Multiple Regression: Omitted Varible Bias and Regression Anatomy

Fernando Hoces la Guardia
07/19/2022

Regression Journey

- Regression as Matching on Groups. Ch2 of MM up to page 68 (not included).
- Regression as Line Fitting and Conditional Expectation. Ch2 of MM, Appendix.
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79 and Appendix.
- Regression Inference, Binary Variables and Logarithms. Ch2 of MM, Appendix + others.

Regression Journey

- Regression as Matching on Groups. Ch2 of MM up to page 68 (not included).
- Regression as Line Fitting and Conditional Expectation. Ch2 of MM, Appendix.
- **Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79 and Appendix.**
- Regression Inference, Binary Variables and Logarithms. Ch2 of MM, Appendix + others.

Today's Lecture

- Omitted Variable Bias (very important)
- Regression Anatomy (not essential)

Omitted Variable Bias (OVB)

Omitted Variable Bias (OVB)

- We are back into the focus on causality!
- The most common regression version of selection bias is called omitted variable bias (OVB).
- Let's go back to the causal question from Dale and Krueger (2002) to motivate this concept.

Back to Earnings and Private/Public College Choice

- In moving from (1) to (2) we were controlling for SAT
- Including SAT had an effect on the coefficient of P_i
- Let's review the change from (4) to (5).
- Including SAT, after controlling for selectivity, seems to not change our causal estimates.
- Today we will formalize this relationship and it will help us understand how other unobservables might affect our causal estimates

TABLE 2.3
Private school effects: Average SAT score controls

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

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Can We Control for Everything?

- In our regressions we would like to control for how much resources had the family of each student.
- A proxy for resources is parental income, but it does not capture other aspects of being rich or poor in resources.
- One example is that two families could have the same income but different family sizes.
  - Imagine a family of 3 and a family of 6 with the same parental income. The larger family has far fewer resources to pay for higher tuition fees than the smaller family.
  - So even controlling for parental income, we would not have Other Things Equal.
- OVB helps us describe what happens when a relevant variable is omitted

What Can We Say About This Bias?

- To understand OVB, let's go back to the simple example of 5 students and two selectivity groups (A and B) for the effect of private college on earnings.

- First, assume that we have all the variables we need and then explore how omitting the variable group (A_i) will bias our estimates.

What Can We Say About This Bias?

- To understand OVB, let's go back to the simple example of 5 students and two selectivity groups (A and B) for the effect of private college on earnings.

- First, assume that we have all the variables we need and then explore how omitting the variable group (A_i) will bias our estimates.

- Let’s label the regression that has the variable (A_i) as the “long” regression (l) and the regression that does not have this variable as the “short” regression (s).

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

Short and Long Regressions: Simple Example 1/2

- From the toy example data on Table 2.1 of MM, we have already compute the regression estimates $\alpha^l = 40,000$, $\beta^l = 10,000$, and $\gamma^l = 60,000$
- Any ideas on how to compute the regression coefficient β^s?

TABLE 2.1
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

Note: Enrollment decisions are highlighted in gray.

Short and Long Regressions: Simple Example 1/2

- From the toy example data on Table 2.1 of MM, we have already compute the regression estimates $\alpha^l = 40,000$, $\boldsymbol{\beta^l = 10,000}$, and $\gamma^l = 60,000$
- Any ideas on how to compute the regression coefficient β^s?
- As we saw yesterday β^s is the simple difference in earnings (Y_i) between treatment $(P_i = 1)$ and control $(P_i = 0)$. From table 2.1 (focusing only on groups A and B) we have that $\boldsymbol{\beta^s = 20,000}$.
- Omitting A_i leads to bias = $\beta^s - \beta^l = 10,000$

TABLE 2.1
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

Note: Enrollment decisions are highlighted in gray.

Short and Long Regressions: Simple Example 2/2

- OVB is define as the difference between effect omitting (on short) minus the effect not omitting (on long). $OVB \equiv \beta^s - \beta^l$. In this toy example is 10k.

- The source of this bias is in attributing to P_i the difference between groups (A and B) captured by A_i.

- We can now establish more formally the two components that connect the coefficients from the long and short regression:

  1. The relationship between the omitted variable (A_i) and treatment (P_i).
  2. The relationship between the outcome (Y_i) and the omitted variable (A_i). This relationship is givent by the parameter γ in the long regression.

