PUBL0055: Introduction to Quantitative Methods

Lecture 4: Regression I

Michal Ovádek & Indraneel Sircar

Lecture Outline

Data and Models

The (Simple) Linear Regression Model

Estimation and Interpretation

Measures of fit

Regression and the difference in means

Conclusion

Data and Models

Motivation

In previous weeks, we have mostly focussed on describing how our dependent variable varies as a function of a binary variable (i.e. difference in means).

Last week, we also saw a statistic for describing the association between two continuous variables (the correlation coefficient).

This week, we introduce regression, which can incorporate both of these types of relationship, and offers a flexible framework for building more sophisticated analyses.

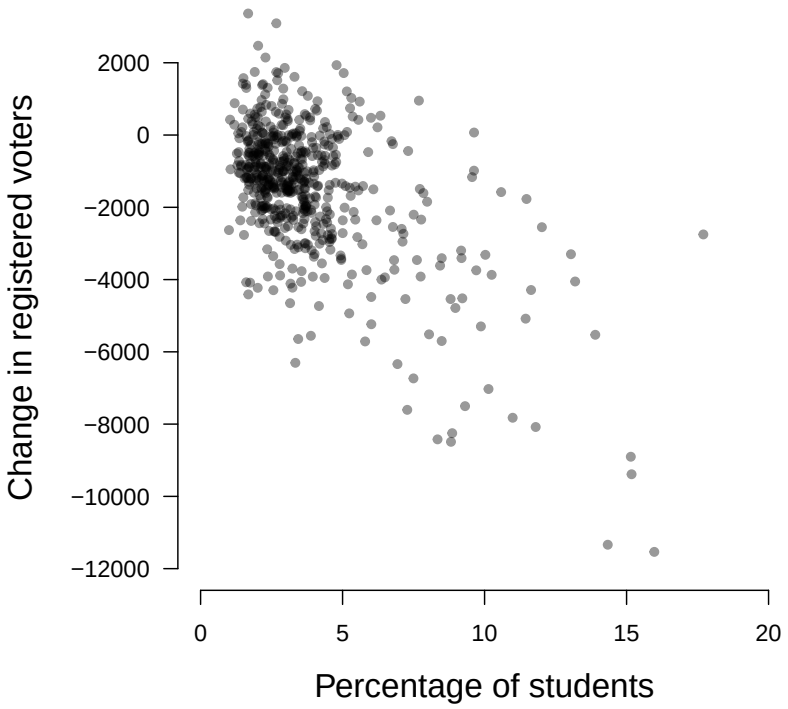
Data

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 (all constituencies in England and Wales).
- 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.

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.

What is a model?

- A model is a simplified abstraction of reality
- Typically, models are used to describe key features or dimensions of some more complicated process
- We do not want a model to capture *every* feature of reality, just those features that help us to describe the phenomena of interest
- "All models are wrong, but some are useful" – George Box
- We will be using statistical models which will always be "wrong", but some will be useful

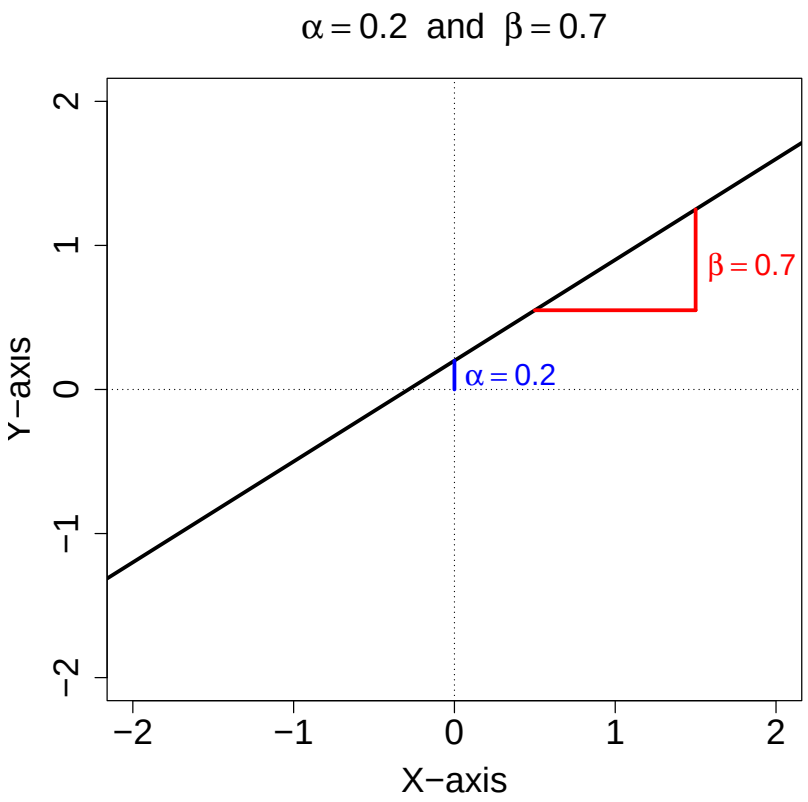
The (Simple) Linear Regression Model

Linear relationships

- The most straightforward way of describing the relationship between two variables is with a line
- 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

