Omitted Variable Bias

November 27-28, 2019

A quick revision of OVB framework

(1) Omitting one relevant variable

True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

We estimate: $y = \alpha_0 + \alpha_1 x_1 + \eta$

$\longrightarrow$ under usual CLRM assumptions:

$$E(\hat{\alpha}_1) = \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{var(x_1)}$$
$$= \beta_1 + \beta_2 \hat{\delta}_1$$

where $\hat{\delta}_1$ is the estimated coefficient of x_1 from the following regression: $x_2 = \delta_0 + \delta_1 x_1 + \xi$

What is the sign of the bias?

		$sign(cov(x_1, x_2))$	
		$+$	$-$
$sign(\beta_2)$	$+$	Positive Bias	Negative Bias
	$-$	Negative Bias	Positive Bias

(2) Omitting two relevant variables

True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

We estimate: $y = \alpha_0 + \alpha_1 x1 + \eta$

$\longrightarrow$ using the same logic as above, under usual CLRM assumptions:

$$\begin{aligned} E(\hat{\alpha_1}) &= \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{var(x_1)} + \beta_3 \frac{cov(x_1, x_3)}{var(x_1)} \\ &= \beta_1 + \beta_2 \hat{\delta}_1 + \beta_3 \hat{\gamma}_1 \end{aligned}$$

where $\hat{\delta}_1$ is the estimated coefficient of x_1 from the following regression: $x_2 = \delta_0 + \delta_1 x_1 + \delta_2 x_3 + \xi$, and $\hat{\gamma}_1$ is the estimated coefficient of x_1 from the following regression:
$x_3 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \nu$.

Example #1: Class Size & Student Performance (Angrist-Lavy, 1999)

- The paper studies the impact of class size on student performance.
- Empirical design: the Maimonides rule in Israel restricts the maximum class size to 40 students, so that if a school has 40+ students in a grade, the state provides funding for an additional teacher.
- Small and likely random difference in enrollment for a grade may hence create large differences in class size (2SLS regression exploits the variation in class size that is due to the law).
- Unit of observation: class; main regressors: class size, school-level index of students' socioeconomic status (*percent disadvantaged*), beginning-of-the year enrollment in the school for each grade; dependent variables: average scores in each class.

TABLE II

OLS ESTIMATES FOR 1991

	5th Grade					
	Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean score		74.3			67.3	
(s.d.)		(8.1)			(9.9)	
Regressors						
Class size	.221			.322		
	(.031)			(.039)		
Percent disadvantaged						
Enrollment						
Root MSE	7.54			9.36		
R^2	.036			.048		
N		2,019				

The unit of observation is the average score in the class. Standard errors are reported in parentheses. Sta

From column (1) to (2) (and from (4) to (5)):

$$E(\hat{\beta}_{cs}) = \beta_{cs} + \beta_{pd} \frac{Cov(cs, pd)}{Var(cs)}$$

$\rightarrow$ Do you expect to have a positive or a negative bias in the estimated coefficient for *class size* comparing (1) to (2) in reading comprehension and (4) to (5) in math scores?

TABLE II
OLS ESTIMATES FOR 1991

	5th Grade					
	Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean score		74.3			67.3	
(s.d.)		(8.1)			(9.9)	
Regressors						
Class size	.221	−.031		.322	.076	
	(.031)	(.026)		(.039)	(.036)	
Percent disadvantaged		−.350			−.340	
		(.012)			(.018)	
Enrollment						
Root MSE	7.54	6.10		9.36	8.32	
R^2	.036	.369		.048	.249	
N		2,019			2,018	

The unit of observation is the average score in the class. Standard errors are reported in parentheses. Sta

From column (1) to (3) (and from (4) to (6)):

$$E(\hat{\beta}_{cs}) = \beta_{cs} + \beta_{pd}\frac{Cov(cs, pd)}{var(cs)} + \beta_{enr}\frac{Cov(cs, enr)}{Var(cs)}$$

$\rightarrow$ Do you expect to have a positive or a negative bias in the estimated coefficient for *class size* comparing (1) to (3) in reading comprehension and (4) to (6) in math score?

