Week 6: Introduction to Linear Regression

Philipp Broniecki

Hertie School of Governance

Statistics 1

1. Midterm
2. Motivation
3. The Linear Regression Model
4. Estimation
5. Coefficient interpretation
6. Hypothesis tests and confidence intervals
7. Take-Home Questions

Overview

Midterm

- Date: Wednesday, 24 October 2018
- Time: 10 a.m. to 12 p.m.
- Place: Forum, r 2.61, 2.32, 2.30, 3.30, 3.61

- You can use: Calculator, a formula collection (containing formulas only)

- We cover all material up until the midterm
- Questions might involve calculations such as confidence intervals, t values, standard errors
- Question will ask applied substantial questions (see last slide today): please answer in full sentences

Overview

Motivation

- Last week we focused on testing relationships between a nominal independent variable (X) and an interval dependent variable (Y)
  - e.g. Are German speaking towns more supportive of higher taxes than French speaking towns?
- Often, we will be interested in testing relationships between two interval variables (X and Y). That is, where both X and Y are continuous.

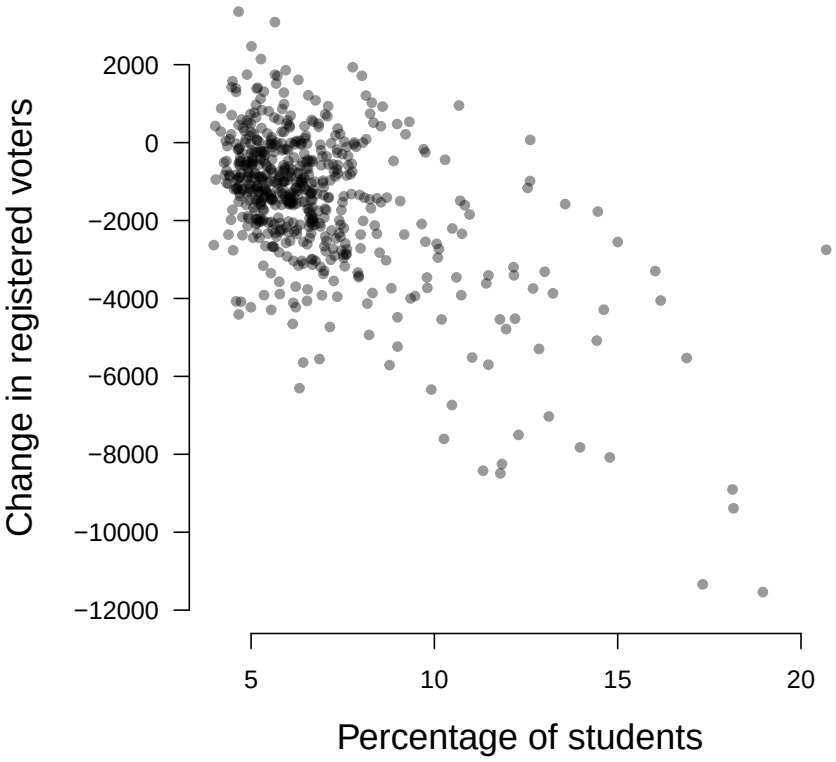
Example I: Students and the electoral register

Students and the electoral register

Before 2015 in the UK, the head of the household could register all members of the household to vote. From 2015, all individuals had to register separately. There were particular concerns that this would lead to many students and young people 'falling off' the electoral register. We collect data on voter registration in 573 UK constituencies to evaluate this concern.

- Unit of analysis: 573 parliamentary constituencies.
- Dependent variable (Y): *Change* in the number of registered voters in a constituency (from 2010 to 2015).
- Independent variable (X): Percentage of a constituency's population who are full time students.

Example I: Students and the electoral register



- What can we tell from looking at this plot?
- Is there a positive or a negative relationship between X and Y?
- Linear regression will help us to make more precise statements about relationships like this.

Key steps in linear regression

- Estimation: Which line can we draw through the data points that 'best' summarises the relationship between X and Y?
- Fitted values: What is our best guess for the conditional value of Y, given a particular value of X?
- Hypothesis testing: How can we test whether the true slope of the line is equal to 0?
- Confidence intervals: What are the plausible values for the slope of the true regression line?

