Regression II

Lecture 6

Quantitative Methods for Political Economy
7SSPN107

Recap: the population regression model

The population regression line describes the average outcome for a given value of the regressor.

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

Recap: the population regression model

The population regression line describes the average outcome for a given value of the regressor.

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

With $E(u_i \mid X_i) = 0$, the conditional mean is

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i.$$

β_0 is the intercept and β_1 is the slope. The error u_i captures all other factors influencing Y_i.

Recap: ordinary least squares

OLS estimates β_0 and β_1 by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the squared difference between predicted and observed outcomes

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Recap: ordinary least squares

OLS estimates β_0 and β_1 by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the squared difference between predicted and observed outcomes

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

The solutions are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Recap: ordinary least squares

OLS estimates β_0 and β_1 by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimise the squared difference between predicted and observed outcomes

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

The solutions are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

The sample regression line gives fitted values and residuals:

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \qquad \hat{u}_i = Y_i - \widehat{Y}_i.$$

Recap: fit and residual spread

The total spread around the mean can be expressed as $\mathrm{TSS} = \mathrm{ESS} + \mathrm{SSR}$

$$\text{Total sum of squares:} \quad \mathrm{TSS} = \sum_i (Y_i - \bar{Y})^2,$$

$$\text{Explained sum of squares:} \quad \mathrm{ESS} = \sum_i (\widehat{Y}_i - \bar{Y})^2,$$

$$\text{Sum of squared residuals:} \quad \mathrm{SSR} = \sum_i (Y_i - \widehat{Y}_i)^2.$$

Recap: fit and residual spread

The total spread around the mean can be expressed as $\mathrm{TSS} = \mathrm{ESS} + \mathrm{SSR}$

$$\text{Total sum of squares:} \quad \mathrm{TSS} = \sum_i (Y_i - \bar{Y})^2,$$

$$\text{Explained sum of squares:} \quad \mathrm{ESS} = \sum_i (\widehat{Y}_i - \bar{Y})^2,$$

$$\text{Sum of squared residuals:} \quad \mathrm{SSR} = \sum_i (Y_i - \widehat{Y}_i)^2.$$

R^2 is the fraction of sample variation in Y explained by the fitted line:

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{SSR}}{\mathrm{TSS}}.$$

Recap: fit and residual spread

The total spread around the mean can be expressed as $\mathrm{TSS} = \mathrm{ESS} + \mathrm{SSR}$

$$\text{Total sum of squares:} \quad \mathrm{TSS} = \sum_i (Y_i - \bar{Y})^2,$$

$$\text{Explained sum of squares:} \quad \mathrm{ESS} = \sum_i (\widehat{Y}_i - \bar{Y})^2,$$

$$\text{Sum of squared residuals:} \quad \mathrm{SSR} = \sum_i (Y_i - \widehat{Y}_i)^2.$$

R^2 is the fraction of sample variation in Y explained by the fitted line:

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{SSR}}{\mathrm{TSS}}.$$

The standard error of the regression estimates the standard deviation of the error:

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

Recap: assumptions and sampling distributions

- Zero conditional mean: $E(u_i \mid X_i) = 0$. The other factors influencing Y_i average to zero at every value of X_i.

Recap: assumptions and sampling distributions

- Zero conditional mean: $E(u_i \mid X_i) = 0$. The other factors influencing Y_i average to zero at every value of X_i.
- Independent, identically distributed observations: the pairs (X_i, Y_i) are independent draws from the same population.

Recap: assumptions and sampling distributions

- Zero conditional mean: $E(u_i \mid X_i) = 0$. The other factors influencing Y_i average to zero at every value of X_i.
- Independent, identically distributed observations: the pairs (X_i, Y_i) are independent draws from the same population.
- Large outliers are unlikely: X_i and Y_i have finite kurtosis.