TABLE II

OLS ESTIMATES FOR 1991

	5th Grade					
	Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean score		74.3			67.3	
(s.d.)		(8.1)			(9.9)	
Regressors						
Class size	.221	−.031	−.025	.322	.076	.019
	(.031)	(.026)	(.031)	(.039)	(.036)	(.044)
Percent disadvantaged		−.350	−.351		−.340	−.332
		(.012)	(.013)		(.018)	(.018)
Enrollment			−.002			.017
			(.006)			(.009)
Root MSE	7.54	6.10	6.10	9.36	8.32	8.30
R^2	.036	.369	.369	.048	.249	.252
N		2,019			2,018	

The unit of observation is the average score in the class. Standard errors are reported in parentheses. Sta

- OLS estimates without controls show a strong positive correlation between class size and student performance.
- Once we control for *percent disadvantaged* (PD) however, the positive association largely disappears, and in some cases, becomes negative.
- *Why?* Positive association between class size and performance largely accounted for by the association between larger classes and higher PD.
- Do you see any threat to identification here? (Selection bias both within and between schools)

Example #2: Does the marriage market matter for women? (Bursztyn et al., 2017)

- Do women “act wife"? That is, do they avoid career-enhancing actions because these actions could signal personality traits (like ambition) which are undesirable in the marriage market?
- Newly admitted MBA students filled out questionnaire on job preferences and personality traits, and a random selection of students thought their answers would be shared with classmates.
- When they believed classmates would not see responses, married and unmarried women answered similarly. However, when they believed their answers would be observed, single women reported lower desired compensation level, less willingness to travel, work fewer hours, and they also reported less professional ambition and tendency for leadership.

Outcome variable: class participation

	Grades Data				
	Participation	Exams and Problem Sets	Midterm Exam	Final Exam	Problem Sets
			A. Women, No Controls		
Unmarried	-5.72**	-0.44	0.81	-1.84	0.55*
	(2.30)	(1.02)	(1.35)	(1.74)	(0.30)
Dependent Variable Mean	77.64	79.66	82.06	68.22	95.57
Observations	561	561	561	561	561
R-squared	0.01	0.00	0.00	0.00	0.01

Control variables (that are added in next specification) are: GMAT score, and years of working experience.

But let's see what happens if we were instead to omit these variables (i.e., what we simply observe by looking at panel A):

$$E(\hat{\beta}_{um}) = \beta_{um} + \beta_{GMAT}\frac{Cov(um, GMAT)}{Var(um)} + \beta_{ye}\frac{Cov(um, ye)}{Var(um)}$$

Do you expect a positive or a negative bias for the estimated coefficient of *unmarried*?

Grades Data

	Participation	Exams and Problem Sets	Midterm Exam	Final Exam	Problem Sets
			A. Women, No Controls		
Unmarried	-5.72**	-0.44	0.81	-1.84	0.55*
	(2.30)	(1.02)	(1.35)	(1.74)	(0.30)
Dependent Variable Mean	77.64	79.66	82.06	68.22	95.57
Observations	561	561	561	561	561
R-squared	0.01	0.00	0.00	0.00	0.01
			B. Women, With Controls		
Unmarried	-5.59***	-0.81	-1.21	-1.10	0.05
	(2.15)	(0.95)	(1.38)	(1.62)	(0.25)
Dependent Variable Mean	77.64	80	82	68	96
Observations	544	544	544	544	544
R-squared	0.27	0.27	0.17	0.33	0.45

Outcome variable: performance in exam and problem sets

Grades Data

	Participation	Exams and Problem Sets	Midterm Exam	Final Exam	Problem Sets
			A. Women, No Controls		
Unmarried	-5.72**	-0.44	0.81	-1.84	0.55*
	(2.30)	(1.02)	(1.35)	(1.74)	(0.30)
Dependent Variable Mean	77.64	79.66	82.06	68.22	95.57
Observations	561	561	561	561	561
R-squared	0.01	0.00	0.00	0.00	0.01

Control variables (that are added in next specification) are: GMAT score, and years of working experience.

Again, what we observe in panel A of the table is given by:

$$E(\hat{\beta}_{um}) = \beta_{um} + \beta_{GMAT}\frac{Cov(um, GMAT)}{Var(um)} + \beta_{ye}\frac{Cov(um, ye)}{Var(um)}$$

How do you expect the sign of the bias to be? Of course, you should still state the same assumptions about the sign of the covariance terms, but since the dependent variable has changed, we now have to restate the signs of all *betas* in the equation.