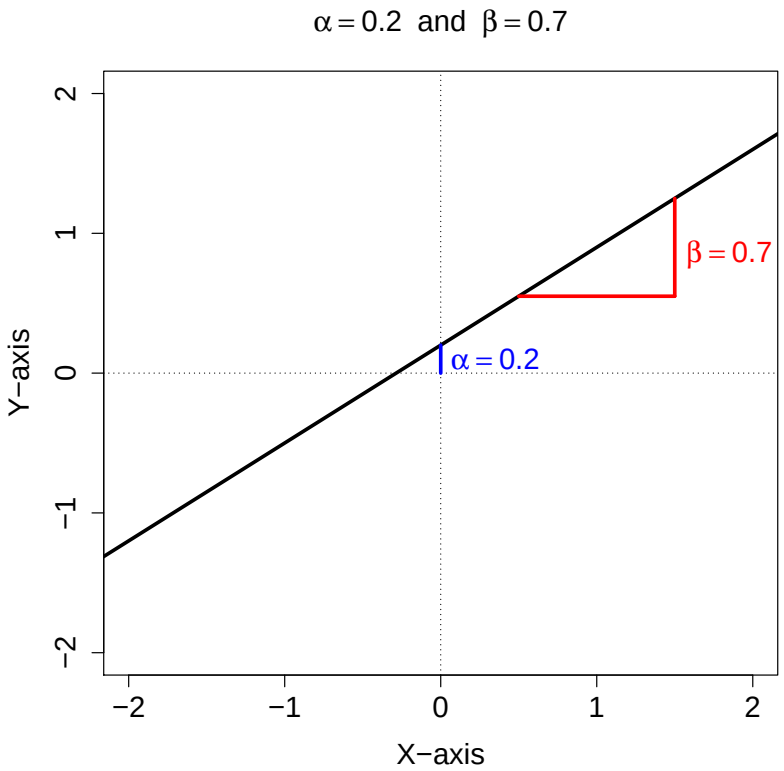
Overview

Linear relationships

- A linear regression model is an approximation of the relationship between our independent variable X and our response variable Y
- In our case, a linear regression model will approximate the true relationship between:
  - the proportion of students, and
  - the change in the number of registered voters

Linear relationships

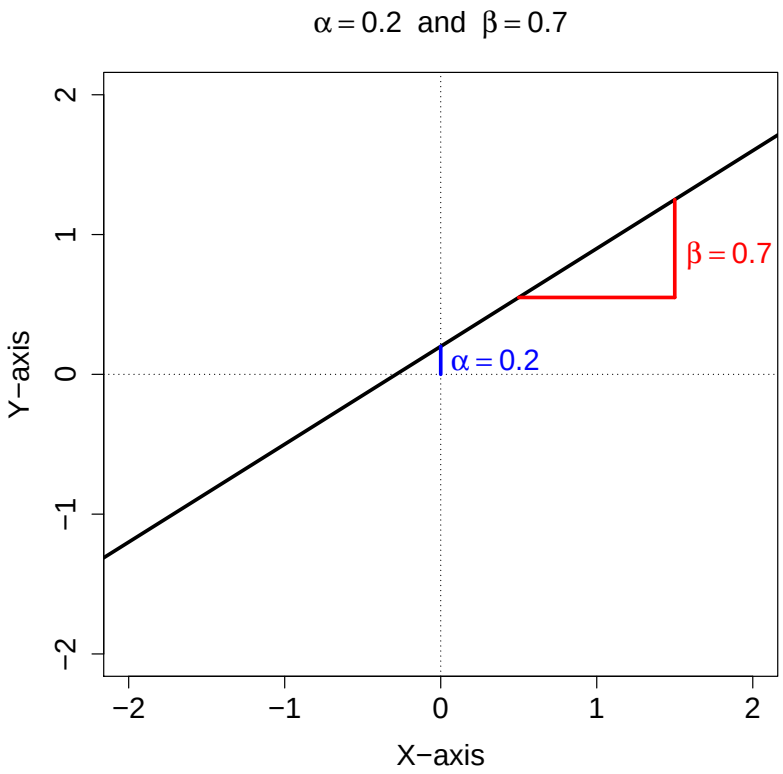
- The most straightforward way of describing the relationship between two variables is with a line
- A line can be represented by this expression: $Y = \alpha + \beta X$

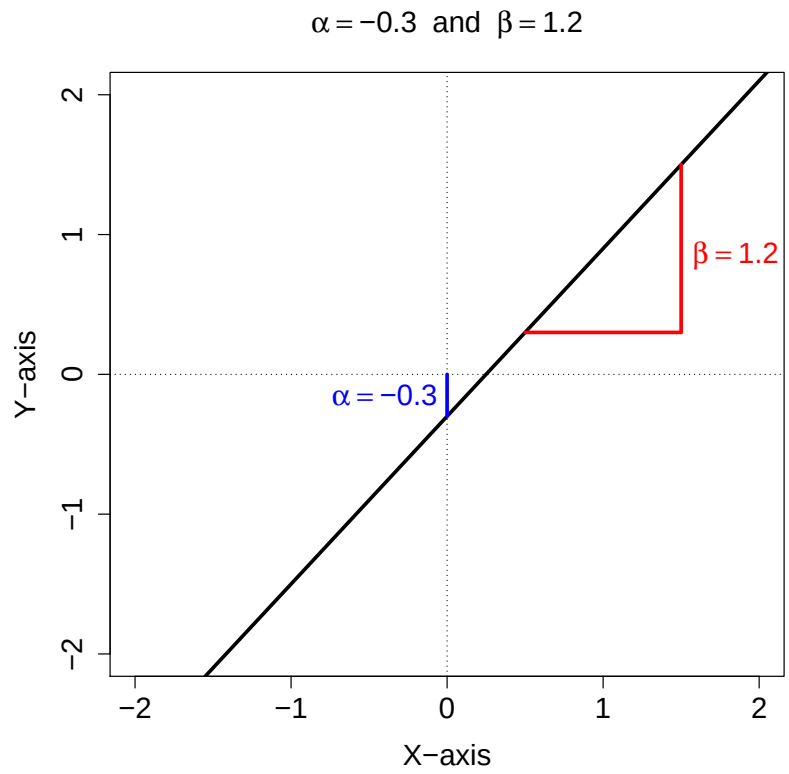


- α is the intercept: the value of Y where $X = 0$
- β is the slope: the amount that Y increases when X increases by one unit
- Here, a one-unit increase in X is associated with a 0.7-unit increase in Y