Recap: assumptions and sampling distributions

- Zero conditional mean: $E(u_i \mid X_i) = 0$. The other factors influencing Y_i average to zero at every value of X_i.
- Independent, identically distributed observations: the pairs (X_i, Y_i) are independent draws from the same population.
- Large outliers are unlikely: X_i and Y_i have finite kurtosis.

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a random sample, so they are random variables. Different samples give different estimates.

Recap: assumptions and sampling distributions

- Zero conditional mean: $E(u_i \mid X_i) = 0$. The other factors influencing Y_i average to zero at every value of X_i.
- Independent, identically distributed observations: the pairs (X_i, Y_i) are independent draws from the same population.
- Large outliers are unlikely: X_i and Y_i have finite kurtosis.

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a random sample, so they are random variables. Different samples give different estimates.

Under the least-squares assumptions, the OLS estimators are unbiased and consistent, with approximately normal sampling distributions in large samples.

5.1

Testing hypotheses about one of the regression coefficients

From estimation to hypothesis testing

Suppose that a sample of 500 workers gives

$$\widehat{Y} = -6 + 2X,$$

where X is education in years and Y is hourly earnings in £/hour.

From estimation to hypothesis testing

Suppose that a sample of 500 workers gives

$$\widehat{Y} = -6 + 2X,$$

where X is education in years and Y is hourly earnings in £/hour.

Is this evidence that the population slope differs from a specified value $\beta_{1,0}$?

From estimation to hypothesis testing

Suppose that a sample of 500 workers gives

$$\widehat{Y} = -6 + 2X,$$

where X is education in years and Y is hourly earnings in £/hour.

Is this evidence that the population slope differs from a specified value $\beta_{1,0}$?

We might be interested in whether the population slope is zero or not, meaning that an additional year of education makes no difference in for mean earnings.

The hypotheses and the t-statistic

Recall the general form of the *t*-statistic:

$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error of the estimate}}.$$

The hypotheses and the t-statistic

Recall the general form of the *t*-statistic:

$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error of the estimate}}.$$

For a two-sided test of the population slope:

$$H_0 : \beta_1 = \beta_{1,0} \qquad \text{versus} \qquad H_1 : \beta_1 \neq \beta_{1,0}.$$

The hypotheses and the t-statistic

Recall the general form of the *t*-statistic:

$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error of the estimate}}.$$

For a two-sided test of the population slope:

$$H_0 : \beta_1 = \beta_{1,0} \qquad \text{versus} \qquad H_1 : \beta_1 \neq \beta_{1,0}.$$

Under H_0, with valid standard errors and a large sample, $t \approx N(0, 1)$.

Computing the standard error

The standard error is an estimator of the standard deviation of the sampling distribution of $\hat{\beta}_1$.

For our sample of $n = 500$ workers, suppose that

$$\sum_i (X_i - \bar{X})^2 = 500,$$
$$\sum_i (X_i - \bar{X})^2 \hat{u}_i^2 = 159\,360.$$

Computing the standard error

The standard error is an estimator of the standard deviation of the sampling distribution of $\hat{\beta}_1$.

For our sample of $n = 500$ workers, suppose that

$$\sum_i (X_i - \bar{X})^2 = 500,$$
$$\sum_i (X_i - \bar{X})^2 \hat{u}_i^2 = 159\,360.$$

Using the standard-error formula:

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{n}{n-2} \frac{\sum_i (X_i - \bar{X})^2 \hat{u}_i^2}{\left[\sum_i (X_i - \bar{X})^2\right]^2}} = \sqrt{\frac{500}{498} \frac{159\,360}{500^2}} = 0.80.$$

Computing the regression t-statistic

For the slope, the general formula becomes

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)}.$$

Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Computing the regression t-statistic

For the slope, the general formula becomes

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)}.$$

Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

With $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$:

$$t = \frac{2 - 0}{0.80} = 2.50.$$

Computing the regression t-statistic

For the slope, the general formula becomes

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)}.$$

Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

With $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$:

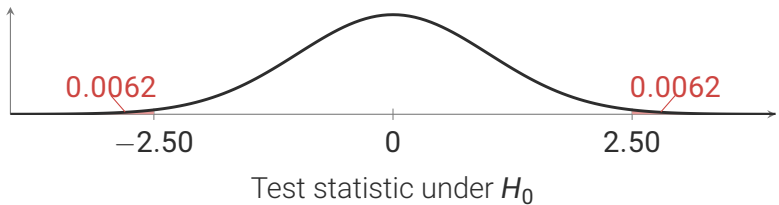
$$t = \frac{2 - 0}{0.80} = 2.50.$$

The estimated slope is 2.50 standard errors above zero.

At the 5% level, a two-sided large-sample test rejects H_0 when $|t| > 1.96$.

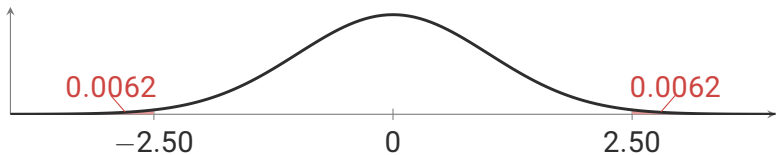
The p-value and the testing decision

Assuming H_0, the p-value is the probability of an estimate $\hat{\beta}_1$ at least as far from $\beta_{1,0}$ as the observed estimate.



The p-value and the testing decision

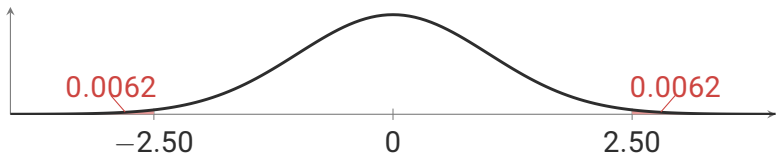
Assuming H_0, the p-value is the probability of an estimate $\hat{\beta}_1$ at least as far from $\beta_{1,0}$ as the observed estimate.



$$p = P_{H_0}(|t| \geq 2.50)$$

The p-value and the testing decision

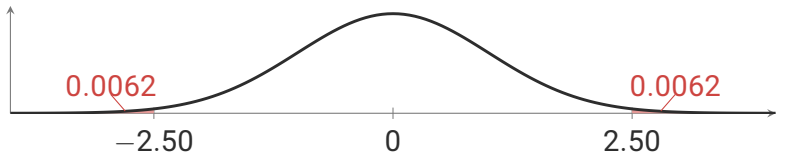
Assuming H_0, the p-value is the probability of an estimate $\hat{\beta}_1$ at least as far from $\beta_{1,0}$ as the observed estimate.



$$p = P_{H_0}(|t| \geq 2.50)$$
$$\approx 2P(Z \leq -2.50) = 2(0.0062) = 0.0124, \qquad Z \sim N(0,1).$$

The p-value and the testing decision

Assuming H_0, the p-value is the probability of an estimate $\hat{\beta}_1$ at least as far from $\beta_{1,0}$ as the observed estimate.



$$p = P_{H_0}(|t| \geq 2.50)$$
$$\approx 2P(Z \leq -2.50) = 2(0.0062) = 0.0124, \qquad Z \sim N(0,1).$$

Since $p < 0.05$, reject $H_0 : \beta_1 = 0$.

Exercise 1: testing a different null value

Keep $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$.

1. Test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$ at 5%.

2. Does this establish that the population slope equals 1?

Exercise 1: testing a different null value

Keep $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$.

1. Test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$ at 5%.

$$t = \frac{2 - 1}{0.80} = 1.25.$$

Since $1.25 < 1.96$, we fail to reject $H_0 : \beta_1 = 1$.

2. Does this establish that the population slope equals 1?

Exercise 1: testing a different null value

Keep $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$.

1. Test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$ at 5%.

$$t = \frac{2 - 1}{0.80} = 1.25.$$

Since $1.25 < 1.96$, we fail to reject $H_0 : \beta_1 = 1$.