Grades Data

	Participation	Exams and Problem Sets	Midterm Exam	Final Exam	Problem Sets
			A. Women, No Controls		
Unmarried	-5.72**	-0.44	0.81	-1.84	0.55*
	(2.30)	(1.02)	(1.35)	(1.74)	(0.30)
Dependent Variable Mean	77.64	79.66	82.06	68.22	95.57
Observations	561	561	561	561	561
R-squared	0.01	0.00	0.00	0.00	0.01
			B. Women, With Controls		
Unmarried	-5.59***	-0.81	-1.21	-1.10	0.05
	(2.15)	(0.95)	(1.38)	(1.62)	(0.25)
Dependent Variable Mean	77.64	80	82	68	96
Observations	544	544	544	544	544
R-squared	0.27	0.27	0.17	0.33	0.45

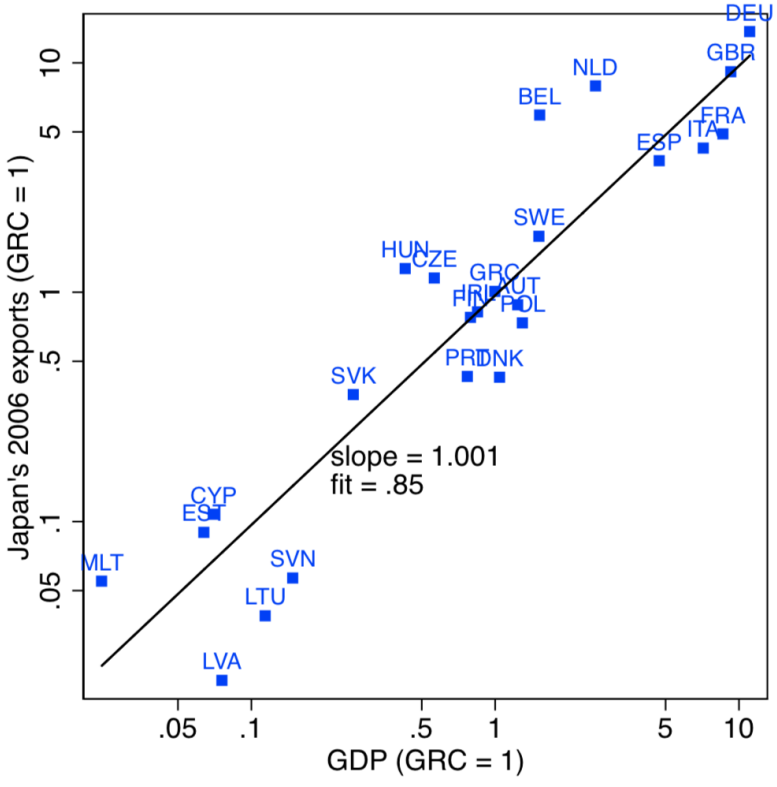
- Observational data results show that while unmarried women perform similarly to married women in class when their performance is kept private from classmates (like in exams and problem sets), they have significantly lower participation grades.
- Single women avoid actions that would help their careers because of marriage market concerns.
- The findings hence point to marriage market signaling as an additional explanation for gender differences in the labor market.