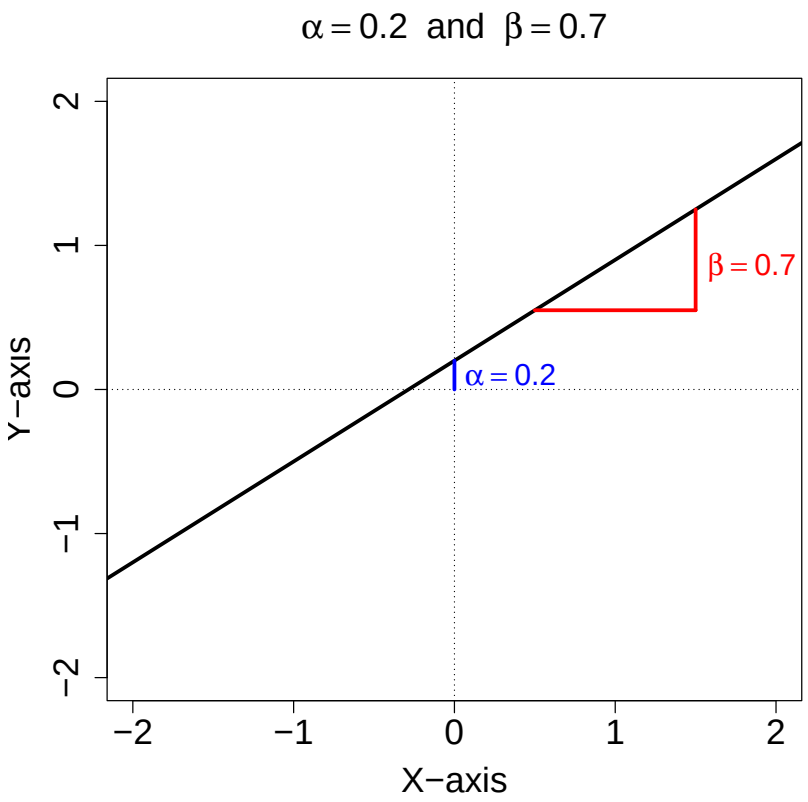
A line can be represented $Y = \alpha + \beta X$



- α is the intercept: the value of Y where $X = 0$

Linear relationships

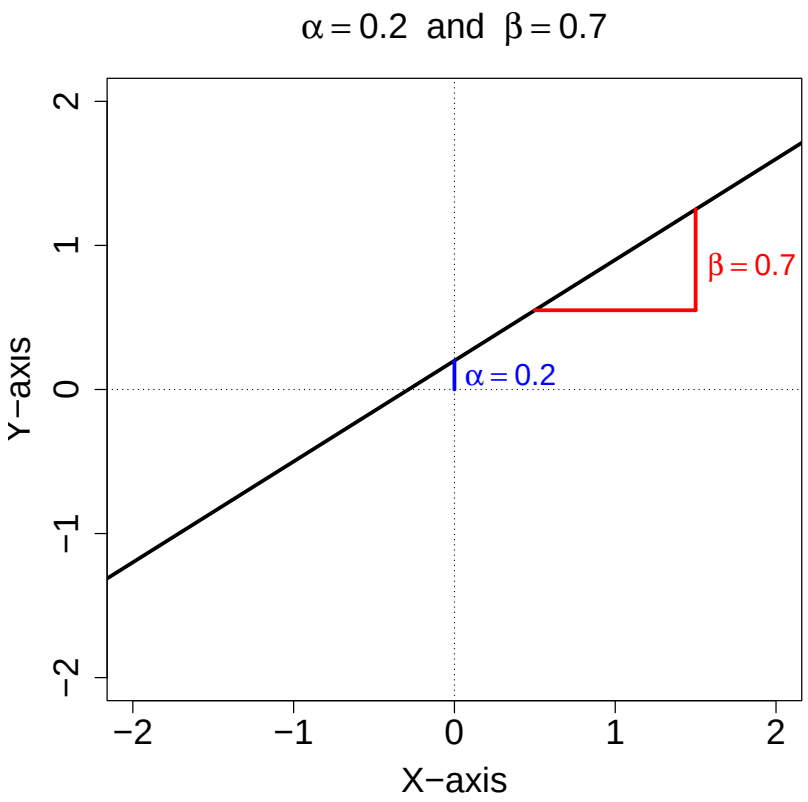
A line can be represented $Y = \alpha + \beta X$



- 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

Linear relationships

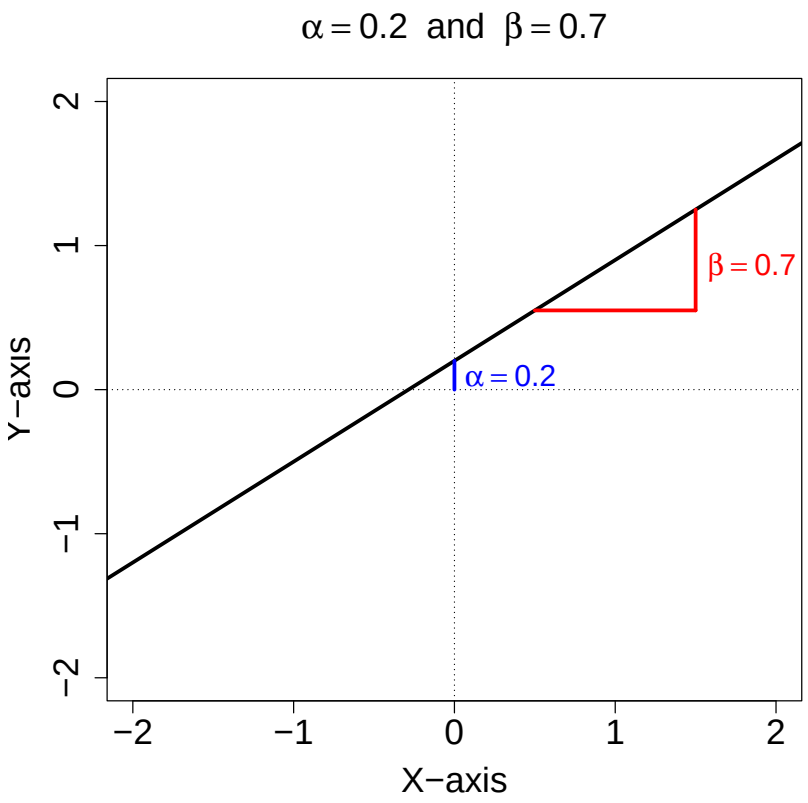
A line can be represented $Y = \alpha + \beta X$