2. Does this establish that the population slope equals 1?

No. The data are compatible with a range of slopes. The test does not establish that one particular value is true.

5.2

Confidence intervals

Sampling uncertainty and confidence intervals

We cannot establish the true value of β_1 because we only observe the sample and not the population.

The OLS estimate $\hat{\beta}_1$ varies across samples and our estimate of the population slope is therefore subject to sampling uncertainty.

Sampling uncertainty and confidence intervals

We cannot establish the true value of β_1 because we only observe the sample and not the population.

The OLS estimate $\hat{\beta}_1$ varies across samples and our estimate of the population slope is therefore subject to sampling uncertainty.

We can use $\hat{\beta}_1$ and its standard error to construct an ***x*%** confidence interval for β_1.

Sampling uncertainty and confidence intervals

We cannot establish the true value of β_1 because we only observe the sample and not the population.

The OLS estimate $\hat{\beta}_1$ varies across samples and our estimate of the population slope is therefore subject to sampling uncertainty.

We can use $\hat{\beta}_1$ and its standard error to construct an ***x*%** confidence interval for β_1.

If we repeatedly drew samples and constructed an ***x*%** confidence interval from each, approximately ***x*%** of those intervals would contain the true β_1.

A 95% confidence interval for the slope

Using the large-sample normal approximation, a 95% confidence interval is

$$\hat{\beta}_1 \pm 1.96\,\mathrm{SE}(\hat{\beta}_1).$$

A 95% confidence interval for the slope

Using the large-sample normal approximation, a 95% confidence interval is

$$\hat{\beta}_1 \pm 1.96\,\mathrm{SE}(\hat{\beta}_1).$$

The critical value **1.96** leaves 2.5% in each tail of the standard normal distribution, with 95% between the cutoffs.

A 95% confidence interval for the slope

Using the large-sample normal approximation, a 95% confidence interval is

$$\hat{\beta}_1 \pm 1.96\,\mathrm{SE}(\hat{\beta}_1).$$

The critical value **1.96** leaves 2.5% in each tail of the standard normal distribution, with 95% between the cutoffs.

We can use the same method to construct a confidence interval for the intercept using $\hat{\beta}_0$ and its standard error.

Example: an interval for the earnings slope

With $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$:

Example: an interval for the earnings slope

With $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$:

$$2 \pm 1.96(0.80) = 2 \pm 1.568 \quad \Rightarrow \quad [0.43,\ 3.57].$$

Example: an interval for the earnings slope

With $\hat{\beta}_1 = 2$ and $\text{SE}(\hat{\beta}_1) = 0.80$:

$$2 \pm 1.96(0.80) = 2 \pm 1.568 \quad \Rightarrow \quad [0.43,\ 3.57].$$

The interval runs from £0.43 to £3.57 per hour for each additional year of education.

Exercise 2: study hours and exam scores

Suppose we are interested in the effect of studying on exam scores. In a sample of 400 students, the estimated slope is 0.80 points per study hour, with standard error 0.50.

1. Calculate a 95% confidence interval.

2. Would a two-sided test reject a zero slope at 5%?

Exercise 2: study hours and exam scores

Suppose we are interested in the effect of studying on exam scores. In a sample of 400 students, the estimated slope is **0.80** points per study hour, with standard error **0.50**.

1. Calculate a 95% confidence interval.

$$0.80 \pm 1.96(0.50) = [-0.18,\ 1.78].$$

2. Would a two-sided test reject a zero slope at 5%?

Exercise 2: study hours and exam scores

Suppose we are interested in the effect of studying on exam scores. In a sample of 400 students, the estimated slope is 0.80 points per study hour, with standard error 0.50.

1. Calculate a 95% confidence interval.

$$0.80 \pm 1.96(0.50) = [-0.18,\ 1.78].$$

2. Would a two-sided test reject a zero slope at 5%?

No. Zero lies inside the interval. Equivalently, $t = 0.80/0.50 = 1.60 < 1.96$.