Example #3: Gravity Equation

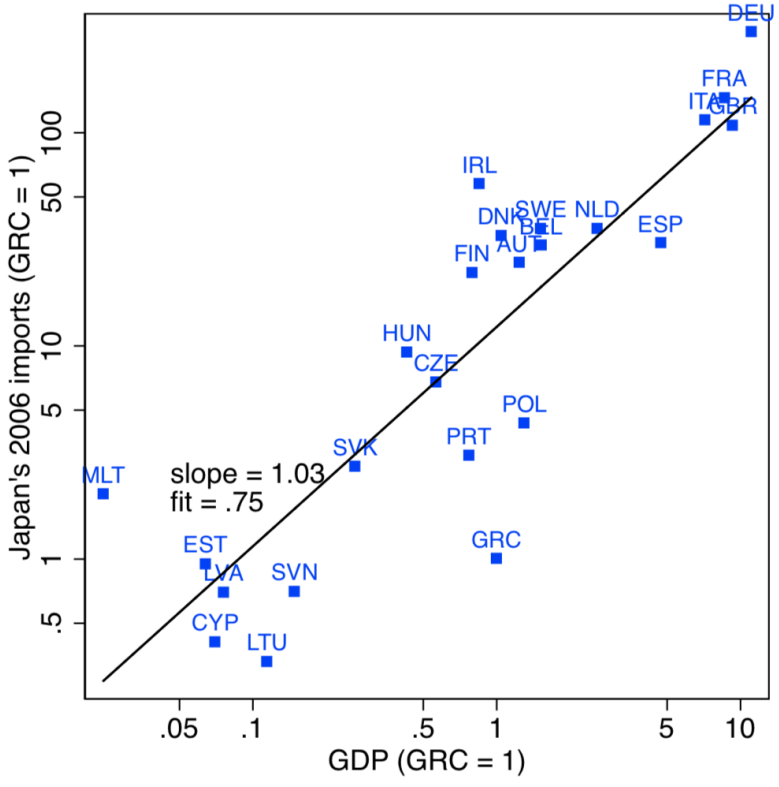
Fact #1

Figure 1 – Trade is proportional to size

(a) Japan's exports to EU, 2006



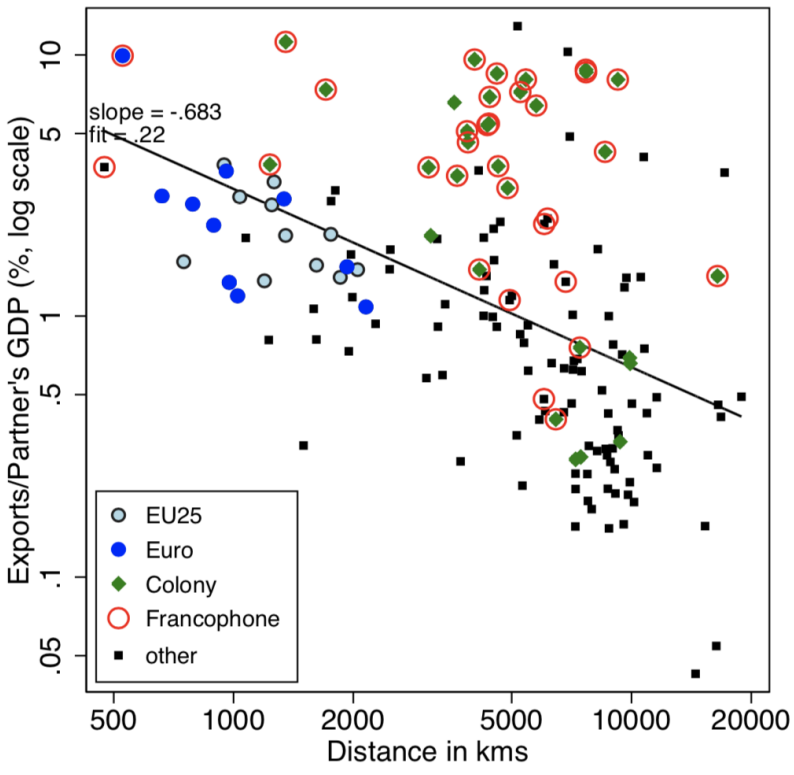
(b) Japan's imports from EU, 2006



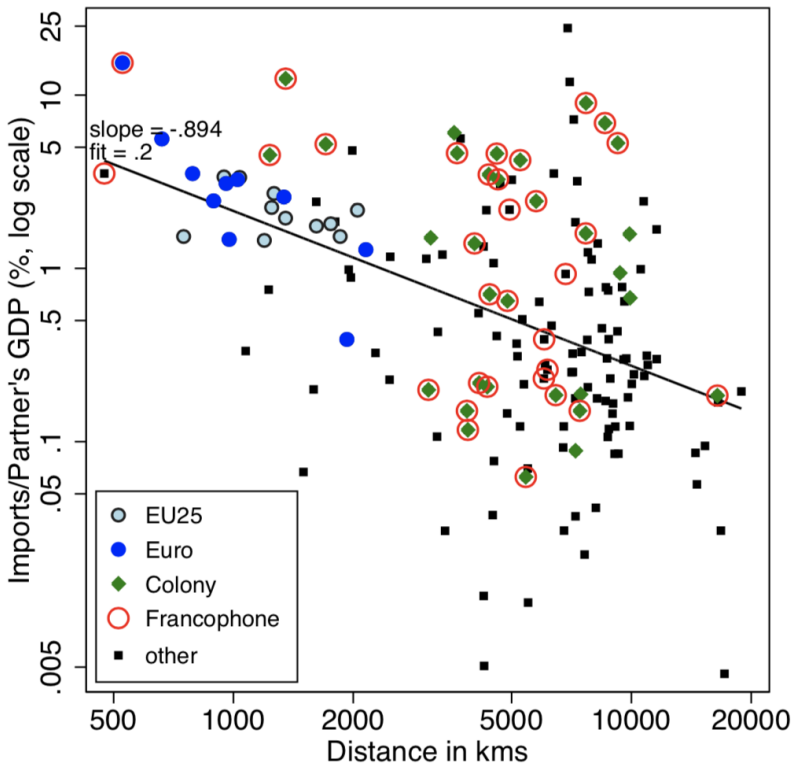
Fact #2

Figure 2 – Trade is inversely proportional to distance

(a) France's exports (2006)



(b) France's imports (2006)



Based on these facts, we run the following regression:

$$log(TradeFlows)_{ij} = \beta_0 + \beta_1 log(Distance)_{ij} + \beta_2 log(GDP)_i + \beta_3 log(GDP)_j + \epsilon_{ij}$$

Dep. Variable: Trade flows (log)	(1)
ldistw	-1.198***
	(0.003)
lgdp_o	0.871***
	(0.001)
lgdp_d	0.706***
	(0.001)
Constant	-4.173***
	(0.033)
Observations	624,145
R-squared	0.509

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Can you think of any omitted variable here?

