Regression I

Lecture 5

Quantitative Methods for Political Economy
7SSPN107

Recap: hypothesis tests and the t-statistic

We are often interested in population parameters. We might, for example, want to test a claim about the population mean:

$$H_0 : \mu_Y = \mu_{Y,0} \qquad \text{versus} \qquad H_1 : \mu_Y \neq \mu_{Y,0}.$$

The t-statistic measures the estimate's distance from the null value in standard errors:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{\text{SE}(\bar{Y})}, \qquad \text{SE}(\bar{Y}) = \frac{s_Y}{\sqrt{n}}.$$

Under H_0 and when n is large, the t-statistic is approximately standard normally distributed: $t \approx N(0, 1)$.

Recap: p-values and significance levels

The p-value is the probability of a test statistic taking a value that is at least as extreme as the observed value, assuming H_0 is true.

$$p \approx 2\Phi(-|t^{act}|) \qquad \text{for a two-sided test.}$$

We can follow a rule that says that we should reject a hypothesis if the p-value is less than a prespecified significance level, e.g., $\alpha = 0.05$. This is equivalent to $|t^{act}| > 1.96$ using the normal approximation.

Failing to reject does not establish that H_0 is true. Conversely, we may falsely reject H_0 while it is true: a Type I error. Its probability is approximately the chosen significance level α.

Recap: three ideas we will use

Conditional expectation, $E(Y \mid X = x)$, is the mean of Y among units with $X = x$.

Covariance describes how X and Y move together:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)], \quad \mu_X = E(X),\ \mu_Y = E(Y).$$

An estimator is a rule applied to a sample. Its value changes across samples, giving it a sampling distribution.

Regression uses these ideas to estimate how the mean of Y varies with X.

A short history of regression

Francis Galton (1886) was a British statistician, meteorologist, inventor of the dog whistle, eugenicist, and responsible for naming regression.

He studied the heights of parents and their adult children and found an approximately linear relationship.

On average, the children of tall or short parents were less tall or short than their parents: their heights tended towards the population average. This is regression towards the mean.

TABLE I.

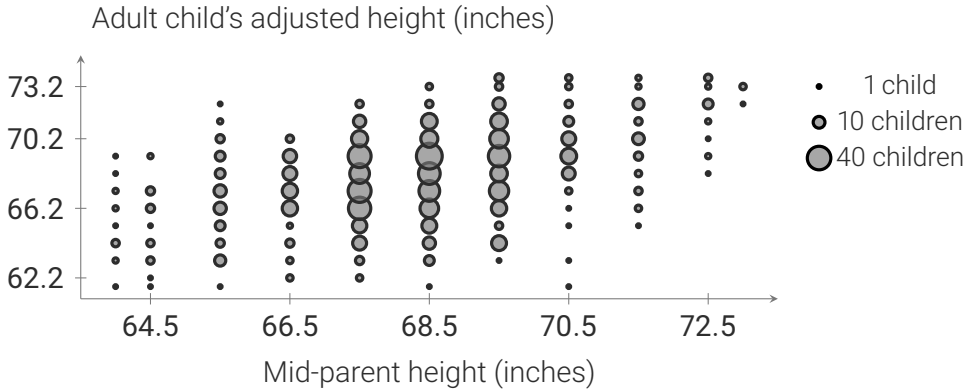
NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

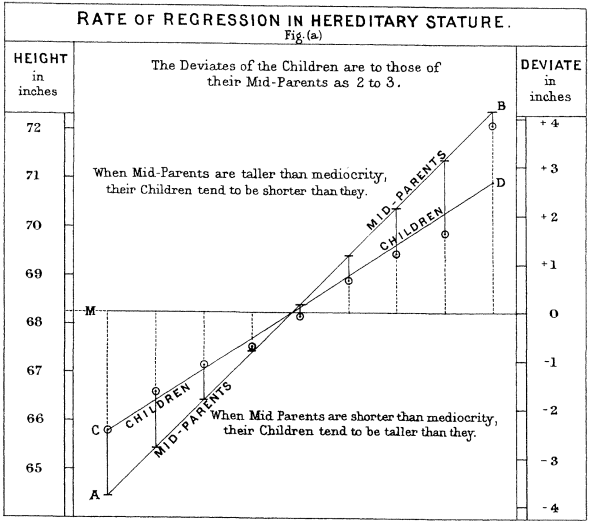
Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Plotting the data in Table I



Regression to the mean



4.