The interval includes both zero and potentially meaningful positive associations.

Statistical and substantive significance

Consider two separate hypothetical studies of education and earnings:

	Study A	Study B
Slope (£/hour per education year)	0.10	2.00
Standard error	0.02	1.20
t for a zero slope	5.00	1.67

Statistical and substantive significance

Consider two separate hypothetical studies of education and earnings:

	Study A	Study B
Slope (£/hour per education year)	0.10	2.00
Standard error	0.02	1.20
t for a zero slope	5.00	1.67

Study A estimates a small association with a great degree of precision.
Study B estimates a larger association with more uncertainty.

Statistical and substantive significance

Consider two separate hypothetical studies of education and earnings:

	Study A	Study B
Slope (£/hour per education year)	0.10	2.00
Standard error	0.02	1.20
t for a zero slope	5.00	1.67

Study A estimates a small association with a great degree of precision.
Study B estimates a larger association with more uncertainty.

Statistical significance depends on both the estimate and its precision, whereas substantive importance depends on the size and context.

5.3

Regression with a binary regressor

What is a binary regressor?

A binary regressor takes only two values, usually **0** and **1**.

What is a binary regressor?

A binary regressor takes only two values, usually **0** and **1**.

Suppose we want to study the effect of a training programme on worker outcomes, Y_i. Let D_i denote whether a worker completed the programme or not:

$$D_i = \begin{cases} 1 & \text{if worker } i \text{ completed the programme,} \\ 0 & \text{if worker } i \text{ did not complete it.} \end{cases}$$

The intercept, slope and group means

For hourly earnings Y_i, we can write

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad E(u_i \mid D_i) = 0.$$

The intercept, slope and group means

For hourly earnings Y_i, we can write

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad E(u_i \mid D_i) = 0.$$

$$E(Y_i \mid D_i = 0) = \beta_0, \qquad E(Y_i \mid D_i = 1) = \beta_0 + \beta_1.$$

The intercept, slope and group means

For hourly earnings Y_i, we can write

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad E(u_i \mid D_i) = 0.$$

$$E(Y_i \mid D_i = 0) = \beta_0, \qquad E(Y_i \mid D_i = 1) = \beta_0 + \beta_1.$$

The intercept, β_0, is mean earnings without training. The slope is the difference in population means between trained and untrained workers.

The intercept, slope and group means

For hourly earnings Y_i, we can write

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad E(u_i \mid D_i) = 0.$$

$$E(Y_i \mid D_i = 0) = \beta_0, \qquad E(Y_i \mid D_i = 1) = \beta_0 + \beta_1.$$

The intercept, β_0, is mean earnings without training. The slope is the difference in population means between trained and untrained workers.

OLS estimates these coefficients using the sample means:

$$\hat{\beta}_0 = \bar{Y}_{D=0}, \qquad \hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}.$$

Example: training and earnings

Suppose that a sample of 600 workers has mean earnings of £18.00 without training and £21.20 with training.

Example: training and earnings

Suppose that a sample of 600 workers has mean earnings of £18.00 without training and £21.20 with training.

$$\widehat{Y} = 18 + 3.20D.$$

Suppose the standard error of $\hat{\beta}_1$ is 1.20.

Example: training and earnings

Suppose that a sample of 600 workers has mean earnings of £18.00 without training and £21.20 with training.

$$\widehat{Y} = 18 + 3.20D.$$

Suppose the standard error of $\hat{\beta}_1$ is 1.20.

$$t = \frac{3.20}{1.20} \approx 2.67, \qquad 95\%\ \text{CI} = 3.20 \pm 1.96(1.20) = [0.85,\ 5.55].$$

Example: training and earnings

Suppose that a sample of 600 workers has mean earnings of £18.00 without training and £21.20 with training.

$$\widehat{Y} = 18 + 3.20D.$$

Suppose the standard error of $\hat{\beta}_1$ is 1.20.

$$t = \frac{3.20}{1.20} \approx 2.67, \qquad 95\%\ \text{CI} = 3.20 \pm 1.96(1.20) = [0.85,\ 5.55].$$

Training participants earn more on average. Note that we do not yet know if this difference is driven by treatment or selection.