Linear relationships

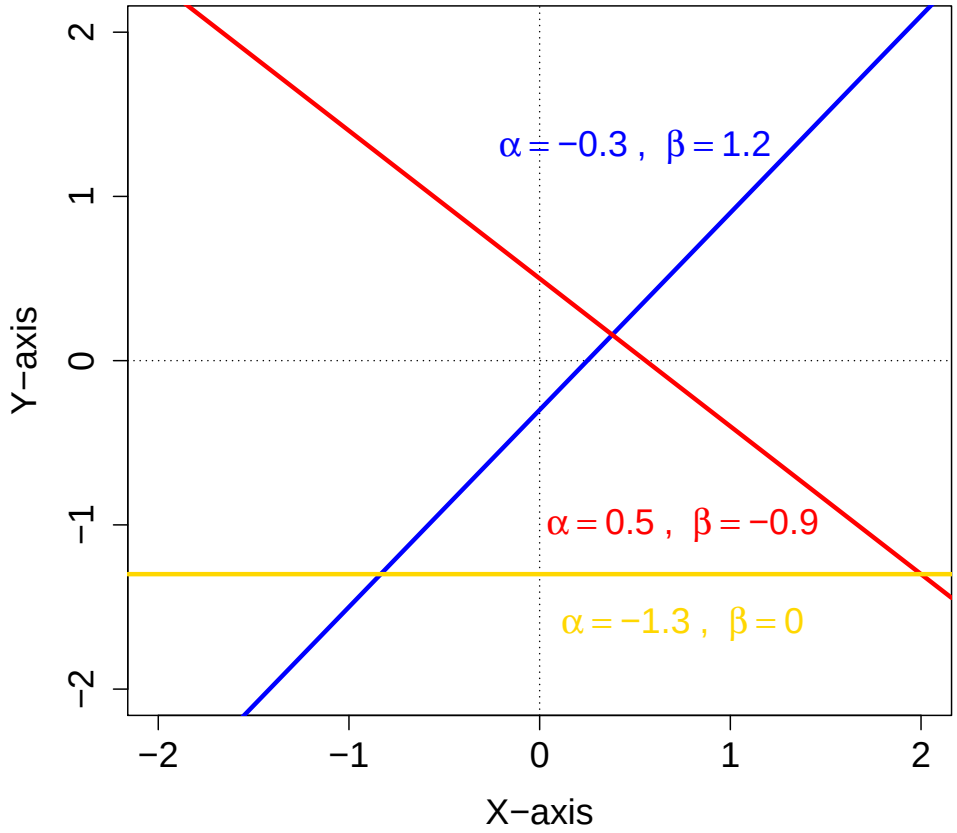
- Different values of α and β uniquely define different lines





Linear relationships

- Different values of α and β uniquely define different lines



- Our goal is to estimate the line that 'best' fits our data

The linear regression model

- The simplest way to summarize the relationship between two variables is to assume that they are linearly related
- We can express this with the bivariate linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where:

- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept or constant.
- β_1 is the slope.
- u_i is the error term or residuals.

β_0 and β_1 are known as the coefficients of the regression line.

The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

We can think of this as a simplified statement about how Y varies:

- The first two terms are the systematic part of the equation
  - β_0 gives the average value of Y when X is equal to 0
  - β_1 gives the average change in Y that results from a one-unit change in X
  - These form the population regression line: the relationship that holds, on average, between X and Y in the population
- The final term is the random part of the equation
  - u_i represents all the other factors aside from X that determine the value of Y

The linear regression model: Voter registration

- In our voter registration example
  - Y_i – change in number of registered voters in constituency i
  - X_i – percentage of students in constituency i
  - What does β_1 represent?
    - the average effect of a one unit change in the percentage of students on change in registration
  - What does β_0 represent?
    - the average change in registration for a constituency with 0% students

The linear regression model

- In our example, the population regression line is:

$$\text{Change in Registered Voters}_i = \beta_0 + \beta_1 * \text{Proportion of Students}_i$$

- β_1 is the slope of population regression line

$$\beta_1 = \frac{\Delta \text{Registration}}{\Delta \text{Students}}$$

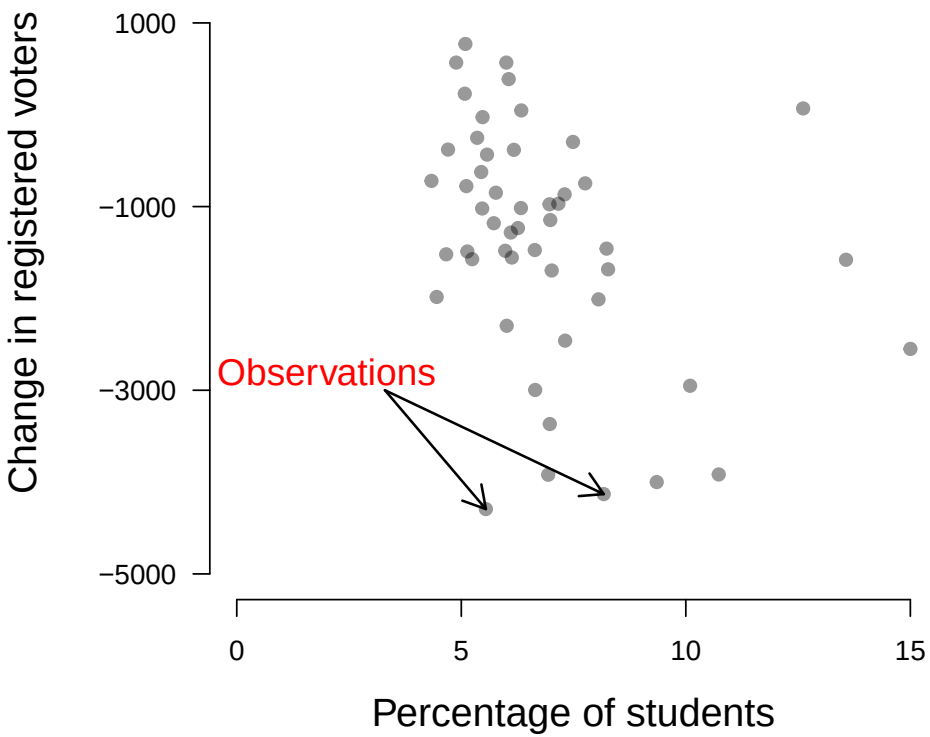
This is tells us how much we should expect the dependent variable (change in voter registration) to change for a unit increase in the proportion of students.

