Lecture 4 Regression

Filipa Sá

King's College London

Semester 1, 2019/20

Readings — regression

- Angrist and Pischke chapter 2
- Dale and Krueger (2002), "Estimating the Payoffs of Attending a More Selective College: an Application of Selection on Observables and Unobservables", *Quarterly Journal of Economics*, vol. 117, no. 4, pages 1491-1527.
- Background: Wooldridge chapters 2-4, 6, 7 and 9

What does regression do?

- When we don't have random assignment, we can use regression
- This compares the treatment and control subjects who have the same observed characteristics
- Causal inference based on regression assumes that, once we have made observed variables equal across treatment and control groups, selection bias from things that we can't observe is mostly eliminated as well.
- What if there are important things that we cannot observe and control for?
    - In that case, regression does not eliminate selection bias
    - The selection bias generated by inadequate controls is called **omitted variables bias (OVB)**

Omitted Variable Bias (OVB)

- Return to the Dale and Krueger study of the effect of attending a private university on earnings
- Students who attend private university may differ from those who attend public university in ways that cannot be measured. For example, they may be more able and motivated.
- Suppose that the "long regression", which controls for ability, is:

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

- Suppose that we do not have a good measure of ability and estimate the "short regression":

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

- Because students who attend private school tend to have higher ability, we expect $\beta^s > \beta^l$

Omitted Variable Bias (OVB)

- We can derive a formula for the relation between the coefficients in the short and long regressions
- The OLS slope coefficient for the short regression is given by:

$$\beta^s = \frac{Cov(Y_i, P_i)}{V(P_i)}$$

- Substituting the long model for Y_i in this expression:

$$\begin{aligned}
\beta^s &= \frac{Cov(Y_i, P_i)}{V(P_i)} \\
&= \frac{Cov(\alpha^l + \beta^l P_i + \gamma A_i + \varepsilon_i^l, P_i)}{V(P_i)} \\
&= \frac{\beta^l V(P_i) + \gamma Cov(A_i, P_i) + Cov(\varepsilon_i^l, P_i)}{V(P_i)} \\
&= \beta^l + \pi_1 \gamma
\end{aligned}$$

Omitted Variable Bias (OVB)

- The OVB equation:

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

- π_1 is the coefficient on P_i in a regression of A_i on P_i (called *auxiliary regression*):

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

- This equation tells us that the relation between the coefficients in the short and long regression has two components:
  - The relation between the omitted variable (A_i) and the treatment variable (P_i), measured by π_1
  - The relation between the omitted variable (A_i) and the outcome variable (Y_i), measured by the coefficient γ in the long regression

Omitted Variable Bias (OVB)

- In words, this equation tells us that:
  Effect of P_i in short = Effect of P_i in long + Relation between A_i and $P_i \times$ Effect of A_i in long
- The omitted variable bias is the difference between β^s and β^l and is given by:
  OVB = Relation between A_i and $P_i \times$ Effect of A_i in long

Side note - properties of covariances

- To derive the OVB equation, we use the following properties of covariance:
  - $Cov(\alpha + \beta X_1, X_2) = Cov(\alpha, X_2) + \beta Cov(X_1, X_2)$
  - $Cov(\alpha, X) = 0$
  - $Cov(X, X) = V(X)$
  - $Cov(\varepsilon_i^l, P_i) = 0$ because residuals are uncorrelated with the regressors that made them (ε_i^l is the residual of a regression that includes P_i)

Omitted Variable Bias (OVB)

- Dale and Krueger attempt to measure ability by including SAT scores, and selectivity group dummies for the universities that students applied for and the universities they were admitted to. They also control for parental income.
- Estimated equation:

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + \varepsilon_i$$

- Suppose that an omitted variable in this equation is family size (FS_i).
  OVB = Relation between FS_i and $P_i \times$ Effect of FS_i in long

Omitted Variable Bias (OVB)

- Differences between Harvard and U-Mass graduates arise in part from differences in family size between the two groups of students (this is the relation between FS_i and P_i) and the fact that coming from a smaller family is associated with higher earnings (this is the relation between FS_i and Y_i in the long regression)

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

- π_1 is the coefficient on P_i in a regression of FS_i on P_i and the other control variables (*auxiliary regression*):

$$FS_i = \pi_0 + \pi_1 P_i + \sum_{j=1}^{150} \theta_j GROUP_{ji} + \pi_2 SAT_i + \pi_3 \ln PI_i + u_i$$