Exercise 3: changing the reference group

Our fitted line is $\widehat{Y} = 18 + 3.20D$, where $D = 1$ means trained.

Define $N = 1 - D$, so $N = 1$ means not trained.

1. Write the fitted line using N.

2. Do the predictions or the two-sided test of no difference change?

Exercise 3: changing the reference group

Our fitted line is $\widehat{Y} = 18 + 3.20D$, where $D = 1$ means trained.

Define $N = 1 - D$, so $N = 1$ means not trained.

1. Write the fitted line using N.

$$\widehat{Y} = 18 + 3.20(1 - N) = 21.20 - 3.20N.$$

2. Do the predictions or the two-sided test of no difference change?

Exercise 3: changing the reference group

Our fitted line is $\widehat{Y} = 18 + 3.20D$, where $D = 1$ means trained.

Define $N = 1 - D$, so $N = 1$ means not trained.

1. Write the fitted line using N.

$$\widehat{Y} = 18 + 3.20(1 - N) = 21.20 - 3.20N.$$

2. Do the predictions or the two-sided test of no difference change?

No. The reference group and slope sign change. The fitted values, standard error and absolute t-statistic stay the same.

5.4

Heteroskedasticity and standard errors

The conditional mean and variance of errors

Our least-squares assumptions include a condition on the conditional mean of the errors:

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

The conditional mean and variance of errors

Our least-squares assumptions include a condition on the conditional mean of the errors:

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

They allow the error variance to vary with X. We can additionally assume

$$\text{Var}(u_i \mid X_i) = \sigma^2 \qquad \text{for all } X_i.$$

The conditional mean and variance of errors

Our least-squares assumptions include a condition on the conditional mean of the errors:

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

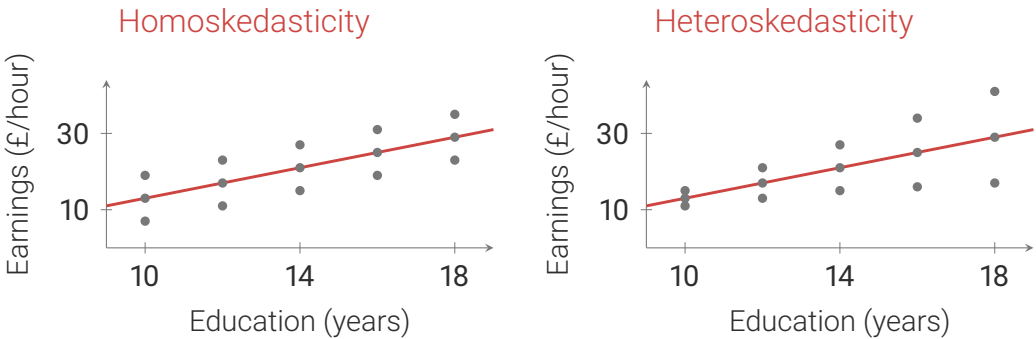
They allow the error variance to vary with X. We can additionally assume

$$\text{Var}(u_i \mid X_i) = \sigma^2 \qquad \text{for all } X_i.$$

Constant conditional variance is homoskedasticity.

When the conditional variance varies with X, we have heteroskedasticity.

The spread of earnings can vary with education



The spread around the line is constant on the left. On the right, it increases with education.

Heteroskedasticity and OLS estimates

Under the other least-squares assumptions, OLS remains unbiased and consistent when the error variance depends on $\boldsymbol{X}$.

Heteroskedasticity and OLS estimates

Under the other least-squares assumptions, OLS remains unbiased and consistent when the error variance depends on $\boldsymbol{X}$.