Linear Regression with One Regressor

Causal inference and prediction

Education may affect hourly earnings by equipping students with skills or qualifications.

- Causal inference: how do earnings change if someone's education were to be increased by an extra year?
- Prediction: what earnings would we predict for a worker with a given education level?

A regression line can help with both questions.

A causal interpretation needs assumptions about how education is determined. Random assignment of education would help make workers comparable.

4.1

The linear regression model

Population regression line

For workers aged 30, let X_i be completed years of education and Y_i gross hourly earnings (£/hour).

The population regression line describes average earnings among workers for a given level of education:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{population regression line}} + u_i.$$

The error u_i measures the gap between earnings predicted by the population regression line and actual earnings.

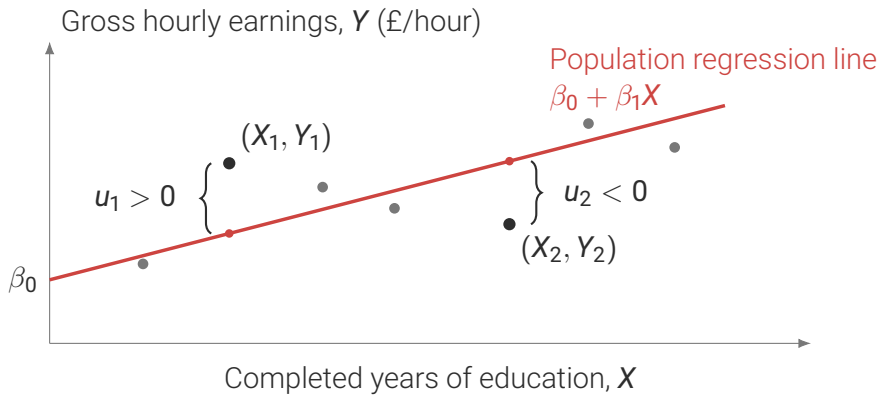
Y is the outcome (dependent variable); X is the regressor (independent variable).

What do the intercept, slope and error mean?

β_0 and β_1 are unknown population coefficients.

- Intercept β_0: the line's predicted hourly earnings at zero years of education.
- Slope β_1: the change in expected hourly earnings associated with one more year of education.
- Error u_i: a worker's actual earnings minus the population line.

Education and earnings



Exercise 1: interpreting a slope

Suppose the education-earnings population line has slope $\beta_1 = 2$.

1. What is the expected difference in hourly earnings for a worker with 3 more years of education?

$$\text{earnings difference} = 2 \times 3 = 6 \text{ £/hour.}$$

2. Does this show that giving someone 3 more years of education would raise their hourly earnings by £6?

No. It compares workers who may differ in dimensions other than education. Prior ability or family background, for example, may affect both education and earnings.

4.2

Ordinary least squares

Using a sample to estimate the population line

Suppose we are given data on these five workers. How should we estimate β_0 and β_1?

Worker	Education X_i (years)	Earnings Y_i (£/hour)
A	10	14
B	12	12
C	14	24
D	16	20
E	18	30

Ordinary Least Squares

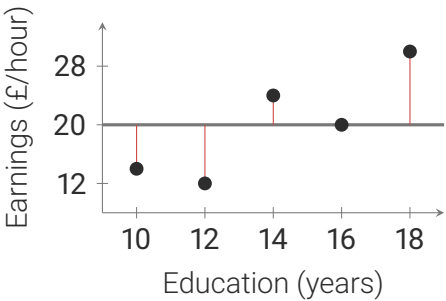
Ordinary least squares (OLS) is a method for estimating β_0 and β_1. It chooses betas that minimise the sum of squared differences between observed and predicted earnings:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right]^2.$$

Recall that previously we simply relied on the sample mean to predict Y_i. Now we allow that prediction to depend on X_i.