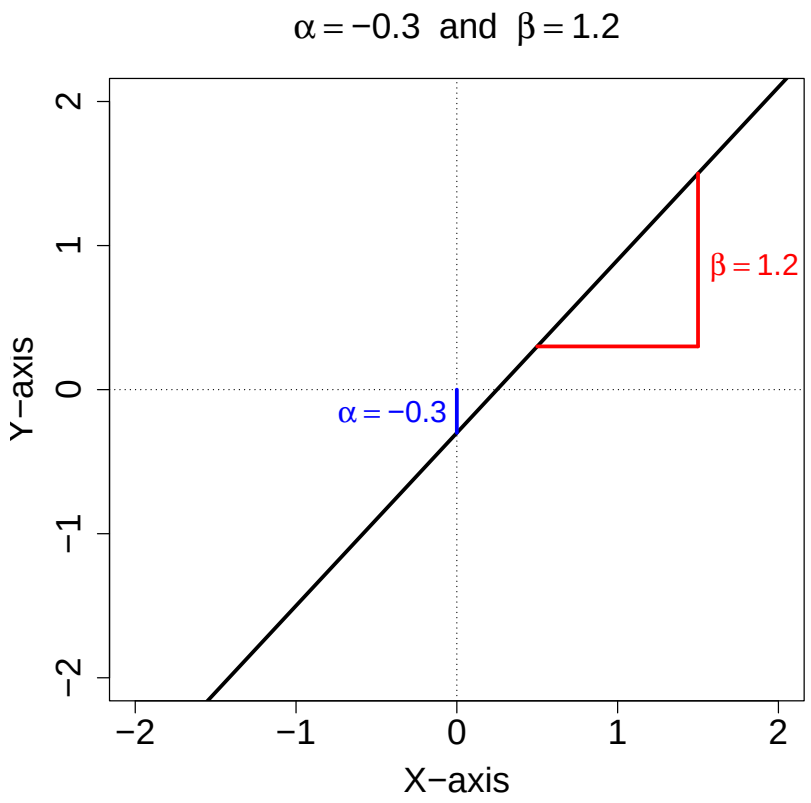


- α is the intercept: the value of Y where $X = 0$
- β is the slope: the amount that Y increases when X increases by one unit
- Here, a one-unit increase in X is associated with a 0.7-unit increase in Y

Linear relationships

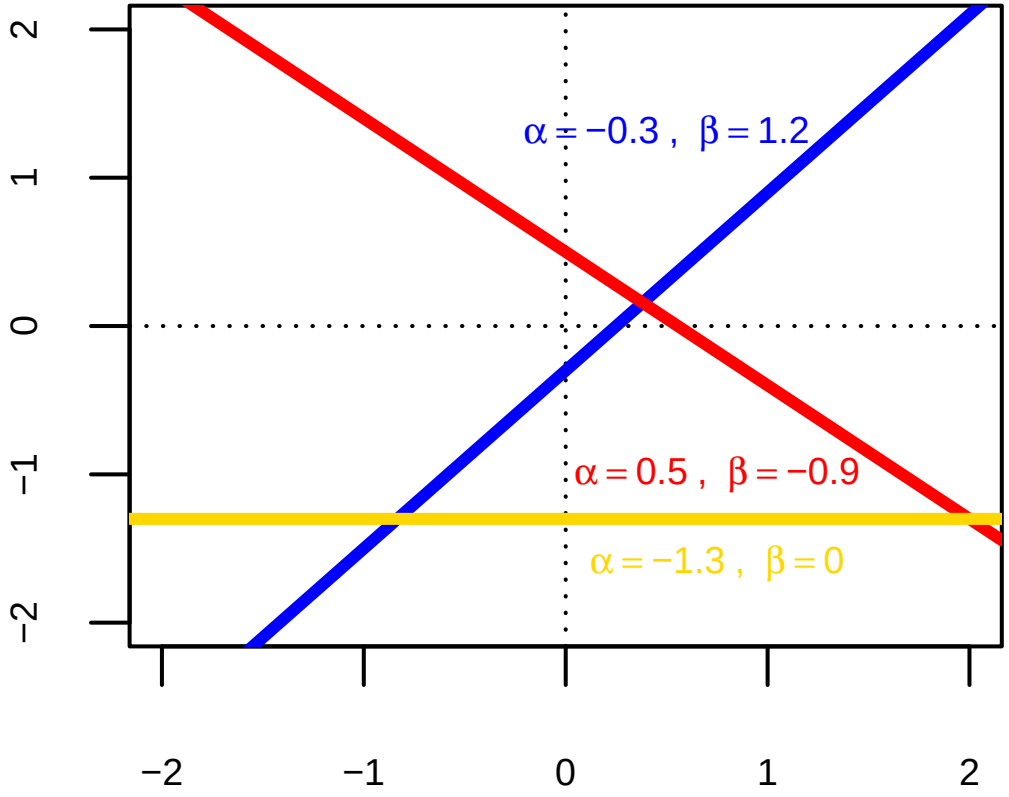
Different values of α and β uniquely define different lines





Linear relationships

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



The linear regression model

A simple way to summarize the relationship between two variables is to assume that they are linearly related.

We can express this with the simple linear regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Observations $i = 1, \dots, n$
- Y is the dependent variable
- X is the independent variable
- α ("alpha") is the intercept or constant
- β ("beta") is the slope
- ϵ_i ("epsilon") is the error term or residual

The linear regression model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

α and β are the coefficients or parameters of the regression line.

- α gives the average value of Y when X is equal to 0
- β gives the average change in Y that results from a 1-unit change in X
- $\rightarrow$ describe the relationship that holds, on average, between X and Y

The linear regression model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

α and β are the coefficients or parameters of the regression line.