- β_0 is the intercept of the population regression line. It tells us the average change in voter registration for constituencies where the proportion of students is equal to zero.

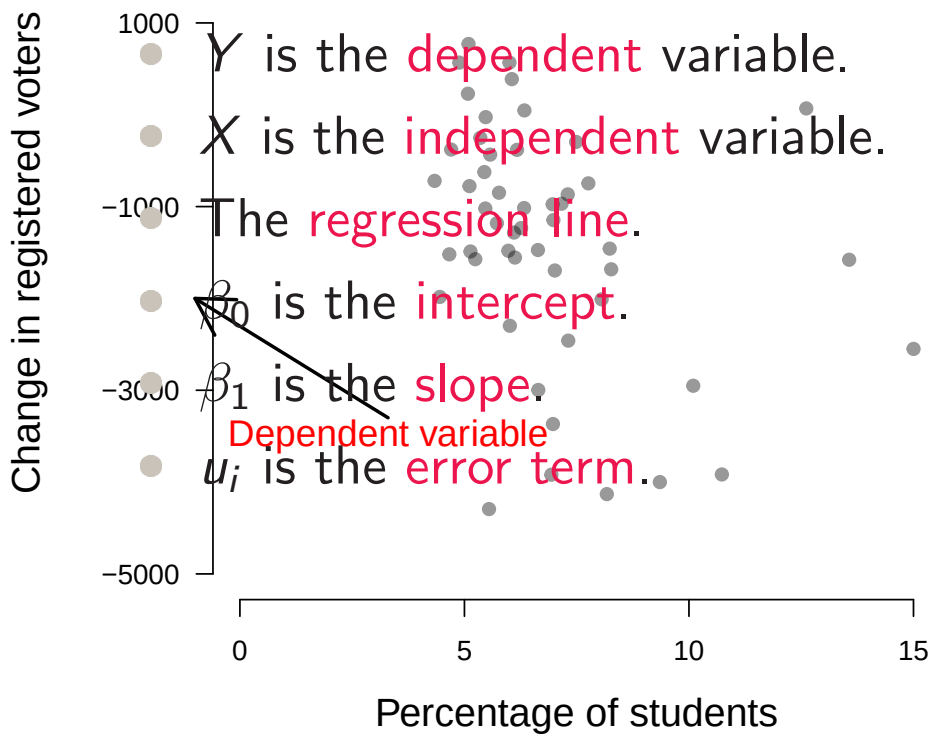
The linear regression model

- Why are β_0 and β_1 "population" parameters?
  - Average relationship between X and Y in the population
- We want to know the population value of β_1 (and β_0).
  - We don't know these values, so must estimate them.
  - We estimate the values using a sample from the population
- Commonly, the most interesting parameter is β_1
- Notation
  - β_0, $\beta_1 \rightarrow$ population parameter values
  - $\hat{\beta}_0$, $\hat{\beta}_1 \rightarrow$ estimated parameter values (or coefficients)

The linear regression line



- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.
- β_0 is the intercept.
- β_1 is the slope.
- u_i is the error term.

