Week 5: Midterm revision session

Jack Blumenau & Philipp Broniecki

University College London

Introduction to Quantitative Methods

1. Administrative information
2. Answer advice
3. Hypothesis testing
4. Simple linear regression
5. Multiple linear regression

Overview

1. Administrative information
2. Answer advice
3. Hypothesis testing
4. Simple linear regression
5. Multiple linear regression

Administrative information

- Midterm will be released at 2pm on November 3rd
- Midterm is due at 2pm on November 8th
- All submissions via Turnitin
- Usual late penalties apply
- Usual extenuating circumstances policies apply

Overview

How much detail do I need to include?

- You will not lose marks for writing fewer than 1000 words

How much detail do I need to include?

- You will not lose marks for writing fewer than 1000 words
- You will lose marks for writing more than 1000 words

How much detail do I need to include?

- You will not lose marks for writing fewer than 1000 words
- You will lose marks for writing more than 1000 words
- Your answer should include sufficient detail to fully answer the question

How much detail do I need to include?

- You will not lose marks for writing fewer than 1000 words
- You will lose marks for writing more than 1000 words
- Your answer should include sufficient detail to fully answer the question
    - Statistical information. e.g. How do we interpret the confidence interval?

How much detail do I need to include?

- You will not lose marks for writing fewer than 1000 words
- You will lose marks for writing more than 1000 words
- Your answer should include sufficient detail to fully answer the question
  - Statistical information. e.g. How do we interpret the confidence interval?
  - Substantive information. e.g. What does this tell us about our research question?

How should I present my answers?

- You need to write in full sentences, not bullet points

How should I present my answers?

- You need to write in full sentences, not bullet points
- You should present output of all statistical tests in a clear and readable format

How should I present my answers?

- You need to write in full sentences, not bullet points
- You should present output of all statistical tests in a clear and readable format
  - Do not copy and paste output from R
  - Do not include screenshots from R
  - Use `screenreg` or make a table in Word

How should I present my answers?

- You need to write in full sentences, not bullet points
- You should present output of all statistical tests in a clear and readable format
  - Do not copy and paste output from R
  - Do not include screenshots from R
  - Use `screenreg` or make a table in Word
- Answer the question! If you are asked to answer a policy relevant question, you should not simply report a p-value without commenting on the substance.

How should I present my answers?

- You need to write in full sentences, not bullet points
- You should present output of all statistical tests in a clear and readable format
  - Do not copy and paste output from R
  - Do not include screenshots from R
  - Use `screenreg` or make a table in Word
- Answer the question! If you are asked to answer a policy relevant question, you should not simply report a p-value without commenting on the substance.
- You can use R to answer any question where you think it might be useful. But if the question tells you to 'show your work', that means you need to show that you know how the values from R were calculated!

Overview

Intuition

- Could a relationship we observe in our data have happened by chance?

Intuition

- Could a relationship we observe in our data have happened by chance?
- What is the probability that there is no relationship even though we observed it in our sample?

Intuition

- Could a relationship we observe in our data have happened by chance?
- What is the probability that there is no relationship even though we observed it in our sample?

1. Is the sample mean different from some hypothesised value?

Intuition

- Could a relationship we observe in our data have happened by chance?
- What is the probability that there is no relationship even though we observed it in our sample?

1. Is the sample mean different from some hypothesised value?
2. Are the means in subgroups of our data different? E.g., is average income in Scotland different from income in Wales?

Intuition

- Could a relationship we observe in our data have happened by chance?
- What is the probability that there is no relationship even though we observed it in our sample?

1. Is the sample mean different from some hypothesised value?
2. Are the means in subgroups of our data different? E.g., is average income in Scotland different from income in Wales?
3. Is effect of some X variable on some Y variable different from 0?

Hypothesis test sequence

- State the hypothesis and the null hypothesis
- Calculate a 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

Test for the sample mean: hypothesis

Is the die loaded

Each outcome on a die is equally likely. Thus, the average outcome from throwing a fair die often is 3.5. If we take a die and throw it 100 times and and get an average of 3.46, is that evidence for a loaded die or not?

