65423

------------------------------------------------------------------------------
      loguhe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       immig |  -.055513   .0061851    -8.98   0.000    -.0676358   -.0433901
     yearsEd |   .095556   .0007958   120.08   0.000     .0939963    .0971157
     married |  .2638873   .0056791    46.47   0.000     .2527563    .2750183
        agep |  .0086363   .0008882     9.72   0.000     .0068954    .0103772
       _cons |  1.379621   .0414398    33.29   0.000     1.298399    1.460843
```

```
85 .         gen beta_long=_b[immig]

86 .         gen gamma_long=_b[yearsEd]

87 .
88 . **Regression of omitted on included (aux reg)**
89 .
90 . reg yearsEd immig married age

      Source |       SS           df       MS      Number of obs   =    67,179
-------------+----------------------------------   F(3, 67175)     =    979.45
       Model |  29564.8411         3  9854.94703   Prob > F        =    0.0000
    Residual |   675892.57    67,175  10.0616683   R-squared       =    0.0419
-------------+----------------------------------   Adj R-squared   =    0.0419
       Total |  705457.411    67,178  10.5013161   Root MSE        =     3.172

------------------------------------------------------------------------------
     yearsEd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       immig |  -1.349747   .0295329   -45.70   0.000    -1.407631   -1.291862
     married |   .8672798   .0273309    31.73   0.000     .8137113    .9208482
        agep |  -.0155344   .0043061    -3.61   0.000    -.0239744   -.0070944
       _cons |   14.20029   .1933065    73.46   0.000     13.82141    14.57917
------------------------------------------------------------------------------

91 .         gen delta=_b[immig]

92 .
93 . **check OVB formula**
94 .
95 .         gen short_chk = beta_long + delta*gamma_long

96 .
97 .         sum short_chk beta_short beta_long gamma_long delta

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   short_chk |     67,179   -.1844894           0  -.1844894  -.1844894
  beta_short |     67,179   -.1844893           0  -.1844893  -.1844893
   beta_long |     67,179    -.055513           0   -.055513   -.055513
  gamma_long |     67,179        .095           0       .095       .095
       delta |     67,179   -1.349747           0  -1.349747  -1.349747

