Lecture 3 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.

A tale of two colleges

- In the US, students can choose to go to a private college and pay over £20,000 a year in fees or go to a public university in their home state and pay less than £7,000.
- Is the difference worth it?
- What we would like to know is how much a 40 year old graduate that went to Harvard would have earned if he had gone to the University of Massachusetts instead.
- If we compare the average earnings of those who went to Harvard and those who went to U-Mass, we find that Harvard graduates earn significantly more.
- But this is comparing apples with oranges — Harvard grads are a special and select group (typically have higher high school grades and are more motivated).

A tale of two colleges

- But to disentangle the pure Harvard effect, we would need to hold constant all other factors that matter for earnings — gender, high school scores, diligence, family connections etc. Some of these factors are not observable.
- A way to get at the causal effect is to look at students who were admitted to both Harvard and U-Mass, but decided to go to U-Mass (for example, because they won a scholarship or have a relative who went to U-Mass, etc.).
- Because they were admitted to Harvard, they have the motivation and ability to do well, but chose to go to U-Mass because of some chance event.

A tale of two colleges

- This is what Dale and Krueger (2002) do — instead of identifying all factors that affect earnings and college choice, they look at the characteristics of the colleges that students applied to and were admitted.
- Data from College & Beyond (C&B)
  - Information on more than 14,000 students who enrolled in a group of moderately to highly selective US colleges and universities.
  - They enrolled in 1976. We have information collected before they went to college (SAT — standardised test used for college admissions in the US) and again in 1996.

Regression ingredients

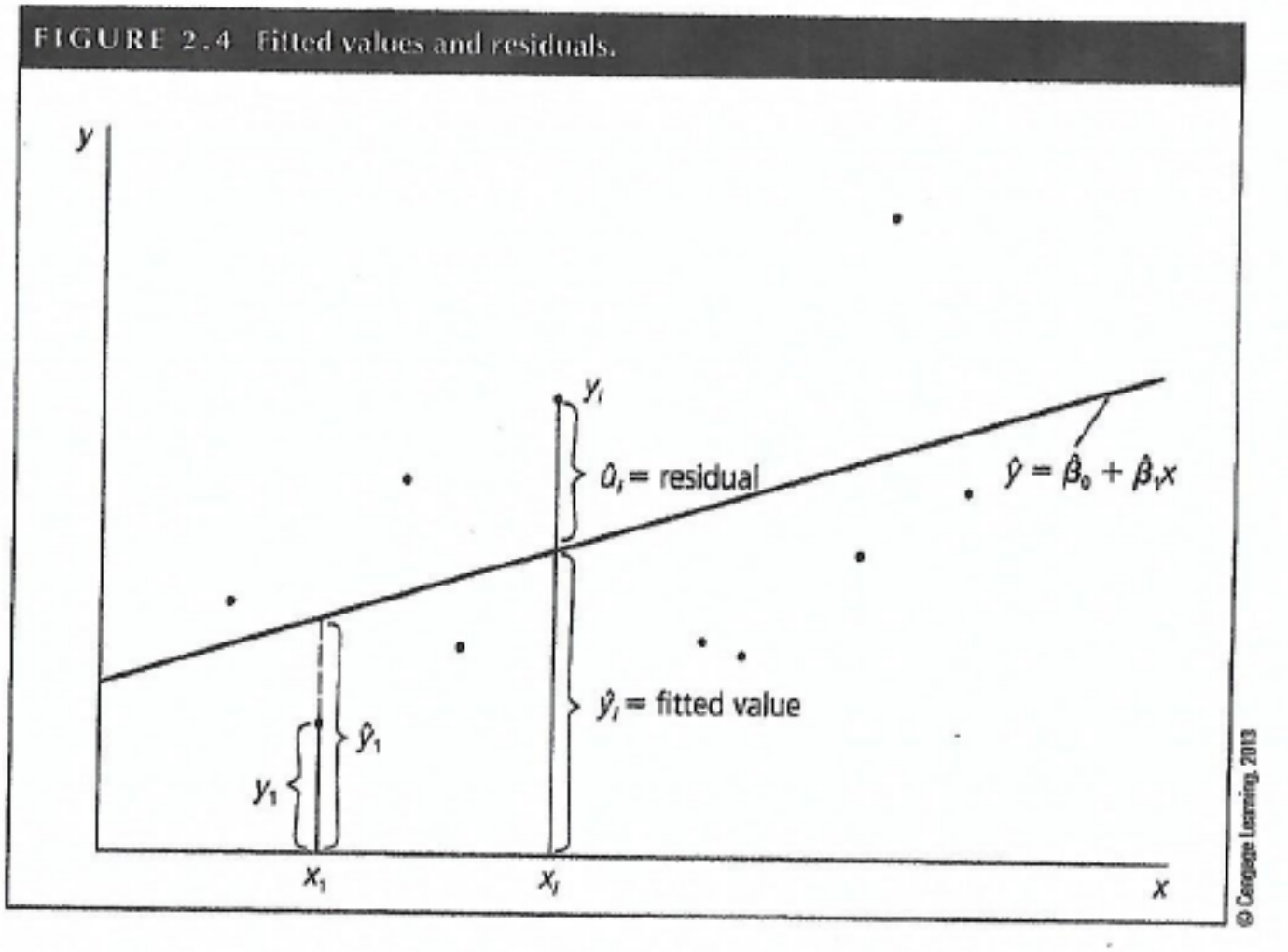
- **Dependent variable** or **outcome variable** (Y_i) — earnings of student i later in life
- **Treatment variable** (P_i) — dummy variable that indicates students who attended a private college or university
- **Control variables** (A_i) — variables that identify the set of schools to which students applied and were admitted
- Regression:

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

Regression ingredients

- Regression coefficients:
  - **Intercept** α
  - **Causal effect of treatment** β
  - Effect of the control variable γ
- The **residual**, also called **error term**, is ε_i. They are the difference between Y_i and the fitted values generated by the model.
- An **Ordinary Least Squares (OLS)** regression chooses values for the coefficients that minimise the sum of the squares of the residuals.

What OLS does



What OLS does

- In a bivariate regression:

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

what OLS does is minimise:

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

- Differentiate wrt α and β and set it equal to zero to obtain the OLS coefficient:

$$\beta = \frac{C(Y_i, P_i)}{V(P_i)}$$

where $C(Y_i, P_i) = E[(Y_i - E(Y_i))(P_i - E(P_i))]$ and $V(P_i) = E(P_i - E(P_i))^2$

- Exercise: derive this expression for β.

Side note: regressions with logs

- Suppose we use the log of earnings (instead of the level) as the dependent variable and focus on the bivariate regression:

$$\ln Y_i = \alpha + \beta P_i + \varepsilon_i$$

- Suppose that β captures the causal effect of going to private school, i.e., the difference between the earnings of students who go to private school ($\ln Y_{1i}$) and the earnings they would have if they had not gone to private school ($\ln Y_{0i}$)

$$\begin{aligned}
\beta &= \ln Y_{1i} - \ln Y_{0i} = \ln\left(\frac{Y_{1i}}{Y_{0i}}\right) \\
&= \ln\left(1 + \frac{Y_{1i} - Y_{0i}}{Y_{0i}}\right) \\
&= \ln(1 + \%\Delta Y) \approx \%\Delta Y
\end{aligned}$$

Side note: regressions with logs

- The regression coefficient β is approximately equal to the percentage change in earnings
- To calculate the exact percentage change, we use the exponential:

$$\begin{aligned} \frac{Y_{1i}}{Y_{0i}} &= \exp(\beta) \\ \frac{Y_{1i} - Y_{0i}}{Y_{0i}} &= \exp(\beta) - 1 \end{aligned}$$

A tale of two colleges

- Regression model:

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

  - The dependent variable is the natural log of earnings, so the coefficient β can be interpreted as the percentage change in earnings from attending a private university

A tale of two colleges

- $GROUP_{ji}$ is a set of 150 selectivity group dummies (indicator variables that take the value 1 if individual i is in group j and 0 otherwise)
  - These indicators are based on the classification of universities into Most Competitive, Highly Competitive, Very Competitive, Competitive, Less Competitive and Noncompetitive (Barron's ranking)
  - The selectivity group dummies reflect the universities that students applied for and the universities they were admitted to
  - For example, one indicator ($GROUP_{1i}$) captures students who applied and were admitted to three Highly Competitive schools, another indicator ($GROUP_{2i}$) captures students who applied to two Highly Competitive schools and one Most Competitive school and were admitted to one of each type, etc.
- Control variables: SAT score (SAT_i) and log of parental income ($\ln PI_i$)

A tale of two colleges

From Mostly Harmless Econometrics: An Empiricist's Companion by Joshua D. Angrist and Jörn-Steffen Pischke. © 2009 Princeton University Press. Used by permission. All rights reserved.

TABLE 2.2
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

A tale of two colleges

- Without controls, earnings are 13.5% higher for private school students
- Controlling for SAT score and family earnings reduces the private school premium
- Introducing the selectivity group dummies (which proxy for ability, motivation, etc.) eliminates the premium altogether

Appendix: deriving the OLS estimator

- In a bivariate regression:

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

what OLS does is minimise:

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

- Differentiate wrt α and set it equal to zero:

$$\begin{aligned} -2E[Y_i - \alpha - \beta P_i] &= 0 \\ \alpha &= E(Y_i) - \beta E(P_i) \qquad ((1)) \end{aligned}$$

- Differentiate wrt β and set it equal to zero:

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

Appendix: deriving the OLS estimator

- Substitute equation (1) in equation (2):

$$E(P_i Y_i) - E(Y_i)E(P_i) + \beta E(P_i)^2 - \beta E(P_i^2) = 0$$

- Solve for β to obtain the OLS coefficient:

$$\begin{aligned} \beta &= \frac{E(P_i Y_i) - E(Y_i)E(P_i)}{E(P_i^2) - E(P_i)^2} \\ \beta &= \frac{C(Y_i, P_i)}{V(P_i)} \end{aligned}$$

where $C(Y_i, P_i) = E[(Y_i - E(Y_i))(P_i - E(P_i))]$ and $V(P_i) = E(P_i - E(P_i))^2$