Test for the sample mean: hypothesis

Is the die loaded

Each outcome on a die is equally likely. Thus, the average outcome from throwing a fair die often is 3.5. If we take a die and throw it 100 times and and get an average of 3.46, is that evidence for a loaded die or not?

- Null Hypothesis: die is fair. The small difference we find is due to chance.
- Hypothesis: The die is loaded. The difference is systematic

Test for the sample mean: t value

- What is the t-statistic?

Test for the sample mean: t value

- What is the t-statistic?

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})}$$

Test for the sample mean: t value

- What is the t-statistic?

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})}$$

Test for the sample mean: t value

- What is the t-statistic?

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})}$$

- The t-statistic is the difference in means. It's units are average distances from the true mean (standard deviations).
- We do not know the standard deviation of the sampling distribution, so we estimate it with the standard error

Test for the sample mean: t value (2)

- The standard error quantifies how much we expect the sample mean to vary from sample to sample

Test for the sample mean: t value (2)

- The standard error quantifies how much we expect the sample mean to vary from sample to sample
- How to get the standard error of the mean $SE(\bar{Y})$?

Test for the sample mean: t value (2)

- The standard error quantifies how much we expect the sample mean to vary from sample to sample
- How to get the standard error of the mean $SE(\bar{Y})$?
- It is computed as the average deviation from our sample mean

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

- where σ_Y is the standard deviation of our sample

Test for the sample mean: t value (2)

- The standard error quantifies how much we expect the sample mean to vary from sample to sample
- How to get the standard error of the mean $SE(\bar{Y})$?
- It is computed as the average deviation from our sample mean

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

- where σ_Y is the standard deviation of our sample
- It approximates the average deviation from the true mean

Test for the sample mean: t value (2)

- The standard error quantifies how much we expect the sample mean to vary from sample to sample
- How to get the standard error of the mean $SE(\bar{Y})$?
- It is computed as the average deviation from our sample mean

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

- where σ_Y is the standard deviation of our sample
- It approximates the average deviation from the true mean
- Formally, it is an *estimate* for the standard deviation of the *sampling distribution*

Test for the sample mean: t value (3)

- First, we need to know the standard deviation of Y (σ_Y)

Test for the sample mean: t value (3)

- First, we need to know the standard deviation of Y (σ_Y)
- The standard deviation of the Y is:

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

Test for the sample mean: t value (3)

- First, we need to know the standard deviation of Y (σ_Y)
- The standard deviation of the Y is:

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

- You cannot compute it from the information we have given you here. You would need to know all Y_i values

Test for the sample mean: t value (3)

- First, we need to know the standard deviation of Y (σ_Y)
- The standard deviation of the Y is:

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

- You cannot compute it from the information we have given you here. You would need to know all Y_i values
- Suppose: $\sigma_Y = 1.69$

Test for the sample mean: t value (4)

- We have all pieces to get the standard error of the mean $SE(\bar{Y})$

$$SE(\bar{Y})$$

Test for the sample mean: t value (4)

- We have all pieces to get the standard error of the mean $SE(\bar{Y})$

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

Test for the sample mean: t value (4)

- We have all pieces to get the standard error of the mean $SE(\bar{Y})$

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}} = \frac{1.69}{\sqrt{100}}$$

Test for the sample mean: t value (4)

- We have all pieces to get the standard error of the mean $SE(\bar{Y})$

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}} = \frac{1.69}{\sqrt{100}} = \frac{1.69}{10}$$

Test for the sample mean: t value (4)

- We have all pieces to get the standard error of the mean $SE(\bar{Y})$

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}} = \frac{1.69}{\sqrt{100}} = \frac{1.69}{10} = 0.17$$

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})}$$

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})}$$

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17}$$

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17} = \frac{-0.04}{0.17}$$

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17} = \frac{-0.04}{0.17} = -0.24$$

- The difference between our observed mean & the null is -0.24 average deviations (standard errors) from the true mean.