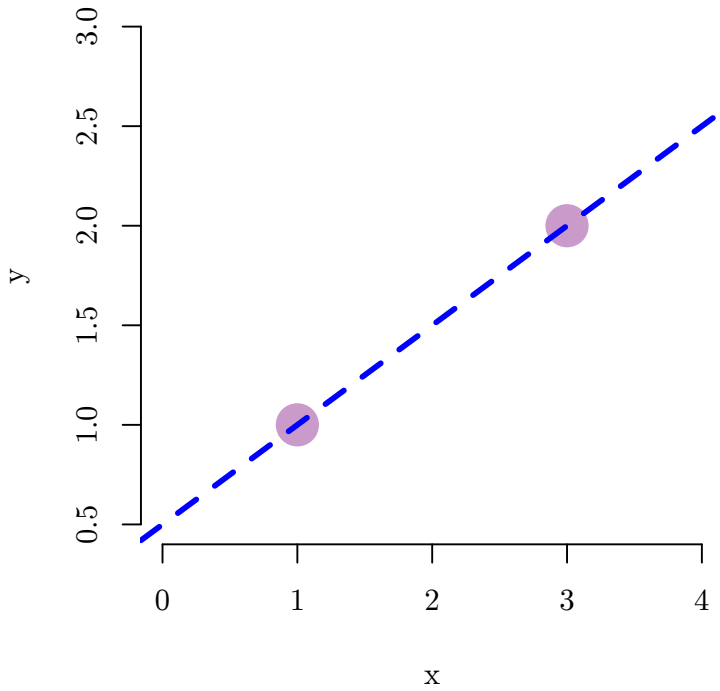


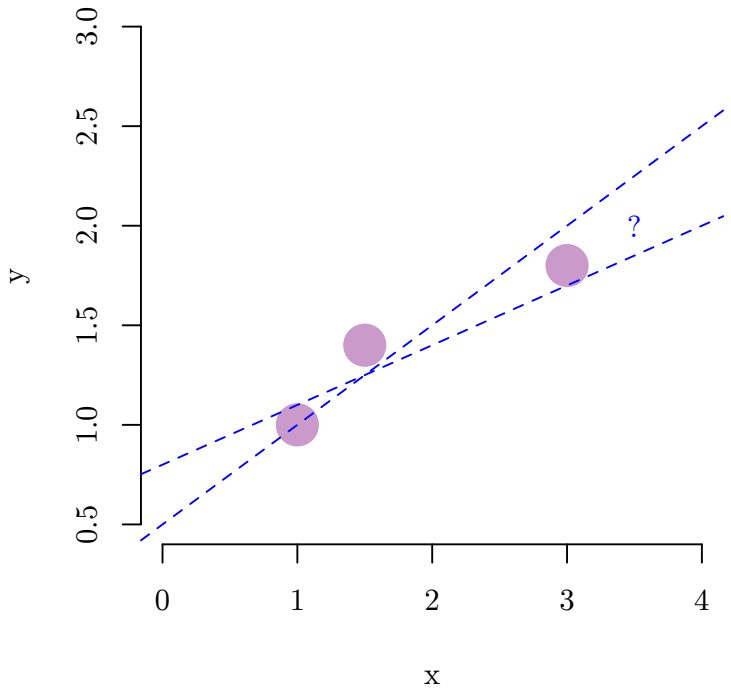
Overview

Estimating β_0 and β_1

We know from basic school geometry:

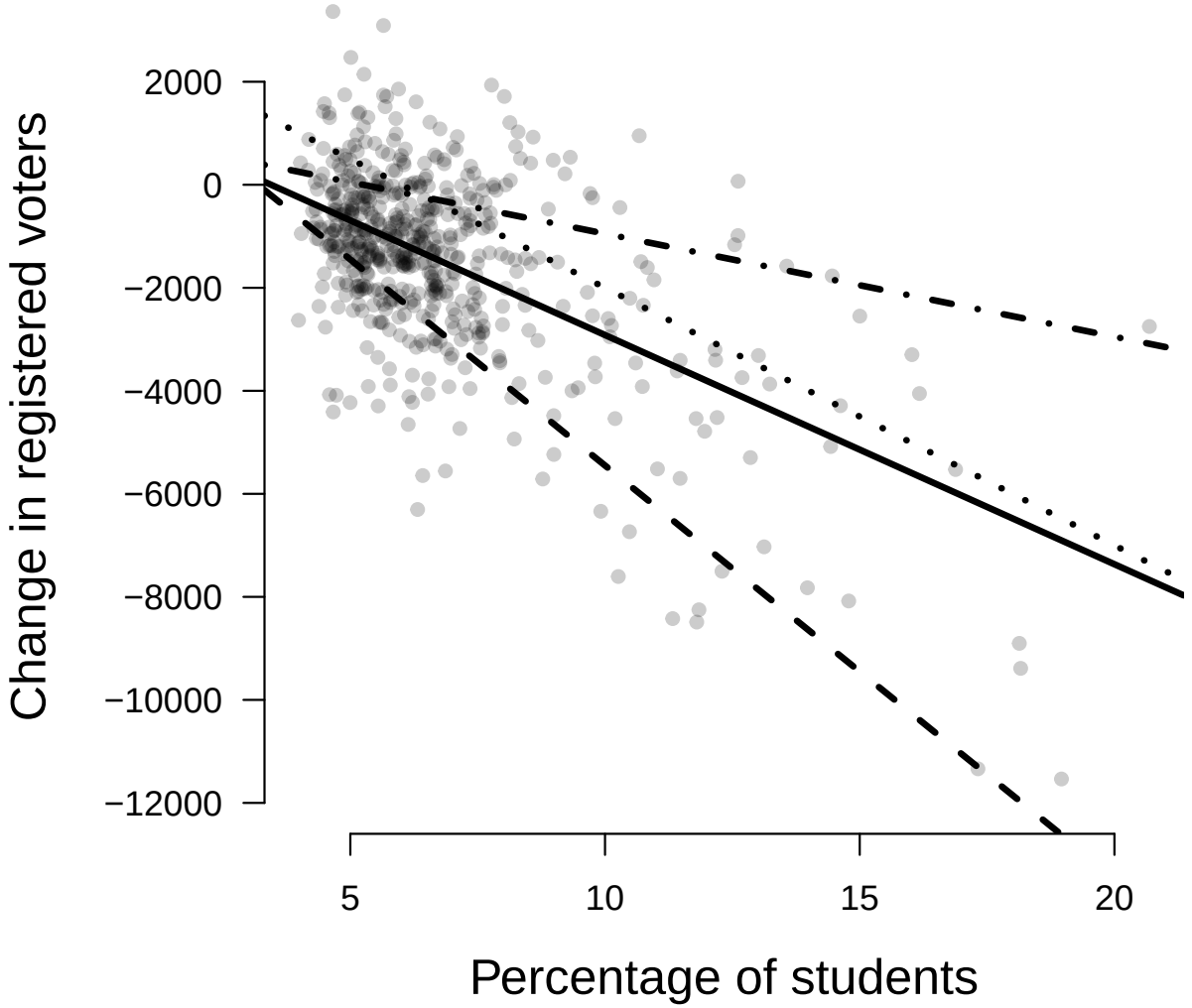
- How to draw a line through two points
- But what do we do when the points are not on the line?