- α gives the average value of Y when X is equal to 0
- β gives the average change in Y that results from a 1-unit change in X
- $\rightarrow$ describe the relationship that holds, on average, between X and Y

ϵ_i is the error term

- ϵ_i allows a unit to deviate from a perfect linear relationship
- $\rightarrow$ represents all factors aside from X that determine the value of Y

The linear regression model (example)

In our voter registration example:

- Y_i – change in number of registered voters in constituency i
- X_i – percentage of students in constituency i
- ϵ_i – all factors influencing registration other than student population

What does α represent?

The linear regression model (example)

In our voter registration example:

- Y_i – change in number of registered voters in constituency i
- X_i – percentage of students in constituency i
- ϵ_i – all factors influencing registration other than student population

What does α represent?

- the average change in registration for a constituency with 0% students

What does β represent?

The linear regression model (example)

In our voter registration example:

- Y_i – change in number of registered voters in constituency i
- X_i – percentage of students in constituency i
- ϵ_i – all factors influencing registration other than student population

What does α represent?

- the average change in registration for a constituency with 0% students

What does β represent?

- the average change in registration associated with a one unit change in the percentage of students

What is a "one-unit" change?

If β represents the change in Y that is associated with a a "one-unit" change in X, we need to know the units in which X is measured.

For example, a "one-unit" increase in...

- ...*age*, measured in years, is *one year*
- ...*height*, measured in inches, is *one inch*
- ...*GDP per capita*, measured in dollars, is *one dollar*

What is a "one-unit" change?

If β represents the change in Y that is associated with a a "one-unit" change in X, we need to know the units in which X is measured.

For example, a "one-unit" increase in...

- ...*age*, measured in years, is *one year*
- ...*height*, measured in inches, is *one inch*
- ...*GDP per capita*, measured in dollars, is *one dollar*

Question: What is a one-unit increase in the "percentage of students"?

What is a "one-unit" change?

If β represents the change in Y that is associated with a a "one-unit" change in X, we need to know the units in which X is measured.

For example, a "one-unit" increase in...

- ...*age*, measured in years, is *one year*
- ...*height*, measured in inches, is *one inch*
- ...*GDP per capita*, measured in dollars, is *one dollar*

Question: What is a one-unit increase in the "percentage of students"?

Answer: A one **percentage point** increase in the percentage of students.

“Percentage” versus “percentage point”

A frequent interpretational error is to confuse *percentage* changes with *percentage point* changes. What’s the difference?

An increase in the percentage of students from 40% to 44% represents:

- An increase of 4 **percentage points**
- An increase of 10 **percent**

When including percentage variables in regression models, we will (almost) always speak about changes in **percentage points**.

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- α & β represent the average relationship between X and Y
  - They are population parameters – values we assume exist in the world
- We would like to know the numerical values that α and β take
- We don't know these values so we must estimate them
- We estimate the values of the parameters from the data
- We use a slightly different notation to indicate estimated parameters
  - α becomes $\hat{\alpha}$, which reads as "alpha hat"
  - β becomes $\hat{\beta}$, which reads as "beta hat"

Fitted values

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

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

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i, \ i = 1, \dots, n$$

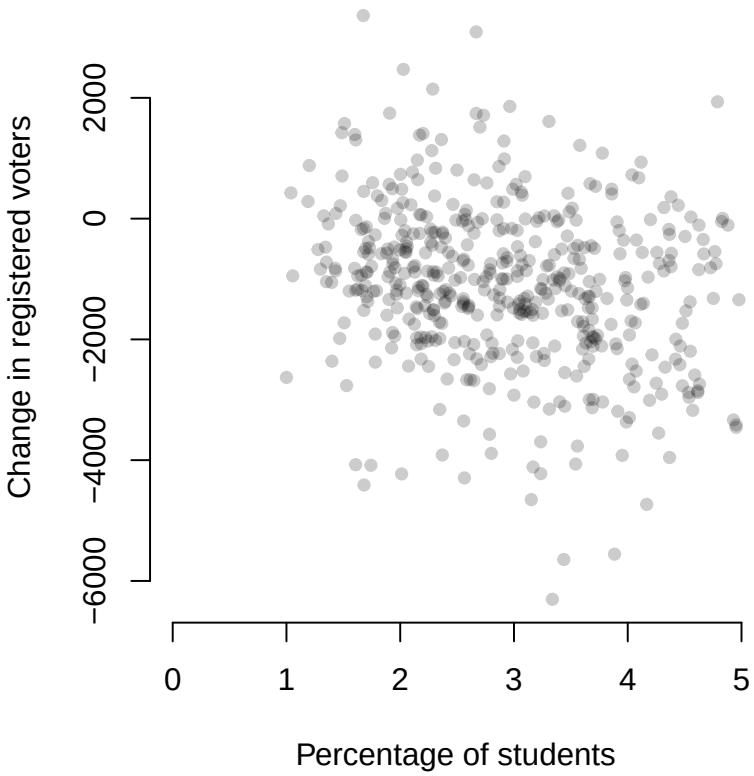
The fitted values tell us what the best guess is for Y for a specific value of X.