- λ is the coefficient on FS_i in the long regression:

$$\ln Y_i = \alpha^l + \beta^l P_i + \sum_{j=1}^{150} \gamma_j^l GROUP_{ji} + \delta_1^l SAT_i + \delta_2^l \ln PI_i + \lambda FS_i + \varepsilon_i^l$$

Omitted Variable Bias (OVB)

- The OVB in this case is likely to be positive:
  - Private school students tend to come from smaller families ($\pi_1 < 0$)
  - Students from smaller families are likely to earn more, no matter where they go to college ($\lambda < 0$)
  - The product of these two terms is positive

Regression sensitivity analysis

- Because we can never be sure that we have controlled for all important variables to eliminate selection bias, it is important to check that the estimation results are not very sensitive (i.e., are *robust*) to the inclusion of additional controls.
- Table 2.3 reports the results of a regression with and without controlling for average SAT score of the set of schools the student applied to and the number of applications sent (selection controls). The number of applications can be important because weaker students apply on average to fewer and less strong schools than stronger applicants.
- Without controls, private school graduates earn 21.2% more than private school graduates. This falls to 13.9% when controlling for SAT score and parental income. The effect disappears after controlling for the average SAT score of schools applied to and the number of applications sent.

Regression sensitivity analysis

- Table 2.5 reports the results of the auxiliary regressions
- Suppose that we omit SAT scores
  - We can calculate the OVB in a regression without selection controls in two ways:
    - Comparing columns (1) and (2) in Table 2.3

$$OVB = Short - Long = 0.212 - 0.152 = 0.06$$

    - Using the OVB formula

$$\begin{aligned} OVB &= Relation\ between\ SAT_i\ and\ P_i \times Effect\ of\ SAT_i\ in\ long \\ &= 1.165 \times 0.051 = 0.06 \end{aligned}$$

The relation between SAT_i and P_i comes from column (1) in Table 2.5 and the effect of SAT_i in long comes from column (2) in Table 2.3.

Regression sensitivity analysis

- Suppose that we omit SAT scores, but include selection controls. The OVB is:
  - Comparing columns (4) and (5) in Table 2.3

$$OVB = Short - Long = 0.034 - 0.031 = 0.003$$

  - Using the OVB formula

$$\begin{aligned} OVB &= Relation\ between\ SAT_i\ and\ P_i \times Effect\ of\ SAT_i\ in\ long \\ &= 0.066 \times 0.036 = 0.003 \end{aligned}$$

    - The relation between SAT_i and P_i comes from column (3) in Table 2.5 and the effect of SAT_i in long comes from column (5) in Table 2.3.

- Once we have controlled for the average SAT score of the schools applied to and the number of applications sent, there is almost no bias from omitting SAT scores.

Regression sensitivity analysis

TABLE 2.3
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)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
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)

Regression sensitivity analysis

TABLE 2.5
Private school effects: Omitted variables bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	
High school rank missing		.556			−.030	
		(.102)			(.023)	
Athlete		−.318			.037	
		(.147)			(.016)	
Average SAT score of schools applied to ÷ 100			.777			.063
			(.058)			(.014)
Sent two applications			.252			.020
			(.077)			(.010)
Sent three applications			.375			.042
			(.106)			(.013)
Sent four or more applications			.330			.079
			(.093)			(.014)

Bad controls

- We have just seen that including more controls can reduce the selection bias. But more controls is not always better.
- For example, we are interested in measuring whether a college degree increases earnings. People can work in two occupations: managerial and manual. Should we control for occupation in a regression of wages on schooling?
- Notation:
  - W_i is a dummy equal to 1 for managerial jobs (0 for manual)
  - Y_i is earnings - outcome variable
  - C_i is a dummy equal to 1 for college graduates - treatment variable
- Having a college degree affects both earnings and occupation:

$$\begin{aligned} Y_i &= C_i Y_{1i} + (1 - C_i) Y_{0i} \\ W_i &= C_i W_{1i} + (1 - C_i) W_{0i} \end{aligned}$$

Bad controls

- Suppose that college graduation is randomly assigned. We can estimate the causal effect of C_i on either Y_i or W_i:

$$\begin{aligned} E[Y_i|C_i &= 1] - E[Y_i|C_i = 0] = E[Y_{1i} - Y_{0i}] \\ E[W_i|C_i &= 1] - E[W_i|C_i = 0] = E[W_{1i} - W_{0i}] \end{aligned}$$