With homoskedasticity, OLS has the smallest sampling variance among linear unbiased estimators: it is efficient.

Heteroskedasticity and OLS estimates

Under the other least-squares assumptions, OLS remains unbiased and consistent when the error variance depends on $\boldsymbol{X}$.

With homoskedasticity, OLS has the smallest sampling variance among linear unbiased estimators: it is efficient.

With heteroskedasticity, OLS is no longer efficient. Other estimators can have smaller sampling variances.

Heteroskedasticity and statistical inference

Standard errors calculated under the assumption of constant error variance can be misleading when that assumption fails.

Heteroskedasticity and statistical inference

Standard errors calculated under the assumption of constant error variance can be misleading when that assumption fails.

Using a homoskedasticity-only standard error,

$$t_{\text{hom}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}_{\text{hom}}(\hat{\beta}_1)}$$

need not have an approximately standard normal distribution under H_0, even in large samples.

Heteroskedasticity and statistical inference

Standard errors calculated under the assumption of constant error variance can be misleading when that assumption fails.

Using a homoskedasticity-only standard error,

$$t_{\text{hom}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}_{\text{hom}}(\hat{\beta}_1)}$$

need not have an approximately standard normal distribution under H_0, even in large samples.

The resulting p-values, hypothesis tests and confidence intervals can therefore be unreliable.

Heteroskedasticity-robust standard errors

Robust standard errors allow the error variance to vary with $\boldsymbol{X}$.

Heteroskedasticity-robust standard errors

Robust standard errors allow the error variance to vary with $\boldsymbol{X}$.

In large samples, they support valid t-tests, p-values and confidence intervals under either homoskedasticity or heteroskedasticity.

Heteroskedasticity-robust standard errors

Robust standard errors allow the error variance to vary with $\boldsymbol{X}$.

In large samples, they support valid t-tests, p-values and confidence intervals under either homoskedasticity or heteroskedasticity.

Changing the standard-error calculation leaves the OLS coefficients, fitted values and residuals unchanged.

Heteroskedasticity-robust standard errors

Robust standard errors allow the error variance to vary with $\boldsymbol{X}$.

In large samples, they support valid t-tests, p-values and confidence intervals under either homoskedasticity or heteroskedasticity.

Changing the standard-error calculation leaves the OLS coefficients, fitted values and residuals unchanged.

The formula used in Section 5.1 is already robust, so the earlier examples already use this approach.

We use robust standard errors throughout this lecture unless a comparison is explicitly stated.

Exercise 4: the same slope, different standard errors

A hypothetical large-sample regression estimates a slope of 2.00.

$$\text{SE}_{\text{hom}} = 0.70, \qquad \text{SE}_{\text{robust}} = 1.20.$$

1. Test a zero slope at 5% using each standard error.

Exercise 4: the same slope, different standard errors

A hypothetical large-sample regression estimates a slope of 2.00.

$$\mathrm{SE}_{\mathrm{hom}} = 0.70, \qquad \mathrm{SE}_{\mathrm{robust}} = 1.20.$$

1. Test a zero slope at 5% using each standard error.

$$t_{\mathrm{hom}} = 2/0.70 = 2.86, \qquad t_{\mathrm{robust}} = 2/1.20 = 1.67.$$

The first rejects. The robust test fails to reject.

6.1

Omitted-variable bias

Education, family background and earnings

Suppose workers whose parents attended university tend to have more education and also earn more for other reasons.

Let $Z_i = 1$ if at least one parent is a university graduate, and 0 otherwise.

Education, family background and earnings

Suppose workers whose parents attended university tend to have more education and also earn more for other reasons.

Let $Z_i = 1$ if at least one parent is a university graduate, and 0 otherwise.

Consider the population model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i, \qquad E(u_i \mid X_i, Z_i) = 0.$$

Education, family background and earnings

Suppose workers whose parents attended university tend to have more education and also earn more for other reasons.