The residuals $\hat{\epsilon}_i$ are

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i, \ i = 1, \dots, n.$$

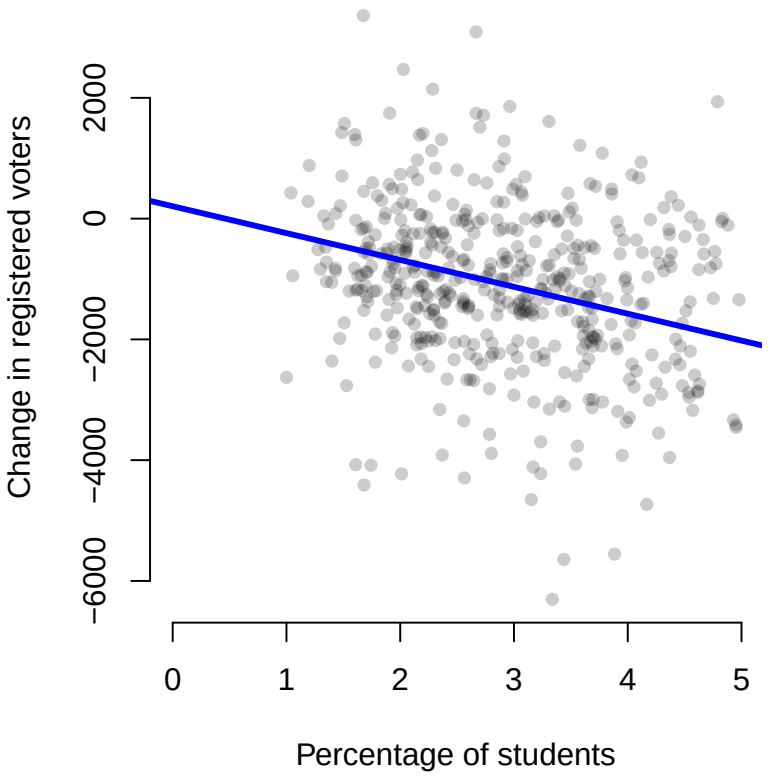
The residuals tell us how far our best guess for each observation is from the value of Y we observe in the sample.

The linear regression line



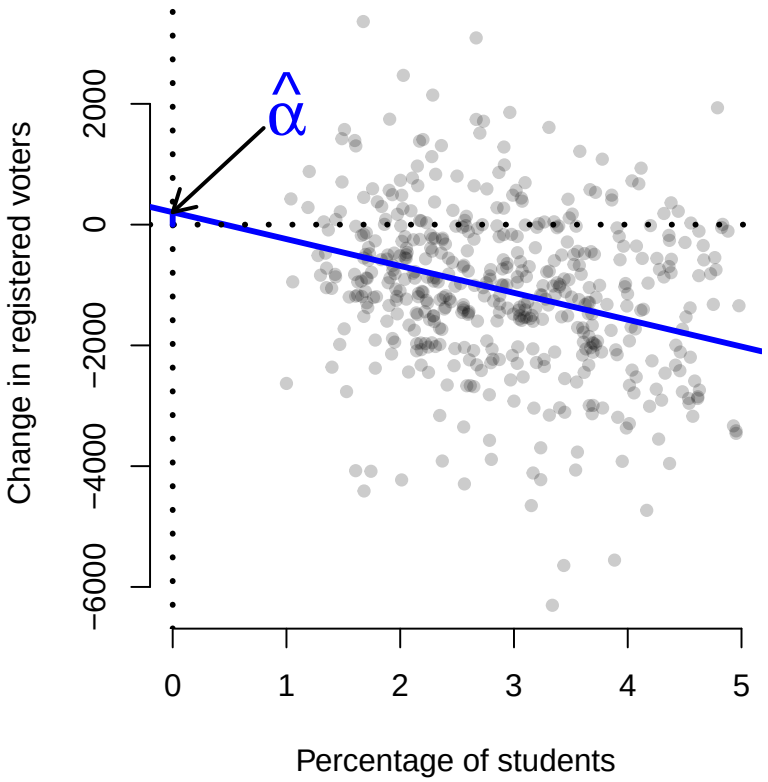
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.

The linear regression line



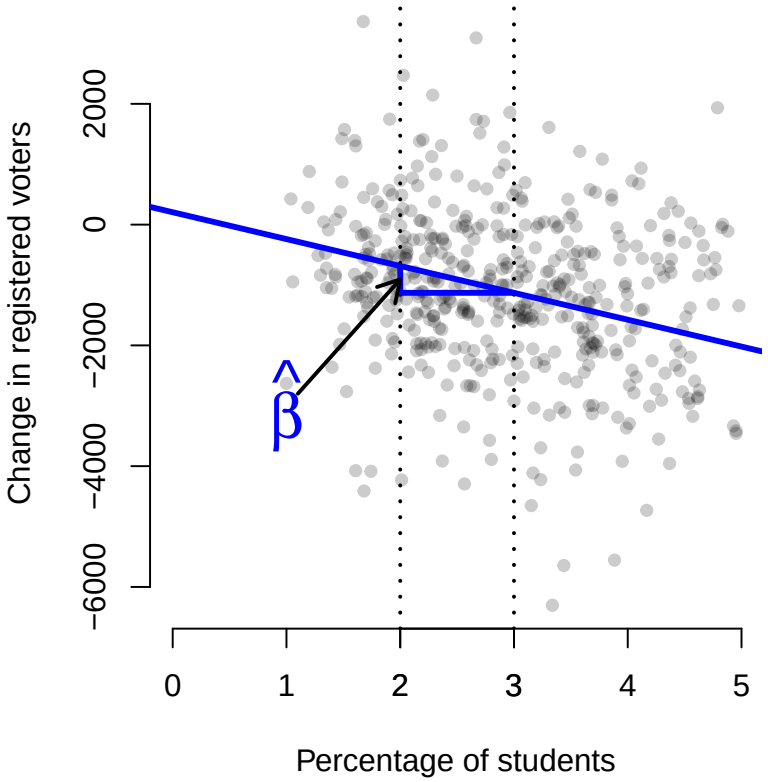
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.