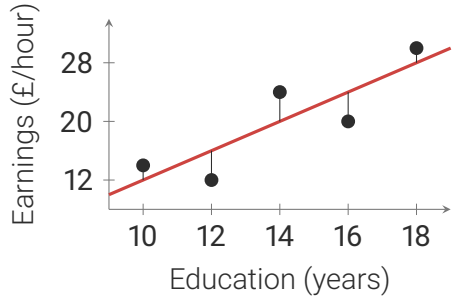
OLS minimises the sum of squared vertical gaps

Constant prediction: **20**



Total squared gaps: **216**

Prediction: $-8 + 2X$



Total squared gaps: **56**

The OLS estimates of the slope and intercept

The OLS coefficients are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{4.5}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \tag{4.6}$$

Exercise 2: calculating the OLS line

For our hypothetical workers, $\bar{X} = 14$ and $\bar{Y} = 20$.

Worker	$X_i - \bar{X}$	$Y_i - \bar{Y}$	Product	$(X_i - \bar{X})^2$
A	-4	-6	24	16
B	-2	-8	16	4
C	0	4	0	0
D	2	0	0	4
E	4	10	40	16
Sum	0	0	80	40

1. Calculate the slope.

$$\hat{\beta}_1 = \frac{80}{40} = 2.$$

Exercise 2: calculating the OLS line

We found $\hat{\beta}_1 = 2$. Recall $\bar{X} = 14$ and $\bar{Y} = 20$.

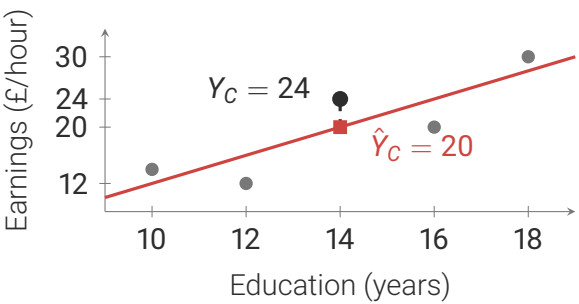
2. Calculate the intercept and write the fitted line.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 20 - 2(14) = -8.$$

$$\hat{Y} = -8 + 2X.$$

The fitted line predicts £2/hour higher earnings for each additional year of education.

Fitted values and residuals: worker C



For C, $X_C = 14$ years:

$$\hat{Y}_C = -8 + 2(14) = 20.$$

$$\hat{u}_C = 24 - 20 = 4.$$

Earnings are £4/hour above the fitted line.

Exercise 3: a fitted value and a residual

Worker B has $X_B = 12$ years of education and earns $Y_B = 12$ £/hour. Our estimated line is $\hat{Y} = -8 + 2X$.

1. Calculate the fitted earnings and residual. Interpret the sign.

$$\hat{Y}_B = -8 + 2(12) = 16, \qquad \hat{u}_B = 12 - 16 = -4.$$

Hourly earnings are £4 below the fitted value.

2. Is the population error also -4 £/hour?

We cannot tell: the true population line is unknown.

4.3

Fit and prediction

How much variation does our line account for?

Sum of squares	Squared gaps added up	Total
Total (TSS)	Observed earnings minus mean	**216**
Explained (ESS)	Fitted earnings minus mean	**160**
Residuals (SSR)	Observed minus fitted earnings	**56**

$$\underbrace{216}_{\text{total}} = \underbrace{160}_{\text{accounted for by line}} + \underbrace{56}_{\text{remaining}} .$$

R-squared and the SER

R-squared: share of variation

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{160}{216} \approx 0.741.$$

The line accounts for 74.1% of earnings variation in our sample.

SER: spread around the line

$$\text{SER} = \sqrt{\frac{\text{SSR}}{n-2}}.$$

Measured in £/hour.

In-sample and out-of-sample prediction

Our fitted line is $\hat{Y} = -8 + 2X$.