Estimating β_0 and β_1

Which of these lines best fits our data?



Estimating β_0 and β_1

- How can we estimate β_0 and β_1 from our data on voter registration?
- The most widely used approach to estimating the parameters of the linear regression model is the ordinary least squares (OLS) method.

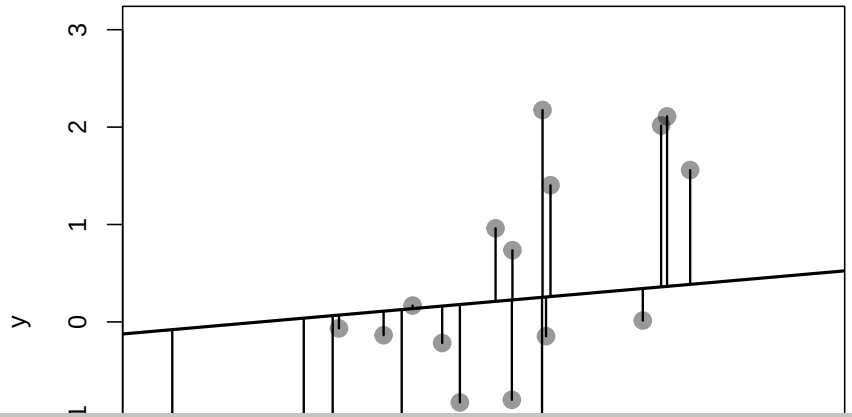
Ordinary Least Squares

- The OLS estimator chooses the regression coefficients so that the estimated line is "as close as possible" to the data.
- It minimizes the sum of the squared differences between the actual values of each observation (Y_i) and the predicted value of each value based on the estimated line ($\hat{Y}_i$).
- Formally, from all possible β_0 and β_1, it chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the following expression:

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

Ordinary Least Squares

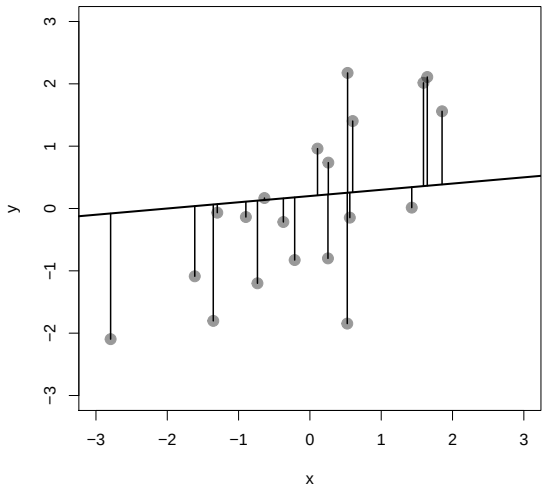
- Let's take some data and plot a random line through the points: Let's take some data and plot the best fitting line through the points:
- For this line, the sum of the squared distances between Y_i and $\hat{Y}_i$:

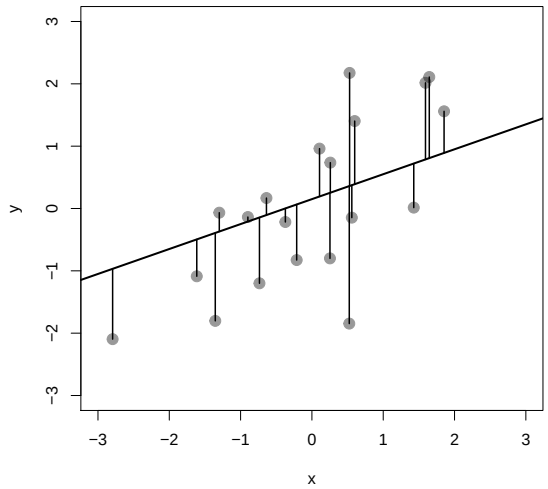


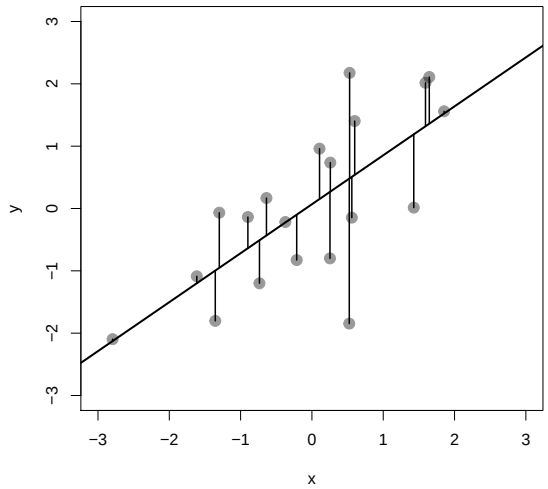
$$\sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i \right) \right]^2 = 30.53$$

- For this line, the sum of the squared distances between

Ordinary Least Squares







- OLS selects the line that minimizes the sum of the squared distances between each point and the line

Ordinary Least Squares

We can estimate the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ easily in R.

```
## Let's call our data.frame "constituencies"
## We can use the lm function (lm = linear model)
my_linear_model <- lm(voters ~ students, data = constituencies)

## We can use the "print" function to view the output
print(my_linear_model)

Call:
lm(formula = voters ~ students, data = constituencies)

Coefficients:
(Intercept)     students
       1533         -445
```

Where (Intercept) $= \hat{\beta}_0$ and students $= \hat{\beta}_1$.