The linear regression line



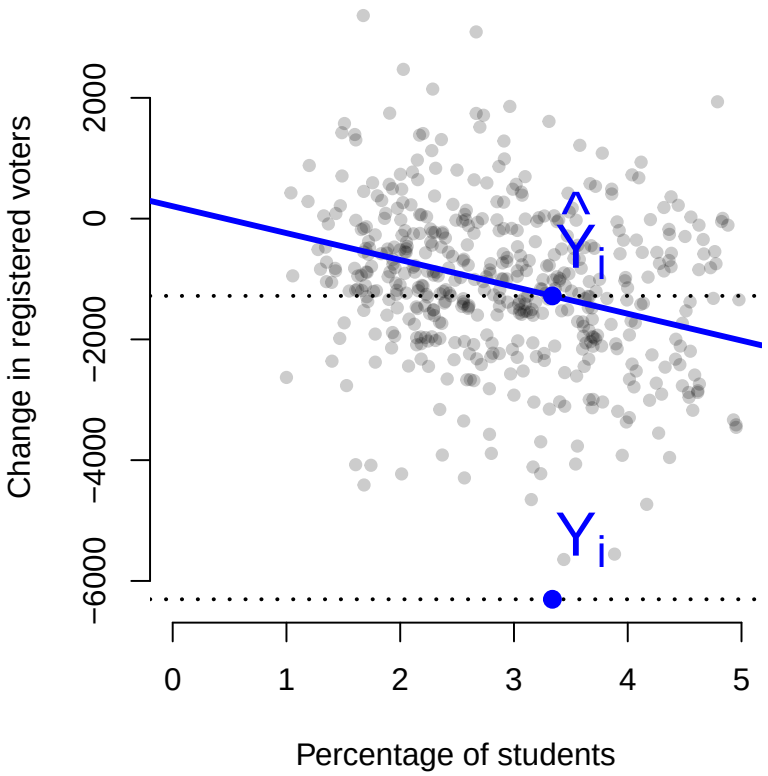
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.
- $\hat{\alpha}$ is the intercept.

The linear regression line



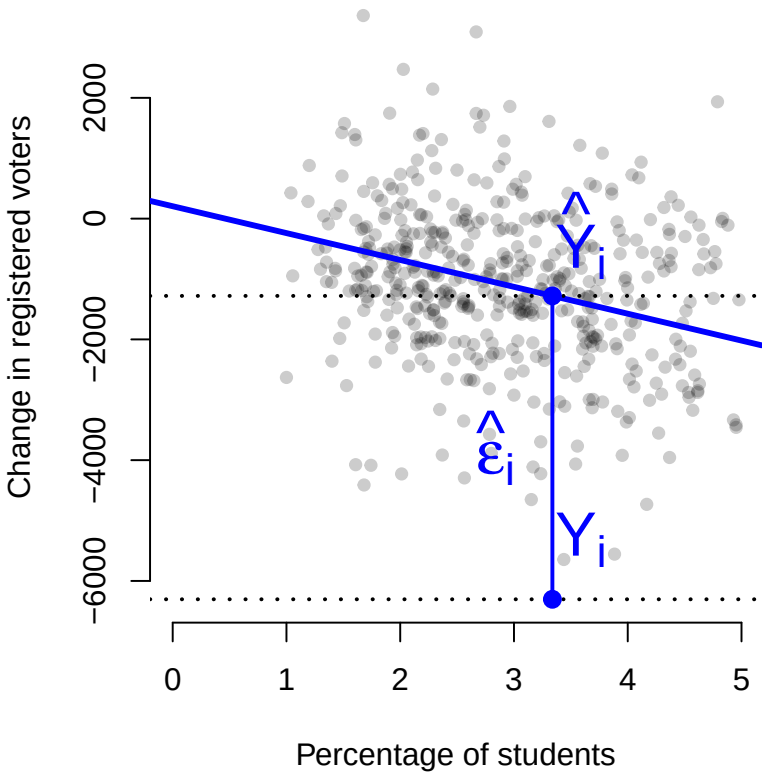
- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.
- $\hat{\alpha}$ is the intercept.
- $\hat{\beta}$ is the slope.

The linear regression line



- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.
- $\hat{\alpha}$ is the intercept.
- $\hat{\beta}$ is the slope.
- $\hat{Y}_i$ the fitted value
- Y_i the observed value of Y

The linear regression line



- Observations $i = 1, \ldots, n$
- Y is the dependent variable.
- X is the independent variable.
- The regression line.
- $\hat{\alpha}$ is the intercept.
- $\hat{\beta}$ is the slope.
- $\hat{Y}_i$ the fitted value
- Y_i the observed value of Y
- $\hat{\epsilon}_i$ is the error term.