In-sample

Worker C was part of the sample that was used to estimate the line. With $X_C = 14$:

$$\hat{Y}_C = -8 + 2(14) = 20.$$

Fitted earnings: £20/hour.
Actual earnings: £24/hour.
Residual: £4/hour.

Out-of-sample

A new worker aged 30 has 15 education years:

$$\hat{Y}_{\text{new}} = -8 + 2(15) = 22.$$

Predicted earnings: £22/hour.

Exercise 4: applying the method to exam scores

We have data on study hours in the week before an exam (X) and scores out of 100 (Y) for a sample of 100 students.

The means are $\bar{X} = 5$ and $\bar{Y} = 60$. The sum of $(X_i - \bar{X})(Y_i - \bar{Y})$ is 800; the sum of $(X_i - \bar{X})^2$ is 200.

1. Find the OLS line.

$$\hat{\beta}_1 = \frac{800}{200} = 4, \quad \hat{\beta}_0 = 60 - 4(5) = 40, \quad \hat{Y} = 40 + 4X.$$

2. Predict the score for a student who studied 8 hours. Interpret the slope.

$$\hat{Y} = 40 + 4(8) = 72 \text{ points}.$$

One more study hour is associated with a 4-point higher score. We cannot assume this association causally without additional assumptions.

4.4

The least-squares assumptions

OLS and causality I

If education were assigned randomly to individuals, then comparing the earnings of individuals with different levels of education would give us the causal effect of education on earnings.

In reality, education is not assigned randomly. Can we still interpret the (positive) association between education and earnings causally?

In general: no, individuals with high and low levels of education may differ in characteristics other than education that also affect earnings (e.g., access to parents' network).

OLS and causality II

In order to interpret the association between education and earnings causally, we need to make a few assumptions.

Recall that the error term u_i measures all factors that affect Y_i other than X_i.

Assumption 1: the combined effect of other factors influencing pay is, on average, equal to zero across education levels:

$$E(u_i \mid X_i) = 0.$$

This rules out that other factors that affect Y_i vary systematically with X_i.

Two further assumptions support the sampling theory

Assumption 2: observations (X_1, Y_1) are independently and identically distributed.

In other words: we assume that workers are sampled independently from the same population we want to study.

Assumption 3: large outliers are unlikely.

This condition helps justify large-sample approximations that we rely on.

4.5

Sampling distributions of OLS estimators

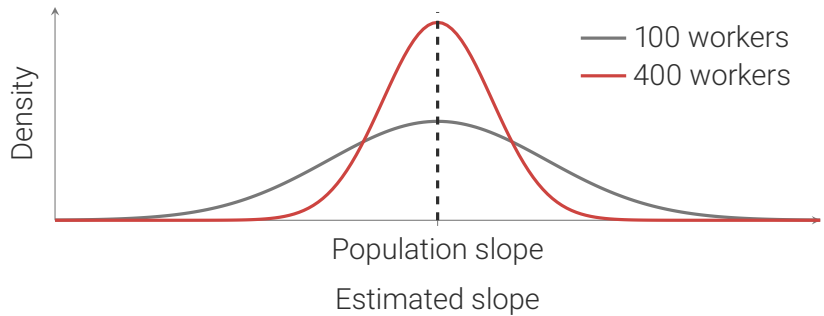
OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables

$\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from randomly drawn samples and are therefore themselves random variables with a sampling distribution.

Under the least-squares assumptions, OLS estimators are:

- Unbiased: over repeated samples, estimates are correct on average.
- Consistent: larger samples concentrate the estimates closer to the population coefficients.
- In large samples, each estimator's sampling distribution is approximately normal.

Larger samples make estimates more precise



Exercise 6: sample size and uncertainty

Suppose the sampling standard deviation of our estimated slope is about 0.40 with 100 workers.

1. What is the sampling standard deviation approximately with 400 workers?

$$0.40 \div \sqrt{400/100} = 0.40 \div 2 = 0.20.$$

2. Does the SER also halve?

No. More observations may improve our estimate of the population line, but workers' earnings will still be spread around the population line, and the SER measures this spread.