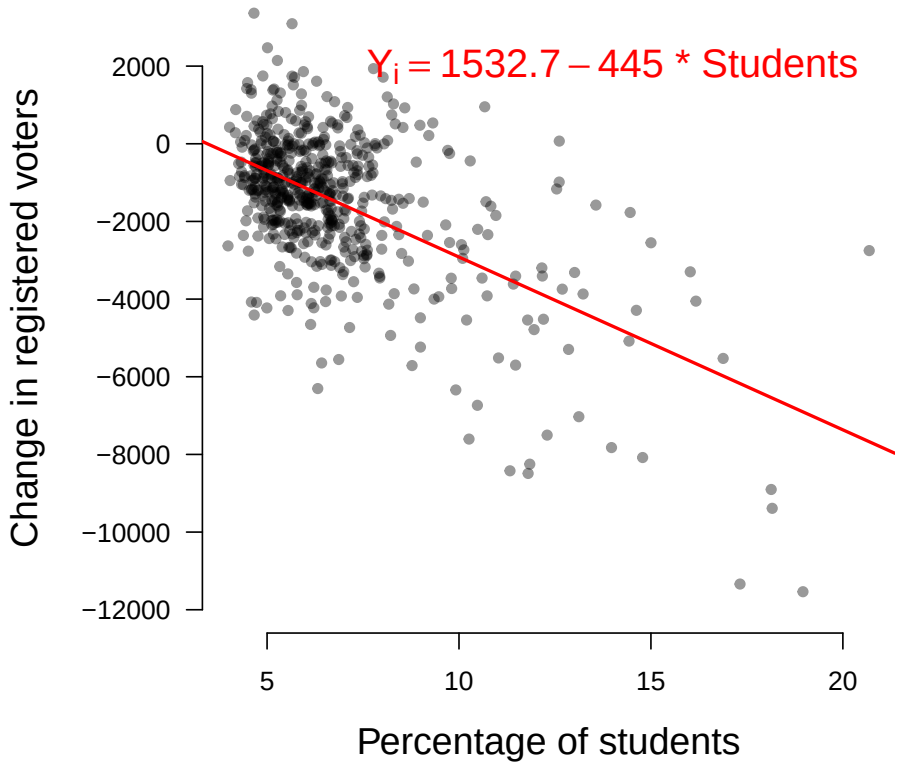
Overview

OLS estimates

The estimated relationship between the percentage of students and change in the number of registered voters is

$$\widehat{\text{Voters}}_i = 1532.7 - 445 \times \text{Students}_i$$

- where *Voters* is the change in registered voters
- *Students* is the percentage of students



OLS estimates: interpretation

- What is the interpretation of $\hat{\beta}_1 = -445$?
  - Generic: A one-unit increase in X is associated with a $\hat{\beta}_1$ change in Y, on average.
  - Specific: A one point increase in the percentage of students in a constituency is associated with a decrease of 445 in the number of registered voters, on average.

OLS estimates: interpretation

- What is the interpretation of $\hat{\beta}_0 = 1532$?
  - Generic: The average value of Y, when X is equal to 0, is $\hat{\beta}_0$
  - Specific: For a hypothetical constituency with 0% students, the model predicts that the number of voters would increase by 1532 between 2010 and 2015.
  - This interpretation of the intercept is not meaningful, as it extrapolates outside the range of the data.

Fitted values

We can also use the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ to calculate fitted or predicted values for any of our sample of X observations.

- The fitted values $\hat{Y}_i$ are:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \ \ i = 1, \ldots, n$$

The fitted values tell us what the best guess is for Y for a specific value of X.

- The residuals $\hat{u}_i$ are

$$\hat{u}_i = Y_i - \hat{Y}_i, \ \ i = 1, \ldots, n.$$

The residuals tell us how far our best guess for each observation is from the value of Y we observe in the sample.

Fitted values

We can also calculate fitted values ($\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$) for *any* arbitrary value of X which may be of interest!

- What is the predicted change in the number of registered voters for a constituency with 10% students?

$$\hat{Y}_i = 1532 - 445 * 10 = -2918$$

- What is the predicted change in the number of registered voters for a constituency with 20% students?

$$\hat{Y}_i = 1532 - 445 * 20 = -7368$$

Overview

Hypothesis tests

$\hat{\beta}_1$ is computed from a sample of data. A different sample would yield different values $\rightarrow$ just as when we computed the sample mean, $\hat{\beta}_1$ is subject to "sampling uncertainty".

- We want to:
  - quantify the sampling uncertainty associated with $\hat{\beta}_1$;
  - use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
  - construct a confidence interval for β_1

Hypothesis tests

- Problem: The government claims that the new system of voter registration does not affect students disproportionately.
- In our sample of data, a 1 point increase in the percentage of students in a constituency is associated with a decrease of 445 in the number of registered voters
  - Is this effect is statistically significantly different from 0?
  - How compatible is our estimate with the (null) hypothesis of the government?

Hypothesis tests

Hypothesis tests in the regression setting are very similar to those we studied last week:

- Specify a hypothesis and a null hypothesis
- Calculate the test-statistic
- Derive the sampling distribution of the test-statistic under the assumption that the null hypothesis is true
- Calculate the p-value
- State a conclusion

Null and alternative hypothesis

- One way to address the claim of the government is to specify the following null and alternative hypothesis:
  - H_0: the percentage of students has no effect on voter registration
  - H_A: the percentage of students has an effect on voter registration

The t-test

- The test statistic for a single regression coefficient is:

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}_1}}$$

  where $\hat{\sigma}_{\hat{\beta}_1}$ is the standard error of $\hat{\beta}_1$.
- Note that in the very common case where the null hypothesis is $\beta_{H_0} = 0$ the t-statistic simplifies to $t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$
- You do not need to know how to calculate the standard error ($\hat{\sigma}_{\hat{\beta}_1}$), but given $\hat{\beta}_1$ and $\hat{\sigma}_{\hat{\beta}_1}$, you need to be able to calculate t

The sampling distribution of the OLS estimator

What is the sampling distribution of t?