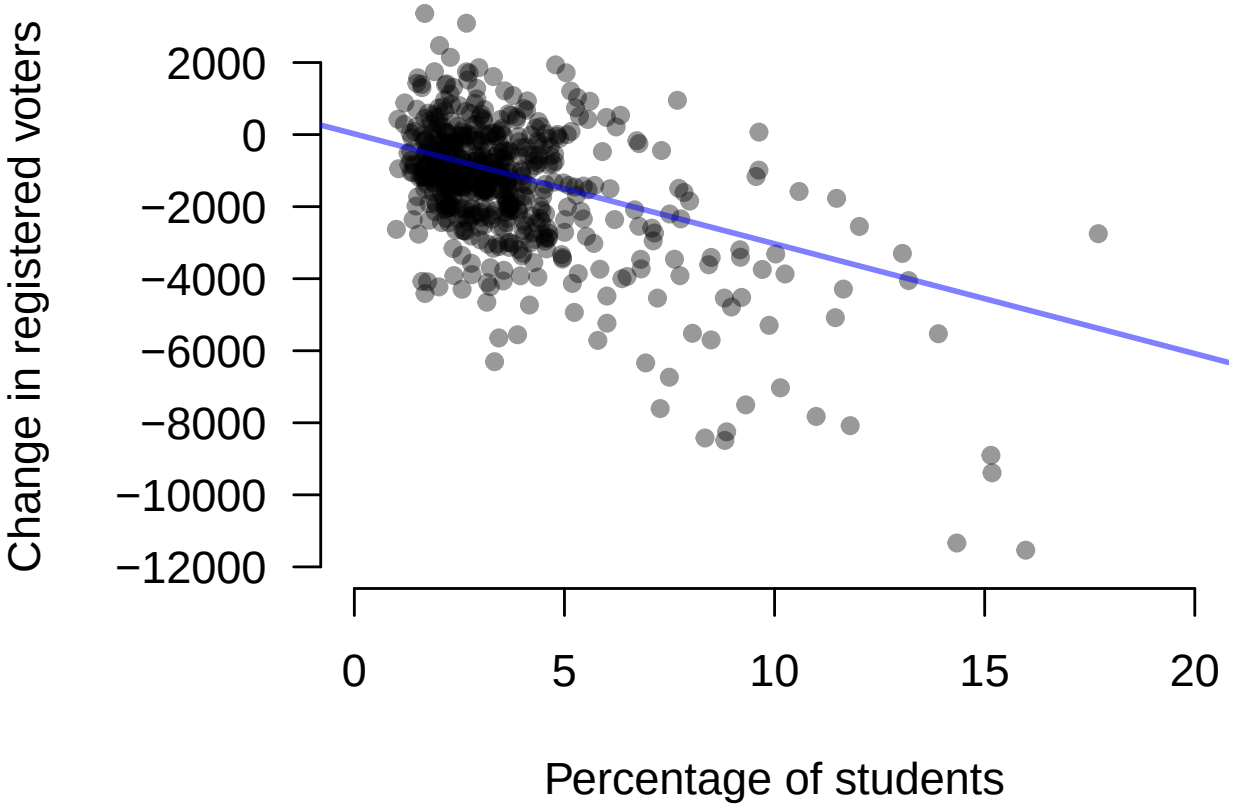
Estimation and Interpretation

Estimating α and β

The main goal of the simple regression model is to estimate a line that "fits" the data. Which of these lines best "fits" our data?

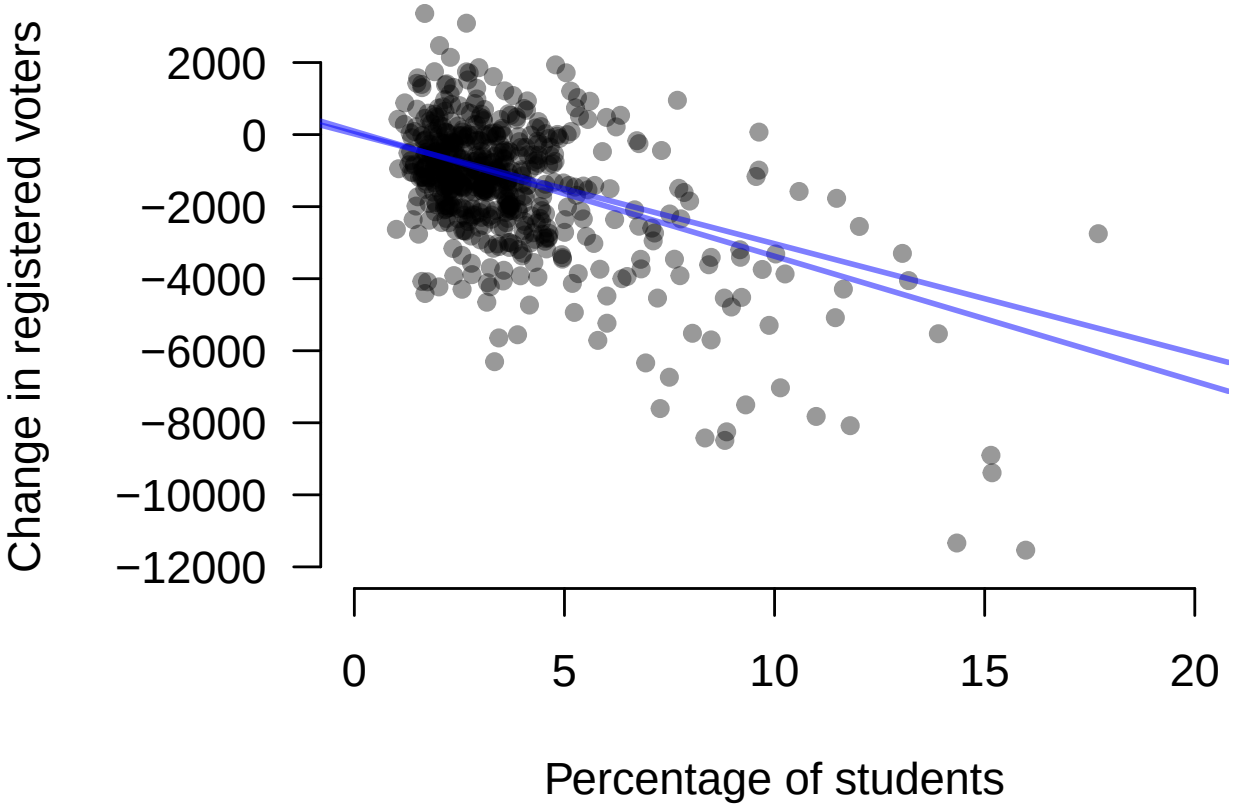
Estimating α and β

The main goal of the simple regression model is to estimate a line that "fits" the data. Which of these lines best "fits" our data?