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17} = \frac{-0.04}{0.17} = -0.24$$

- The difference between our observed mean & the null is -0.24 average deviations (standard errors) from the true mean.
- That's not much! Our sample is large, so if we repeated our trial 100 times:

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17} = \frac{-0.04}{0.17} = -0.24$$

- The difference between our observed mean & the null is -0.24 average deviations (standard errors) from the true mean.
- That's not much! Our sample is large, so if we repeated our trial 100 times:
  - 68 sample means will be within 1 standard error of true mean

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17} = \frac{-0.04}{0.17} = -0.24$$

- The difference between our observed mean & the null is -0.24 average deviations (standard errors) from the true mean.
- That's not much! Our sample is large, so if we repeated our trial 100 times:
  - 68 sample means will be within 1 standard error of true mean
  - 95 would be within 1.96 standard errors of the true mean

Test for the sample mean: t value (5)

- Now we can calculate t

$$t = \frac{\bar{Y} - \mu_0}{SE(\bar{Y})} = \frac{3.46 - 3.5}{SE(\bar{Y})} = \frac{3.46 - 3.5}{0.17} = \frac{-0.04}{0.17} = -0.24$$

- The difference between our observed mean & the null is -0.24 average deviations (standard errors) from the true mean.
- That's not much! Our sample is large, so if we repeated our trial 100 times:
  - 68 sample means will be within 1 standard error of true mean
  - 95 would be within 1.96 standard errors of the true mean
- We therefore know that the null is not that unlikely $\rightarrow$ We fail to reject the null hypothesis

Test for the sample mean: p value

- The p-value gives the probability of observing an absolute value of the test-statistic as large or larger than the one we calculate from our sample (-0.24), *under the assumption that* H_0 *is true*

Test for the sample mean: p value

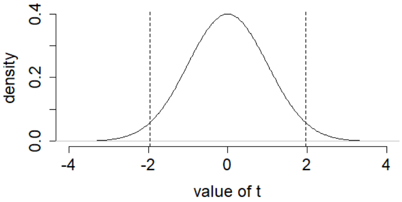
- The p-value gives the probability of observing an absolute value of the test-statistic as large or larger than the one we calculate from our sample (-0.24), *under the assumption that H_0 is true*
  - $\rightarrow$ probability that we mistakenly reject H_0 (false positive)

Test for the sample mean: p value

- The p-value gives the probability of observing an absolute value of the test-statistic as large or larger than the one we calculate from our sample (-0.24), *under the assumption that H_0 is true*
  - $\rightarrow$ probability that we mistakenly reject H_0 (false positive)
- Because n is large ($n = 100$), t follows a normal distribution

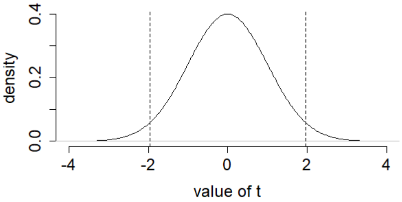
Test for the sample mean: p value

- The p-value gives the probability of observing an absolute value of the test-statistic as large or larger than the one we calculate from our sample (-0.24), *under the assumption that H_0 is true*
  - $\rightarrow$ probability that we mistakenly reject H_0 (false positive)
- Because n is large ($n = 100$), t follows a normal distribution



Test for the sample mean: p value

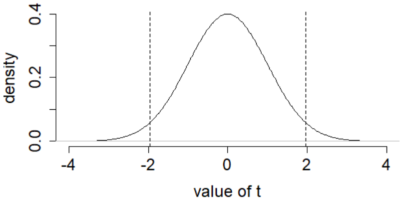
- The p-value gives the probability of observing an absolute value of the test-statistic as large or larger than the one we calculate from our sample (-0.24), *under the assumption that H_0 is true*
  - $\rightarrow$ probability that we mistakenly reject H_0 (false positive)
- Because n is large ($n = 100$), t follows a normal distribution




```
## probability that
## t <= -0.24 or t >= +0.24?
```