OVB Formula: General

$$\text{Effect of included in short} = \text{Effect of included in long} +$$
$$\text{Effect of omitted on outcome, in long} \times$$
$$\text{Relationhip between omitted and included} \times$$

OVB Formula: General

$$\text{Effect of included in short} = \text{Effect of included in long} +$$
$$\text{Effect of omitted on outcome, in long} \times$$
$$\text{Relationhip between omitted and included} \times$$

> *"Short equals long plus effect of omitted in long (on outcome) times the regression of omitted on included"*

OVB Formula: General (Causal)

$$\begin{aligned} \text{Effect of treatment in short} = \text{Effect of treatment in long} + \\ \text{Effect of omitted on outcome, in long} \times \\ \text{Relationhip between omitted and treatment} \times \end{aligned}$$

> *"Short equals long plus effect of omitted in long (on outcome) times the regression of omitted on included"*

OVB Formula in Example 1/3

$$\text{Effect of } P_i \text{ in short} = \text{Effect of } P_i \text{ in long} +$$
$$\text{Effect of } A_i \text{ on } Y_i \text{ (in long)} \times$$
$$\text{Relationship between } A_i \text{ and } P_i$$

> *"Short equals long plus effect of omitted in long (on outcome) times the regression of omitted on included"*

OVB Formula in Example 2/3

$$\text{Effect of } P_i \text{ in short} = \text{Effect of } P_i \text{ in long} + \\ \text{Effect of } A_i \text{ on } Y_i \text{ (in long)} \times \\ \text{Relationship between } A_i \text{ and } P_i$$

OVB Formula in Example 2/3

$$\text{Effect of } P_i \text{ in short} = \text{Effect of } P_i \text{ in long} +$$
$$\text{Effect of } A_i \text{ on } Y_i \text{ (in long)} \times$$
$$\text{Relationship between } A_i \text{ and } P_i$$

$$\beta^s = \beta^l +$$
$$\text{Relationship between } A_i \text{ and } P_i \times$$
$$\gamma$$

$$OVB = \beta^s - \beta^l = \text{Relationship between } A_i \text{ and } P_i \times \gamma$$

OVB Formula in Example 2/3

$$\text{Effect of } P_i \text{ in short} = \text{Effect of } P_i \text{ in long} +$$
$$\text{Effect of } A_i \text{ on } Y_i \text{ (in long)} \times$$
$$\text{Relationship between } A_i \text{ and } P_i$$

$$\beta^s = \beta^l +$$
$$\text{Relationship between } A_i \text{ and } P_i \times$$
$$\gamma$$

$$OVB = \beta^s - \beta^l = \text{Relationship between } A_i \text{ and } P_i \times \gamma$$

The relationship between A_i and P_i can be estimated using an auxiliary regression:

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

OVB Formula in Example 3/3

$$OVB = \beta^s - \beta^l = \pi_1 \times \gamma$$

OVB Formula in Example 3/3

$$OVB = \beta^s - \beta^l = \pi_1 \times \gamma$$

- We know $\gamma = 60,000$, how could we estimate π_1?

OVB Formula in Example 3/3

$$OVB = \beta^s - \beta^l = \pi_1 \times \gamma$$

- We know $\gamma = 60,000$, how could we estimate π_1?
- $\pi_1 = \overline{A}_1 - \overline{A}_0 = 2/3 - 1/2 = 0.1667$
- $OVB = \beta^s - \beta^l = 0.1667 \times 60,000 = 10,000$
- The same we obtained by computing $\beta^s - \beta^l$ before!
- The key idea is that we care about the bias that we cannot observe ($\beta^s - \beta^l$), but we can investigate it by thinking about plausible values for the relationship between omitted and included (π_1) and the effect of omitted in long (γ).

OVB in Dale and Krueger Study 1/3

- Let's discuss how the omitted variable "Family Size" (FS_i) could be generating some OVB.
- What would be the short equation in this case (hint: is not that short)?

OVB in Dale and Krueger Study 1/3

- Let's discuss how the omitted variable "Family Size" (FS_i) could be generating some OVB.

- What would be the short equation in this case (hint: is not that short)?

$$lnY_i = \alpha^s + \beta^s P_i + \sum_{j=1}^{150} \gamma_j^s GROUP_{ji} + \delta_1^s SAT + \delta_2^s lnPI_i + e_i^s$$

- What would be the long equation in this case (hint: long basically means "longer" than short)?

OVB in Dale and Krueger Study 1/3

- Let's discuss how the omitted variable "Family Size" (FS_i) could be generating some OVB.