- In practice, we can estimate these causal effects by regressing Y_i and W_i on C_i

Bad controls

- Bad controls means that a comparison of earnings conditional on W_i does not have a causal interpretation. Suppose we consider only those in managerial jobs:

$$\begin{aligned} E[Y_i|W_i &= 1, C_i = 1] - E[Y_i|W_i = 1, C_i = 0] = \\ E[Y_{1i}|W_{1i} &= 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0] \end{aligned}$$

- Because college graduation is randomly assigned, we have:

$$\begin{aligned} E[Y_{1i}|W_{1i} &= 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0] = \\ E[Y_{1i}|W_{1i} &= 1] - E[Y_{0i}|W_{0i} = 1] = \\ E[Y_{1i} - Y_{0i}|W_{1i} &= 1] + \{E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]\} \end{aligned}$$

- The *causal effect* is $E[Y_{1i} - Y_{0i}|W_{1i} = 1]$ and the *selection bias* is $E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]$

Bad controls

- It is not clear in which direction the bias goes. But if we look at $W_{0i} = 1$, we are looking at especially bright people who got a job as a manager without the benefit of a college degree. If we look at $W_{1i} = 1$, we have a weaker group of those who become managers only by virtue of completing college. In that case, we would expect the selection bias to be negative. A regression of wages on education controlling for occupation would underestimate the returns to a college degree.
- Timing matters:
  - Bad controls are variables that are themselves outcomes and could be the dependent variable in the regression
  - Good controls are variables that can be considered fixed at the time treatment was determined

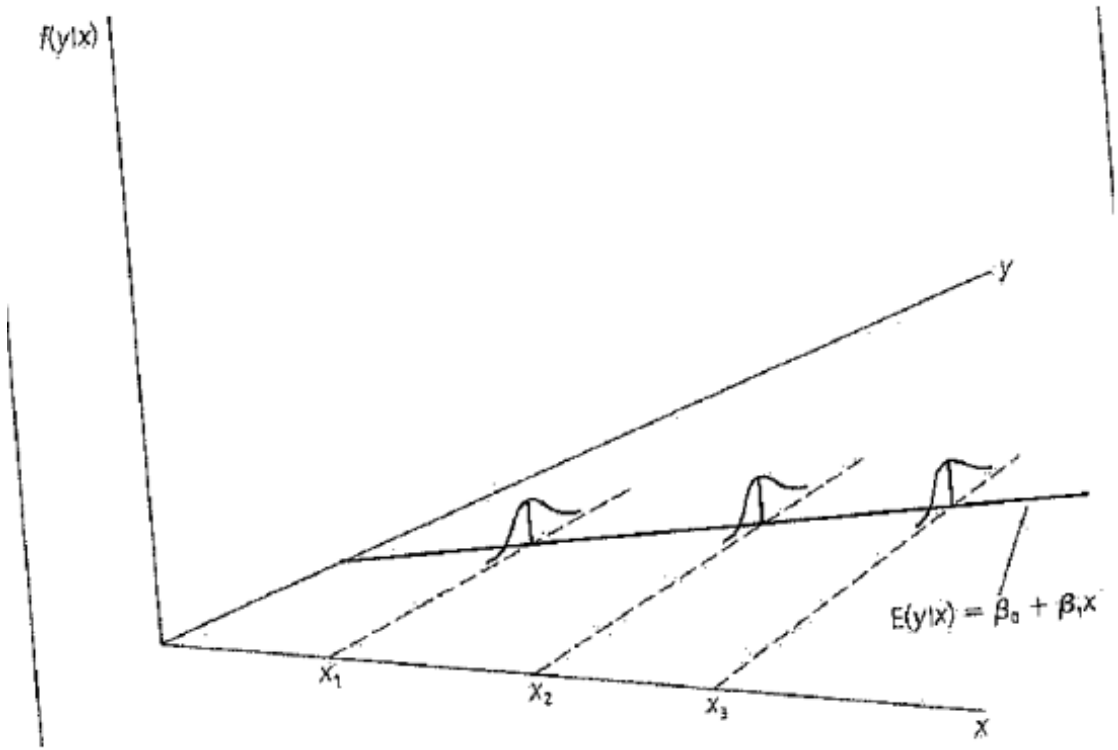
Side note: robust standard errors

- The standard errors reported by Stata using the **regress** command assume that the residuals are unrelated to the regressors. In econometrics, this assumption is known as **homoskedasticity**.
- This assumption may not be satisfied. There is an alternative way to calculate standard errors to take into account possible correlation between the residuals and the regressors (know as **heteroskedasticity**). These alternative standard errors are called **robust** and are now commonly used in empirical work.
- **You should always use robust standard errors**. You do this by adding an option in the Stata regress command:
  reg y x, vce(robust)