Test for the sample mean: p value

- The p-value gives the probability of observing an absolute value of the test-statistic as large or larger than the one we calculate from our sample (-0.24), *under the assumption that H_0 is true*
  - $\rightarrow$ probability that we mistakenly reject H_0 (false positive)
- Because n is large ($n = 100$), t follows a normal distribution



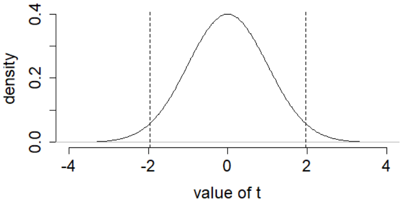

```
## probability that
## t <= -0.24 or t >= +0.24?

2*(1 - pnorm(0.24))
[1] 0.8103303
```

Test for the sample mean: p value (2)

- Alternatively, we can get p using the t distribution with n-1 df
- Df is our number of observations minus 1 degree of freedom for each estimated parameter, i.e. 1 in our case

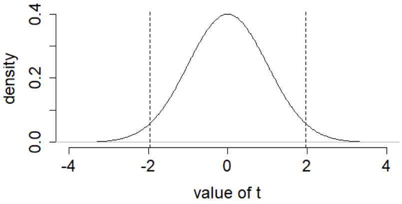
Test for the sample mean: p value (2)

- Alternatively, we can get p using the t distribution with n-1 df
- Df is our number of observations minus 1 degree of freedom for each estimated parameter, i.e. 1 in our case



Test for the sample mean: p value (2)

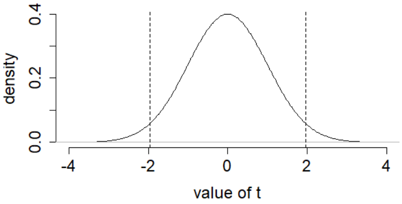
- Alternatively, we can get p using the t distribution with n-1 df
- Df is our number of observations minus 1 degree of freedom for each estimated parameter, i.e. 1 in our case




```
## probability that
## t <= -0.24 or t >= +0.24?
```

Test for the sample mean: p value (2)

- Alternatively, we can get p using the t distribution with n-1 df
- Df is our number of observations minus 1 degree of freedom for each estimated parameter, i.e. 1 in our case




```
## probability that
## t <= -0.24 or t >= +0.24?

2*(1 - pt(0.24, df = 99))
[1] 0.8108265
```

Test for the sample mean: R

- You can carry out the individual steps or you can use the t.test() function

```
t.test( var.name, mu = value of H0 , conf = 0.95 )
```

t-tests for the difference in two means

- Often we are interested in whether the mean for one group is different from the mean for another group
    - Is woman's income different to men's income?
    - Do Democratic and Republican senators receive different amounts of campaign donations?

t-tests for the difference in two means

- Often we are interested in whether the mean for one group is different from the mean for another group
  - Is woman's income different to men's income?
  - Do Democratic and Republican senators receive different amounts of campaign donations?
- t-tests can also be used to compare the means of two groups

t-tests for the difference in two means

- Often we are interested in whether the mean for one group is different from the mean for another group
  - Is woman's income different to men's income?
  - Do Democratic and Republican senators receive different amounts of campaign donations?
- t-tests can also be used to compare the means of two groups
- Requires an interval-level dependent variable (Y) and binary independent variable (X)

t-tests for the difference in two means

- What is the null hypothesis?

t-tests for the difference in two means

- What is the null hypothesis?
  - There is no difference between the means of the two groups in the population

t-tests for the difference in two means

- What is the null hypothesis?
  - There is no difference between the means of the two groups in the population
- The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{SE(Y_{X=0} - Y_{X=1})}$$

t-tests for the difference in two means

- What is the null hypothesis?
  - There is no difference between the means of the two groups in the population
- The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{SE(Y_{X=0} - Y_{X=1})} = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{Y_{X=0}}}{n_{X=0}} + \frac{s^2_{Y_{X=1}}}{n_{X=1}}}}$$