- What would be the short equation in this case (hint: is not that short)?

$$lnY_i = \alpha^s + \beta^s P_i + \sum_{j=1}^{150} \gamma_j^s GROUP_{ji} + \delta_1^s SAT + \delta_2^s lnPI_i + e_i^s$$

- What would be the long equation in this case (hint: long basically means "longer" than short)?

$$lnY_i = \alpha^l + \beta^l P_i + \sum_{j=1}^{150} \gamma_j^l GROUP_{ji} + \delta_1^l SAT + \delta_2^l lnPI_i + \lambda FS_i + e_i^l$$

OVB in Dale and Krueger Study 2/3

$$lnY_i = \alpha^s + \beta^s P_i + \sum_{j=1}^{150} \gamma_j^s GROUP_{ji} + \delta_1^s SAT + \delta_2^s lnPI_i + e_i^s$$

$$lnY_i = \alpha^l + \beta^l P_i + \sum_{j=1}^{150} \gamma_j^l GROUP_{ji} + \delta_1^l SAT + \delta_2^l lnPI_i + \lambda FS_i + e_i^l$$

- What would be the auxiliary regression in this case?

OVB in Dale and Krueger Study 2/3

$$lnY_i = \alpha^s + \beta^s P_i + \sum_{j=1}^{150} \gamma_j^s GROUP_{ji} + \delta_1^s SAT + \delta_2^s lnPI_i + e_i^s$$

$$lnY_i = \alpha^l + \beta^l P_i + \sum_{j=1}^{150} \gamma_j^l GROUP_{ji} + \delta_1^l SAT + \delta_2^l lnPI_i + \lambda FS_i + e_i^l$$

- What would be the auxiliary regression in this case?

$$FS_i = \pi_0 + \pi_1 P_i + \sum_{j=1}^{150} \pi_{3,j} GROUP_{ji} + \pi_4 SAT + \pi_5 lnPI_i + u_i$$

$$OVB = \beta^s - \beta^l = \pi_1 \times \lambda$$

- Time to think about the sign and magnitude of π_1 and λ in this case.

OVB in Dale and Krueger Study 2/3

$$lnY_i = \alpha^s + \beta^{\mathbf{s}} P_i + \sum_{j=1}^{150} \gamma_j^s GROUP_{ji} + \delta_1^s SAT + \delta_2^s lnPI_i + e_i^s$$

$$lnY_i = \alpha^l + \beta^{\mathbf{l}} P_i + \sum_{j=1}^{150} \gamma_j^l GROUP_{ji} + \delta_1^l SAT + \delta_2^l lnPI_i + \lambda FS_i + e_i^l$$

- What would be the auxiliary regression in this case?

$$FS_i = \pi_0 + \pi_{\mathbf{1}} P_i + \sum_{j=1}^{150} \pi_{3,j} GROUP_{ji} + \pi_4 SAT + \pi_5 lnPI_i + u_i$$

$$OVB = \beta^{\mathbf{s}} - \beta^{\mathbf{l}} = \pi_{\mathbf{1}} \times \lambda$$

- Time to think about the sign and magnitude of π_1 and λ in this case.

OVB in Dale and Krueger Study 3/3

- π_1 is likely to be negative and large in magnitude.
- λ higher family sizes might lead to less resources per children and this could have a negative effect on future earnings. Hence $\lambda < 0$
- Hence omitting FS_i will probably lead to a OVB that is positive (estimated effects are larger than true effects) positive.

OVB in Dale and Krueger Study 3/3

- π_1 is likely to be negative and large in magnitude.
- λ higher family sizes might lead to less resources per children and this could have a negative effect on future earnings. Hence $\lambda < 0$
- Hence omitting FS_i will probably lead to a OVB that is positive (estimated effects are larger than true effects) positive.
- Let's think of other potentially omitted variables: received tutoring? parental education?

OVB in Dale and Krueger Study 3/3

- π_1 is likely to be negative and large in magnitude.
- λ higher family sizes might lead to less resources per children and this could have a negative effect on future earnings. Hence $\lambda < 0$
- Hence omitting FS_i will probably lead to a OVB that is positive (estimated effects are larger than true effects) positive.
- Let's think of other potentially omitted variables: received tutoring? parental education?
- One thing that is interesting about this particular example is that most stories that you can think have either $\lambda < 0, \pi_1 < 0$ or $\lambda > 0, \pi_1 > 0$ leading us to suspect that the estimated effect of private college in a regression are likely to be overestimated.

Robustness to Inclusion/Exclussion of Regressors

- In regression, we can never know if we have control for enough variables to eliminate OVB/selection bias.
- Given this, we should always ask how much do the estimated coefficients change when including new variables.
- Confidence on regression estimates of causal effects grow when treatment effects are insensitive to the inclusion of new variables.