98 .
99 .         log close
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/2020/1432apps/LN8log.smcl
  log type:  smcl
 closed on:   2 Mar 2020, 14:32:28
```

Private college redux

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.212	0.152	0.139	0.034	0.031	0.037
	(0.060)	(0.057)	(0.043)	(0.062)	(0.062)	(0.039)
Own SAT Score/100		0.051	0.024		0.036	0.009
		(0.008)	(0.006)		(0.006)	(0.006)
Predicted log(Parental Income)			0.181			0.159
			(0.026)			(0.025)
Female			-0.398			-0.396
			(0.012)			(0.014)
Black			-0.003			-0.037
			(0.031)			(0.035)
Hispanic			0.027			0.001
			(0.052)			(0.054)
Asian			0.189			0.155
			(0.035)			(0.037)
Other/Missing Race			-0.166			-0.189
			(0.118)			(0.117)
High School Top 10 Percent			0.067			0.064
			(0.020)			(0.020)
High School Rank Missing			0.003			-0.008
			(0.025)			(0.023)
Athlete			0.107			0.092
			(0.027)			(0.024)
Average SAT Score of				0.110	0.082	0.077
Schools Applied to/100				(0.024)	(0.022)	(0.012)
Sent Two Application				0.071	0.062	0.058
				(0.013)	(0.011)	(0.010)
Sent Three Applications				0.093	0.079	0.066
				(0.021)	(0.019)	(0.017)
Sent Four or more Applications				0.139	0.127	0.098
				(0.024)	(0.023)	(0.020)

Note: Standard errors are shown in parentheses. The sample size is 14,238.

	Dependent Variable					
	Own SAT score/100			Predicted log(Parental Income)		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	1.165	1.130	0.066	0.128	0.138	0.028
	(0.196)	(0.188)	(0.112)	(0.035)	(0.037)	(0.037)
Female		-0.367			0.016	
		(0.076)			(0.013)	
Black		-1.947			-0.359	
		(0.079)			(0.019)	
Hispanic		-1.185			-0.259	
		(0.168)			(0.050)	
Asian		-0.014			-0.060	
		(0.116)			(0.031)	
Other/Missing Race		-0.521			-0.082	
		(0.293)			(0.061)	
High School Top 10 Percent		0.948			-0.066	
		(0.107)			(0.011)	
High School Rank Missing		0.556			-0.030	
		(0.102)			(0.023)	
Athlete		-0.318			0.037	
		(0.147)			(0.016)	
Average SAT Score of			0.777			0.063
Schools Applied To/100			(0.058)			(0.014)
Sent Two Application			0.252			0.020
			(0.077)			(0.010)
Sent Three Applications			0.375			0.042
			(0.106)			(0.013)
Sent Four or more Applications			0.330			0.079
			(0.093)			(0.014)

Note: Standard errors are shown in parentheses. The sample size is 14,238.

- Test the OVB formula: Take *Short* to be the reg on the private school dummy, with no controls and *Long* to be the reg that adds individual SAT scores. In the first table, above, we see

$$Short - Long = OVB = .212 - .152 = .06$$

  As also can be seen in column 2 of the second table, the effect of SAT in the long regression is .051, while the second table shows the regression of SAT (omitted in short) on the private school dummy (included in short) produces a coefficient of 1.165. {Putting these pieces together, we confirm *OVB= Reg of omitted on included* × *Effect of omitted in Long = 1.165×.051=.06.* Phew!

3 OVB What? Selection Bias!

- Why do we care to go long? Private Y_{0i}'s are better (on average)!

- Regression reduces, maybe even eliminates, the resulting selection bias, *provided we've got the right controls*

- Heres a set-up that makes regression causal
  - Let $Y_{0i} = \alpha + \eta_i$, where $E[Y_{0i}] = \alpha$; assume $Y_{1i} - Y_{0i} = \rho$.
  - This means
$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})P_i = \alpha + \rho P_i + \eta_i \tag{5}$$

- Private college isn't randomly assigned, so
$$E[\eta_i|P_i] \neq 0$$
Tricky: equation (5) is *not* a regression. Indeed, because the CEF of Y_i given P_i is linear, the regression of Y_i on P_i produces
$$E[Y_i|P_i = 1] - E[Y_i|P_i = 0] = \rho + \{E[\eta_i|P_i = 1] - E[\eta_i|P_i = 0]\},$$
in other words: the causal effect of interest plus ....

- Suppose, however, we observe controls X_i that satisfy a *conditional independence assumption* (CIA):
$$E[\eta_i|P_i, X_i] = E[\eta_i|X_i] = \gamma' X_i \tag{6}$$
Equivalently,
$$\eta_i = \gamma' X_i + u_i$$
where $E[u_i X_i] = 0$ *by construction*

- This leads to a regression model
$$Y_i = \alpha + \gamma' X_i + \rho P_i + u_i$$
The CIA, which links regression coefficients with causal parameters, makes regression causal (MHE 3.2 elaborates)
  - The X_i in DK02 is a vector of dummies for Barrons selectivity groups

4 An OVB Classic: Ability Bias

- Schooling coefficients with and without controls for family background, AFQT scores (a measure of ability), and occupation (MHE Table 3.2.1)

Table 3.2.1: Estimates of the returns to education, males

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupational dummies
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey)
The number in the first row is the coefficient on years of education in a weighted least squares regression of education on wages with the indicated controls. The number in parentheses is the associated standard error. The sample is restricted to males, weighted by NLSY sampling weights, and the sample size is 2434.
* Additional controls are mother's/father's years of education, and dummy variables for race and Census region.