Traditionally in the Trade literature, controls include: whether countries share a border (*contig*), having signed a Free Trade Agreement (*fta*), sharing a common language (*comlang*), presence of past or current colonial relationships (*colhist* and *colcur* respectively), etc.

Let's think about the sign of the bias for the estimated coefficient of distance, when our OVB is a measure of contiguity. *sign(bias)* depends on both the sign of the true coefficient for our omitted variable *contig*, and on the sign of the correlation between the omitted variable *contig* and *distance*. Do you expect to have a positive or a negative bias?

Dep. Variable: Trade flows (log)	(1)	(2)
ldistw	-1.198***	-1.124***
	(0.003)	(0.004)
lgdp_o	0.871***	0.871***
	(0.001)	(0.001)
lgdp_d	0.706***	0.706***
	(0.001)	(0.001)
contig		0.907***
		(0.015)
Constant	-4.173***	-4.828***
	(0.033)	(0.036)
Observations	624,145	624,145
R-squared	0.509	0.511

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

What about if we include the dummy variable *FTA*? Do you expect a positive or a negative bias for the estimated coefficient of *distance*? And for *contiguity*?

Dep. Variable: Trade flows (log)	(1)	(2)	(3)
ldistw	-1.198***	-1.124***	-1.065***
	(0.003)	(0.004)	(0.004)
lgdp_o	0.871***	0.871***	0.864***
	(0.001)	(0.001)	(0.001)
lgdp_d	0.706***	0.706***	0.700***
	(0.001)	(0.001)	(0.001)
contig		0.907***	0.893***
		(0.015)	(0.015)
fta			0.597***
			(0.012)
Constant	-4.173***	-4.828***	-5.243***
	(0.033)	(0.036)	(0.037)
Observations	624,145	624,145	624,145
R-squared	0.509	0.511	0.512

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

We now look at what changes if we consider sharing a common language as an additional omitted variable.

Do you expect the estimated coefficients for *distance*, and *fta* to be downward or upward biased? In other words, sharing a common language (which can be interpreted as a proxy for common culture/history), should positively or negatively impact the amount of trade between countries? But also, is *comlang* positively or negatively correlated with distance and FTA?

Dep. Variable: Trade flows (log)	(1)	(2)	(3)	(4)
ldistw	-1.198***	-1.124***	-1.065***	-1.043***
	(0.003)	(0.004)	(0.004)	(0.004)
lgdp_o	0.871***	0.871***	0.864***	0.880***
	(0.001)	(0.001)	(0.001)	(0.001)
lgdp_d	0.706***	0.706***	0.700***	0.713***
	(0.001)	(0.001)	(0.001)	(0.001)
contig		0.907***	0.893***	0.759***
		(0.015)	(0.015)	(0.015)
fta			0.597***	0.568***
			(0.012)	(0.012)
comlang				0.585***
				(0.008)
Constant	-4.173***	-4.828***	-5.243***	-5.812***
	(0.033)	(0.036)	(0.037)	(0.037)
Observations	624,145	624,145	624,145	624,145
R-squared	0.509	0.511	0.512	0.516

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Finally, let's look at the inclusion of both *colhist* and *colcur* in the regression (as if they were our two omitted relevant variables). What changes do you expect in the estimated coefficients for *distance*, *fta* and *comlang*?

Dep. Variable: Trade flows (log)	(1)	(2)	(3)	(4)	(5)
ldistw	-1.198***	-1.124***	-1.065***	-1.043***	-1.039***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
lgdp_o	0.871***	0.871***	0.864***	0.880***	0.869***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
lgdp_d	0.706***	0.706***	0.700***	0.713***	0.703***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
contig		0.907***	0.893***	0.759***	0.713***
		(0.015)	(0.015)	(0.015)	(0.015)
fta			0.597***	0.568***	0.603***
			(0.012)	(0.012)	(0.012)
comlang				0.585***	0.424***
				(0.008)	(0.008)
colhist					1.712***
					(0.015)
colcur					1.117***
					(0.075)
Constant	-4.173***	-4.828***	-5.243***	-5.812***	-5.656***
	(0.033)	(0.036)	(0.037)	(0.037)	(0.037)
Observations	624,145	624,145	624,145	624,145	624,145
R^2	0.509	0.511	0.512	0.516	0.522

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1