- Where $s^2_{Y_{X=0}}$ and $s^2_{Y_{X=1}}$ are the sample variances for each group

t-tests for the difference in two means

- What is the null hypothesis?
  - There is no difference between the means of the two groups in the population
- The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{SE(Y_{X=0} - Y_{X=1})} = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{Y_{X=0}}}{n_{X=0}} + \frac{s^2_{Y_{X=1}}}{n_{X=1}}}}$$

- Where $s^2_{Y_{X=0}}$ and $s^2_{Y_{X=1}}$ are the sample variances for each group
  - The variance (s^2_Y) is just the standard deviation (s_Y) squared

t-tests for the difference in two means

- What is the null hypothesis?
  - There is no difference between the means of the two groups in the population
- The test statistic for the difference in means (for a null hypothesis of no difference) is

$$t = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{SE(Y_{X=0} - Y_{X=1})} = \frac{\bar{Y}_{X=0} - \bar{Y}_{X=1}}{\sqrt{\frac{s^2_{Y_{X=0}}}{n_{X=0}} + \frac{s^2_{Y_{X=1}}}{n_{X=1}}}}$$

- Where $s^2_{Y_{X=0}}$ and $s^2_{Y_{X=1}}$ are the sample variances for each group
  - The variance (s^2_Y) is just the standard deviation (s_Y) squared
- $n_{X=0}$ and $n_{X=1}$ are the number of observations for each group

Test for the difference in means: critical value of t

- Assuming that sample size is large (> 30), the critical t value is 1.96
- To know the exact critical value, we need to know the degrees of freedom (df)
- You could do it in R using the t.test() function which computes the correct number of df for you

Test for the difference in means: p value

- Once we know the correct t value, getting the p value is the same as in the t-test for the sample mean if the sample is large
- If the sample is small, use R's t.test() function

Test for the difference in means: R

- You need a continuous dependent variable (DV)
- A binary independent variable (IV)
- Unless stated otherwise, the null is usually there is no difference in means. Hence, mu $= 0$

```
t.test(DV ~ IV, mu = 0, conf = 0.95)
```

Overview

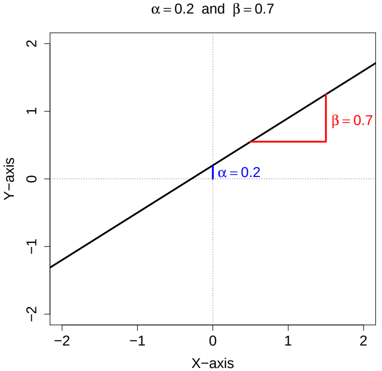
1. Administrative information
2. Answer advice
3. Hypothesis testing
4. Simple linear regression
5. Multiple linear regression

Simple linear regression: intuition

- How are two phenomena (X and Y) related?

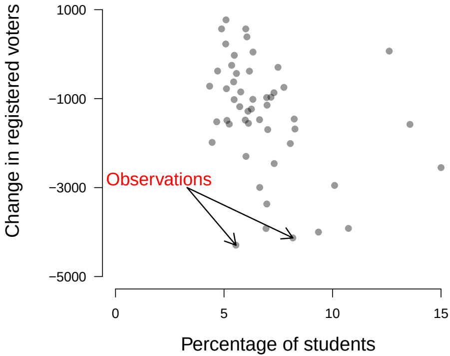
Linear relationships

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



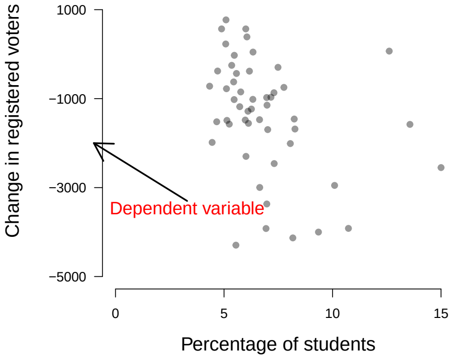
- α is the intercept: the value of Y where $X = 0$
- β is the slope: the amount that Y increases when X increases by one unit
- Here, a one-unit increase in X is associated with a 0.7-unit increase in Y