Let $Z_i = 1$ if at least one parent is a university graduate, and 0 otherwise.

Consider the population model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i, \qquad E(u_i \mid X_i, Z_i) = 0.$$

If we leave out Z_i, the education coefficient can capture both education and the related differences in family background.

When does an omitted variable bias the slope?

In this model, leaving out Z produces omitted-variable bias in the education slope when both conditions hold:

When does an omitted variable bias the slope?

In this model, leaving out Z produces omitted-variable bias in the education slope when both conditions hold:

- Z helps determine earnings after accounting for education: $\beta_2 \neq 0$.

When does an omitted variable bias the slope?

In this model, leaving out Z produces omitted-variable bias in the education slope when both conditions hold:

- Z helps determine earnings after accounting for education: $\beta_2 \neq 0$.
- Z is correlated with education: $\text{Cov}(X, Z) \neq 0$.

When does an omitted variable bias the slope?

In this model, leaving out Z produces omitted-variable bias in the education slope when both conditions hold:

- Z helps determine earnings after accounting for education: $\beta_2 \neq 0$.
- Z is correlated with education: $\text{Cov}(X, Z) \neq 0$.

An omitted earnings determinant that is uncorrelated with education does not create this bias in the education slope.

The direction and size of omitted-variable bias

Under the two-regressor model, the simple-regression slope converges to

$$\beta_1 + \beta_2 \frac{\text{Cov}(X, Z)}{\text{Var}(X)}.$$

Suppose, hypothetically, that

$$\beta_1 = 1.20, \qquad \beta_2 = 4, \qquad \frac{\text{Cov}(X, Z)}{\text{Var}(X)} = 0.20.$$

The direction and size of omitted-variable bias

Under the two-regressor model, the simple-regression slope converges to

$$\beta_1 + \beta_2 \frac{\text{Cov}(X, Z)}{\text{Var}(X)}.$$

Suppose, hypothetically, that

$$\beta_1 = 1.20, \qquad \beta_2 = 4, \qquad \frac{\text{Cov}(X, Z)}{\text{Var}(X)} = 0.20.$$

$$\text{omitted-variable bias} = 4(0.20) = 0.80, \qquad \text{simple slope} \longrightarrow 2.00.$$

The direction and size of omitted-variable bias

Under the two-regressor model, the simple-regression slope converges to

$$\beta_1 + \beta_2 \frac{\text{Cov}(X, Z)}{\text{Var}(X)}.$$

Suppose, hypothetically, that

$$\beta_1 = 1.20, \qquad \beta_2 = 4, \qquad \frac{\text{Cov}(X, Z)}{\text{Var}(X)} = 0.20.$$

$$\text{omitted-variable bias} = 4(0.20) = 0.80, \qquad \text{simple slope} \longrightarrow 2.00.$$

Omitting family background biases the education slope upward.

Exercise 5: predicting the direction of bias

Suppose prior ability raises earnings even at a given education level.

1. More able workers tend to acquire more education. What happens if we omit ability?

2. Suppose ability is instead uncorrelated with education. Does omitting it still bias the education slope in this linear model?

Exercise 5: predicting the direction of bias

Suppose prior ability raises earnings even at a given education level.

1. More able workers tend to acquire more education. What happens if we omit ability?

The omitted variable has a positive earnings coefficient and a positive correlation with education. The bias is upward.

2. Suppose ability is instead uncorrelated with education. Does omitting it still bias the education slope in this linear model?

Exercise 5: predicting the direction of bias

Suppose prior ability raises earnings even at a given education level.

1. More able workers tend to acquire more education. What happens if we omit ability?

The omitted variable has a positive earnings coefficient and a positive correlation with education. The bias is upward.

2. Suppose ability is instead uncorrelated with education. Does omitting it still bias the education slope in this linear model?

No. The covariance term is zero, so this source of omitted-variable bias disappears. Other sources of bias may remain.