Robustness: Dale and Krueger Study 1/2

- Moving from column (1) to (2):
  - (1) was omitting SAT_i, and (2) is the long version of (1)
  - $OVB = \beta^s - \beta^l = 0.212 - 0.152 = 0.06$
  - How about computing the same but using the OVB formula?
  - We need the auxiliary regression (page 76 of MM): $\pi_1 = 1.165$
  - Where is the "effect of omitted in long" (λ) ?

TABLE 2.3
Private school effects: Average SAT score controls

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

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Robustness: Dale and Krueger Study 1/2

- Moving from column (1) to (2):
  - (1) was omitting SAT_i, and (2) is the long version of (1).
  - $OVB = \beta^s - \beta^l = 0.212 - 0.152 = 0.06$
  - How about computing the same but using the OVB formula?
  - We need the auxiliary regression (page 76 of MM): $\pi_1 = 1.165$
  - Where is the "effect of omitted in long" (λ) ?
  - $\lambda = 0.051$
  - $OVB = \pi_1 \times \lambda = 1.165 \times 0.051 = 0.06$!

TABLE 2.3
Private school effects: Average SAT score controls

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

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Robustness: Dale and Krueger Study 2/2

- Moving from column (4) to (5):
  - (4) was omitting SAT_i, and (5) is the long version of (4)
  - $OVB = \beta^s - \beta^l = 0.034 - 0.031 = 0.003$
  - How about computing the same but using the OVB formula?
  - We need the auxiliary regression (page 76 of MM): $\pi_1 = 0.066$
  - Where is the (λ)?

Table 2.3
Private school effects: Average SAT score controls

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

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Robustness: Dale and Krueger Study 2/2

- Moving from column (4) to (5):
  - (4) was omitting SAT_i, and (5) is the long version of (4)
  - $OVB = \beta^s - \beta^l = 0.034 - 0.031 = 0.003$
  - How about computing the same but using the OVB formula?
  - We need the auxiliary regression (page 76 of MM): $\pi_1 = 0.066$
  - Where is the (λ)?
  - $\lambda = 0.036$
  - $OVB = \pi_1 \times \lambda = 0.066 \times 0.036 = 0.0024!$
  - Differences are due to rounding of small numbers
  - Most of the change comes from π_1

TABLE 2.3
Private school effects: Average SAT score controls

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

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Proof of OVB Formula

$$\beta_1^s = \frac{Cov(X_{1i}, Y_{1i})}{Var(X_{1i})}$$

Substitute for Y_i using equation for long.

Proof of OVB Formula

$$\beta_1^s = \frac{Cov(X_{1i}, Y_{1i})}{Var(X_{1i})}$$

$$\beta^s = \frac{Cov(X_{1i}, \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l)}{Var(X_{1i})}$$

$$= \frac{\beta^l Var(X_{1i}) + \gamma Cov(X_{1i}, X_{2i}) + Cov(X_{1i}, e_i^l)}{Var(X_{1i})}$$

Substitute for $\boldsymbol{Y_i}$ using equation for long.

But what is a key property of any residuals?

Proof of OVB Formula

$$\beta_1^s = \frac{Cov(X_{1i}, Y_{1i})}{Var(X_{1i})}$$

Substitute for Y_i using equation for long.

$$\beta^s = \frac{Cov(X_{1i}, \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l)}{Var(X_{1i})}$$

$$= \frac{\beta^l Var(X_{1i}) + \gamma Cov(X_{1i}, X_{2i}) + Cov(X_{1i}, e_i^l)}{Var(X_{1i})}$$

But what is a key property of any residuals?

$$\beta^s = \frac{\beta^l Var(X_{1i}) + \gamma Cov(X_{1i}, X_{2i})}{Var(X_{1i})}$$

$$= \beta^l + \gamma \frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})}$$

What is that last term?
(think auxiliary regression)

Proof of OVB Formula

$$\beta_1^s = \frac{Cov(X_{1i}, Y_{1i})}{Var(X_{1i})}$$

Substitute for Y_i using equation for long.

$$\beta^s = \frac{Cov(X_{1i}, \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l)}{Var(X_{1i})}$$

$$= \frac{\beta^l Var(X_{1i}) + \gamma Cov(X_{1i}, X_{2i}) + Cov(X_{1i}, e_i^l)}{Var(X_{1i})}$$

But what is a key property of any residuals?

$$\beta^s = \frac{\beta^l Var(X_{1i}) + \gamma Cov(X_{1i}, X_{2i})}{Var(X_{1i})}$$

$$= \beta^l + \gamma \frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})}$$

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

What is that last term?
(think auxiliary regression)

Acknowledgments

- Ed Rubin's Graduate Econometrics
- MM