Side note: robust standard errors

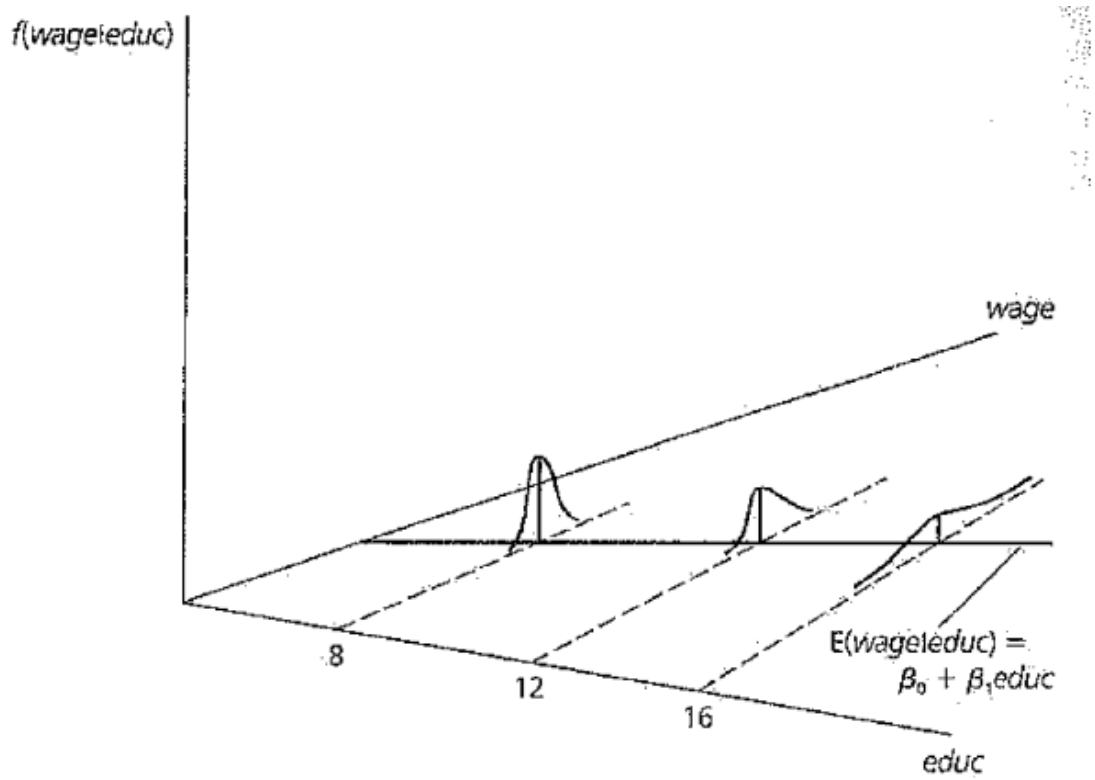
- Homoskedasticity implies that the variability of the dependent variable does not depend on the regressor

$$
\begin{aligned}
y &= \beta_0 + \beta_1 X + u \\
u &= y - \beta_0 - \beta_1 X \\
Var(u|x) &= Var(y|x)
\end{aligned}
$$



Side note: robust standard errors

- In some cases, the homoskedasticity assumption is not likely to hold. For example, the variability of earnings may be larger for people who attend private universities because they have a wider variety of job opportunities. This is why we use robust standard errors.



Appendix

- Show that $Cov(\varepsilon_i^l, P_i) = 0$
- This comes directly from the derivation of the OLS estimator that we saw last week
- Consider the long regression:

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

what OLS does is minimise:

$$E[Y_i - \alpha^l - \beta^l P_i - \gamma A_i]^2$$

- Differentiate with respect to β^l and set it equal to zero:

$$\begin{aligned} -2E[P_i(Y_i - \alpha^l - \beta^l P_i - \gamma A_i)] &= 0 \\ E(P_i \varepsilon_i^l) &= 0 \end{aligned}$$

- Note that $Cov(\varepsilon_i^l, P_i) = E(P_i \varepsilon_i^l)$ because $E(\varepsilon_i^l) = 0$