The linear regression line



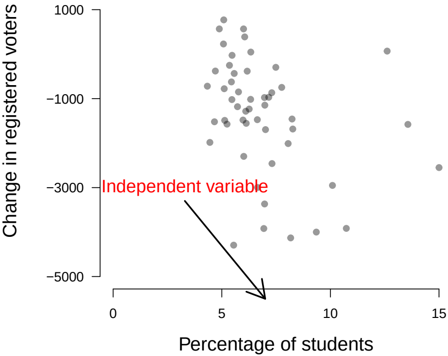
- Observations $i = 1, \ldots, n$

The linear regression line



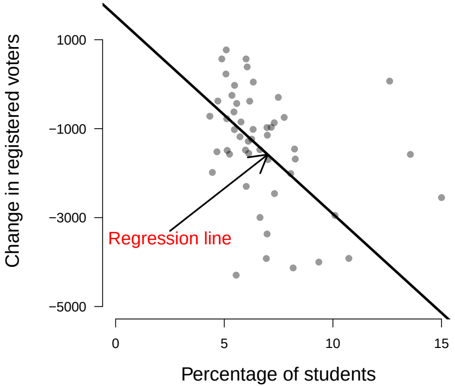
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.

The linear regression line



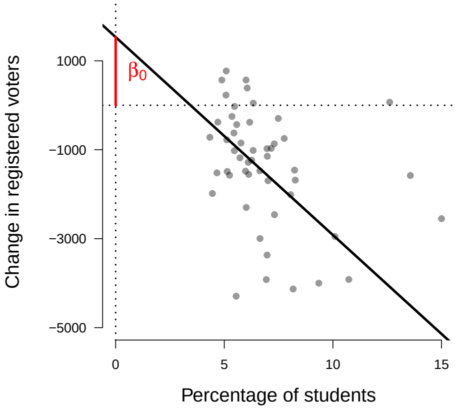
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.

The linear regression line



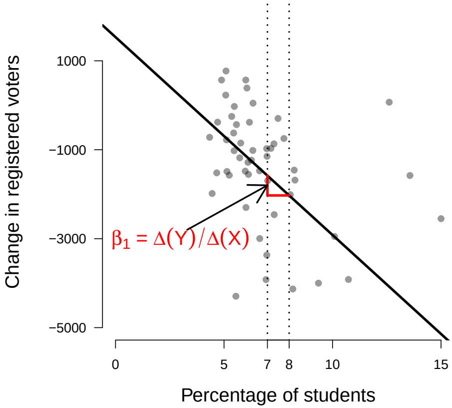
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.

The linear regression line



- Observations $i = 1, \dots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.
- β_0 is the intercept.

The linear regression line



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

Application to voter registration

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

DV: Δvoters	$\hat{\beta}_k$, $(\hat{\sigma}_{\hat{\beta}_k})$
(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

DV: Δvoters	$\hat{\beta}_k$, ($\hat{\sigma}_{\hat{\beta}_k}$)
(Intercept)	1532.69
	(192.41)
students	-444.97
	(26.99)
R^2	0.32
N.	573

- To test the government's hypothesis:

Application to voter registration

DV: Δvoters	$\hat{\beta}_k$, $(\hat{\sigma}_{\hat{\beta}_k})$
(Intercept)	1532.69
	(192.41)
students	-444.97
	(26.99)
R^2	0.32
N.	573

- To test the government's hypothesis:

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

Application to voter registration

DV: Δvoters	$\hat{\beta}_k$, $(\hat{\sigma}_{\hat{\beta}_k})$
(Intercept)	1532.69
	(192.41)
students	-444.97
	(26.99)
R^2	0.32
N.	573

- To test the government's hypothesis:

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

Application to voter registration

DV: Δvoters	$\hat{\beta}_k$, $(\hat{\sigma}_{\hat{\beta}_k})$
(Intercept)	1532.69
	(192.41)
students	−444.97
	(26.99)
R^2	0.32
N.	573

- To test the government's hypothesis:

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

Application to voter registration

DV: Δvoters	$\hat{\beta}_k$, $(\hat{\sigma}_{\hat{\beta}_k})$
(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
```

Overview

Multiple linear regression: intuition

- We can control for confounders with multiple linear regression

More than two independent variables

```
## Specify the model with 3 independent variables
linear_model_3 <- lm(AfD ~ christian + east
+ migrantfraction , data = results)

## Output in a nice format
screenreg(list(linear_model_1, linear_model_2, linear_model_3))

===================================================
                 Model 1      Model 2      Model 3
---------------------------------------------------
(Intercept)        21.29 ***     7.82 ***   11.78 ***
                   (0.76)       (1.30)      (1.90)
christian          -0.16 ***     0.03        0.00
                   (0.01)       (0.02)      (0.02)
eastTRUE                        11.77 ***    9.14 ***
                                (0.99)      (1.35)
migrantfraction                             -0.09 **
                                            (0.03)
---------------------------------------------------
R^2                 0.36         0.56        0.58
Adj. R^2            0.35         0.56        0.57
Num. obs.         299          299         299
===================================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

More than two independent variables

```
## Specify the model with 3 independent variables
linear_model_3 <- lm(AfD ~ christian + east
+ migrantfraction , data = results)

## Output in a nice format
screenreg(list(linear_model_1, linear_model_2, linear_model_3))

===================================================
                 Model 1      Model 2      Model 3
---------------------------------------------------
(Intercept)       21.29 ***    7.82 ***    11.78 ***
                  (0.76)      (1.30)       (1.90)
christian         -0.16 ***    0.03         0.00
                  (0.01)      (0.02)       (0.02)
eastTRUE                      11.77 ***     9.14 ***
                              (0.99)       (1.35)
migrantfraction                            -0.09 **
                                           (0.03)
---------------------------------------------------
R^2                0.36        0.56         0.58
Adj. R^2           0.35        0.56         0.57
Num. obs.        299         299          299
===================================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

- The coefficient on `migrantfraction` ($\hat{\beta}_3$) is negative and significant

More than two independent variables

```
## Specify the model with 3 independent variables
linear_model_3 <- lm(AfD ~ christian + east
+ migrantfraction , data = results)

## Output in a nice format
screenreg(list(linear_model_1, linear_model_2, linear_model_3))

=====================================================
                 Model 1      Model 2      Model 3
-----------------------------------------------------
(Intercept)       21.29 ***    7.82 ***    11.78 ***
                  (0.76)      (1.30)       (1.90)
christian         -0.16 ***    0.03         0.00
                  (0.01)      (0.02)       (0.02)
eastTRUE                      11.77 ***     9.14 ***
                              (0.99)       (1.35)
migrantfraction                            -0.09 **
                                           (0.03)
-----------------------------------------------------
R^2                0.36        0.56         0.58
Adj. R^2           0.35        0.56         0.57
Num. obs.        299         299          299
=====================================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

- The coefficient on `migrantfraction` ($\hat{\beta}_3$) is negative and significant
- The coefficient on `east` ($\hat{\beta}_2$) is smaller in model 3

More than two independent variables

```
## Specify the model with 3 independent variables
linear_model_3 <- lm(AfD ~ christian + east
+ migrantfraction , data = results)

## Output in a nice format
screenreg(list(linear_model_1, linear_model_2, linear_model_3))

===================================================
                 Model 1      Model 2     Model 3
---------------------------------------------------
(Intercept)       21.29 ***    7.82 ***   11.78 ***
                  (0.76)      (1.30)      (1.90)
christian         -0.16 ***    0.03        0.00
                  (0.01)      (0.02)      (0.02)
eastTRUE                      11.77 ***    9.14 ***
                              (0.99)      (1.35)
migrantfraction                           -0.09 **
                                          (0.03)
---------------------------------------------------
R^2                0.36        0.56        0.58
Adj. R^2           0.35        0.56        0.57
Num. obs.        299         299         299
===================================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

- The coefficient on `migrantfraction` ($\hat{\beta}_3$) is negative and significant
- The coefficient on east ($\hat{\beta}_2$) is smaller in model 3
- The R^2 has increased