- When n is small (< 30), t follows a t-distribution with n - 2 df.
- When n is large (> 30) the Central Limit Theorem implies that t will follow the standard normal distribution
- Most regression packages always use the t distribution as the normal distribution is only correct for large sample sizes

$\rightarrow$ we can normally assume that t will follow the standard normal, unless n is very small

Application to voter registration

- For the regression of registration on the percentage of students we obtain:

	voters
(Intercept)	1532.69
	(192.41)
students	−444.97
	(26.99)
R^2	0.32
N.	573

where the numbers in brackets are the standard errors of the coefficients.

Application to voter registration

	voters
(Intercept)	1532.69
	(192.41)
students	−444.97
	(26.99)
R^2	0.32
N.	573

- To test the government's hypothesis:

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-445 - 0}{27} \approx -16.48$$

- Can we reject the null hypothesis at $\alpha = 0.05$?

Application to voter registration

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-445 - 0}{27} \approx -16.48$$

- The probability of observing a value of the t-statistic outside the interval $[-1.96, 1.96]$ is less than five percent under the standard normal distribution.
- As the t-statistic is clearly outside this interval, the probability that H_0 is correct is less than five percent.
- We can therefore reject the government's claim at the five percent significance level.

Application to voter registration

R will automatically calculate the correct test-statistic for you:

```
summary(my_linear_model)

Residuals:
    Min      1Q  Median      3Q     Max
-5163.4  -787.0   -21.7   924.5  4921.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1532.69     192.41   7.966 8.93e-15 ***
students     -444.97      26.99 -16.489  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1525 on 571 degrees of freedom
Multiple R-squared:  0.3226,Adjusted R-squared:  0.3214
F-statistic: 271.9 on 1 and 571 DF,  p-value: < 2.2e-16
```

Statistical significance

- In the vast majority of t-tests the null hypothesis is that the coefficient is equal to zero.
- In this case the null hypothesis is often not even stated and you will encounter statements such as:
  - The coefficient is significant at the XX percent level
  - The coefficient is significant at conventional levels
- In all of these statements the implicit null hypothesis (or simply "null") is that the coefficient of interest is equal to zero.

Statistical significance

- We should not forget that t-test can nevertheless be used to test also other null hypotheses.
- For example, can we reject the null that the true effect of the percentage of students on voter registration is -460?

$$\frac{\hat{\beta}_1 - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-445 - (-460)}{27} = 0.556 \tag{1}$$

- As 0.556 falls within the interval (-1.97, 1.97), we fail to reject the new null hypothesis that $\beta_{H_0} = -460$

Computing p-values

- We can also determine precisely how unlikely the government's hypothesis is given our estimates.
- P-value: the probability that we would observe an absolute test-statistic as large or larger than the one we observe under the assumption that the null hypothesis is true.
- Modern regression packages will always report the p-value for any t-statistic (and also for other regression statistics).

Application to voter registration

R will automatically calculate the correct p-value for you:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5163.4  -787.0   -21.7   924.5  4921.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1532.69     192.41   7.966 8.93e-15 ***
students     -444.97      26.99 -16.489  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1525 on 571 degrees of freedom
Multiple R-squared:  0.3226,Adjusted R-squared:  0.3214
F-statistic: 271.9 on 1 and 571 DF,  p-value: < 2.2e-16
```

- What does a p-value of $<$ 2e-16 mean?
- 2e-16 $=$ 0.0000000000000002
- $\rightarrow$ it is *very* unlikely that we would observe this test-statistic if the null hypothesis were true

Confidence intervals for regression coefficients

- We can also estimate confidence intervals for $\hat{\beta}_1$:

$$95\% \text{ Confidence interval} : \hat{\beta}_1 \pm 1.96 * SE(\hat{\beta}_1)$$

$$99\% \text{ Confidence interval} : \hat{\beta}_1 \pm 2.58 * SE(\hat{\beta}_1)$$

- In the case of our regression the 95 percent confidence interval:

$$\text{Lower bound:} = -445 - 1.96 \times 27 = -497.92$$
$$\text{Upper bound:} = -445 + 1.96 \times 27 = -392.08$$

- Intuition: The confidence interval contains all values of the parameter that cannot be rejected at the five percent significance level given our estimate.

Overview

Take-Home-Questions I

- What is the slope parameter substantially?
- What it the intercept parameter substantially?
  - When do we not interpret the intercept?
- Formulate a usual null hypothesis of a slope parameter

Take-Home-Questions II

Voting age and turnout

We are interested in the effect of the legal voting age on turnout. We measured turnout in 671 different places with varying voting ages.

- Y — turnout (as percentage of voting age population)
- X — voting age (in years); minimum in our sample $= 15$ and maximum $= 24$
- $\hat{\beta}_0 = 55$; $\hat{\sigma}_{\hat{\beta}_0} = 35$
- $\hat{\beta}_1 = 0.9$; $\hat{\sigma}_{\hat{\beta}_1} = 0.3$

(1) Interpret the result substantially; (2) Discuss the intercept; (3) What's our best guess of turnout when the minimum voting age is 16?; (4) What is the difference in turnout b/w voting ages 16 and 18?