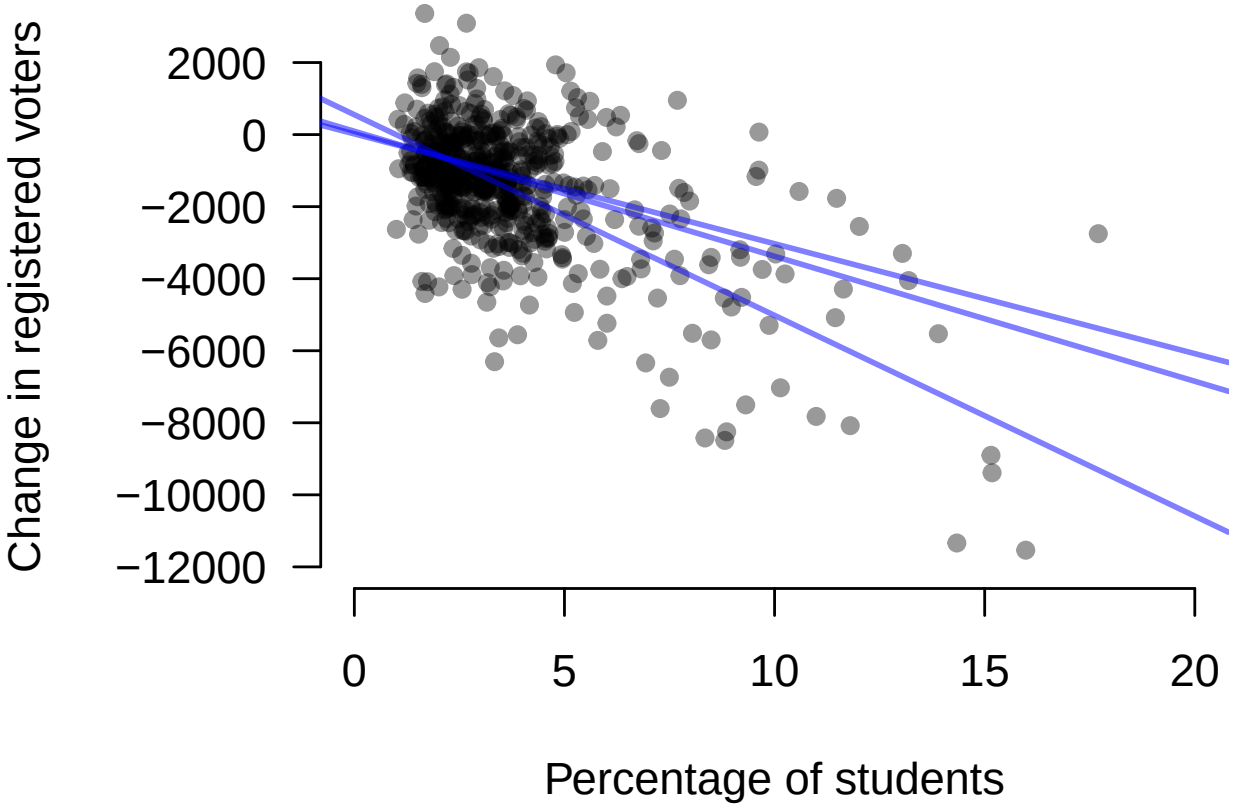
Estimating α and β

The main goal of the simple regression model is to estimate a line that "fits" the data. Which of these lines best "fits" our data?



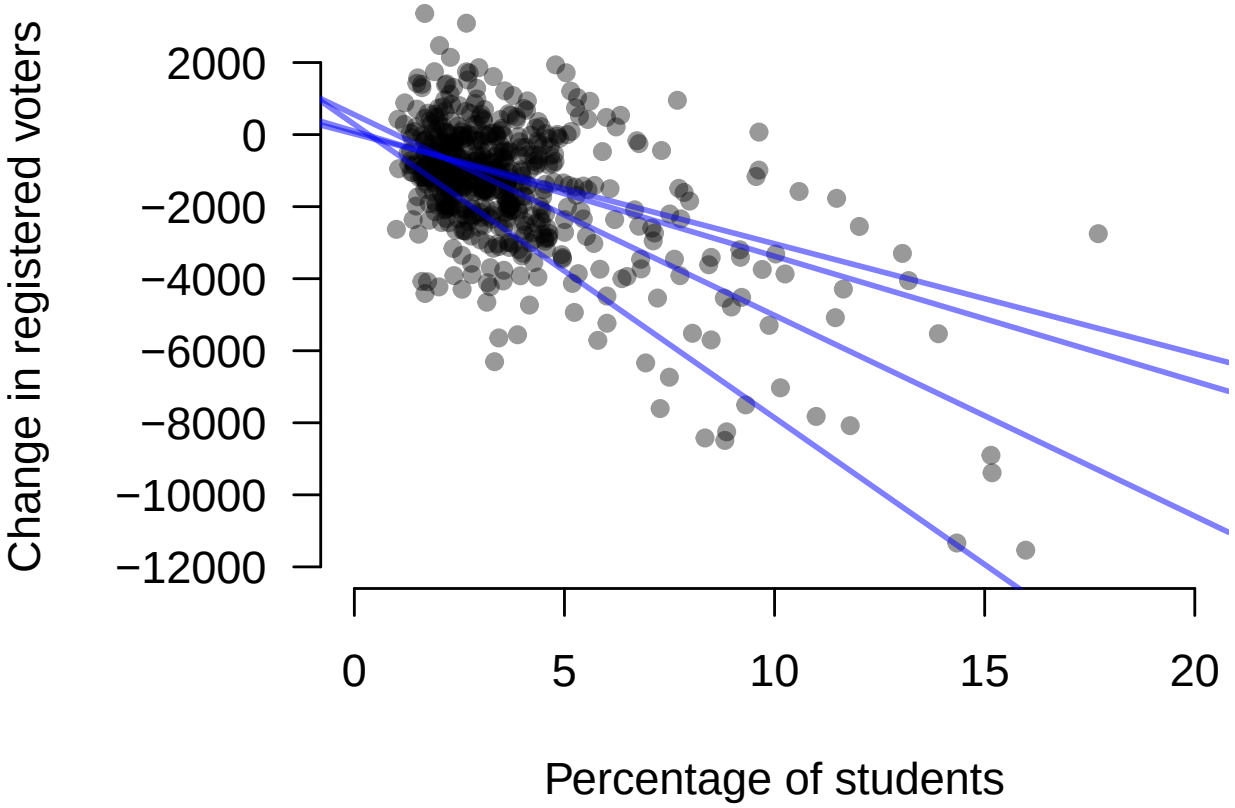
Estimating α and β

The main goal of the simple regression model is to estimate a line that "fits" the data. Which of these lines best "fits" our data?



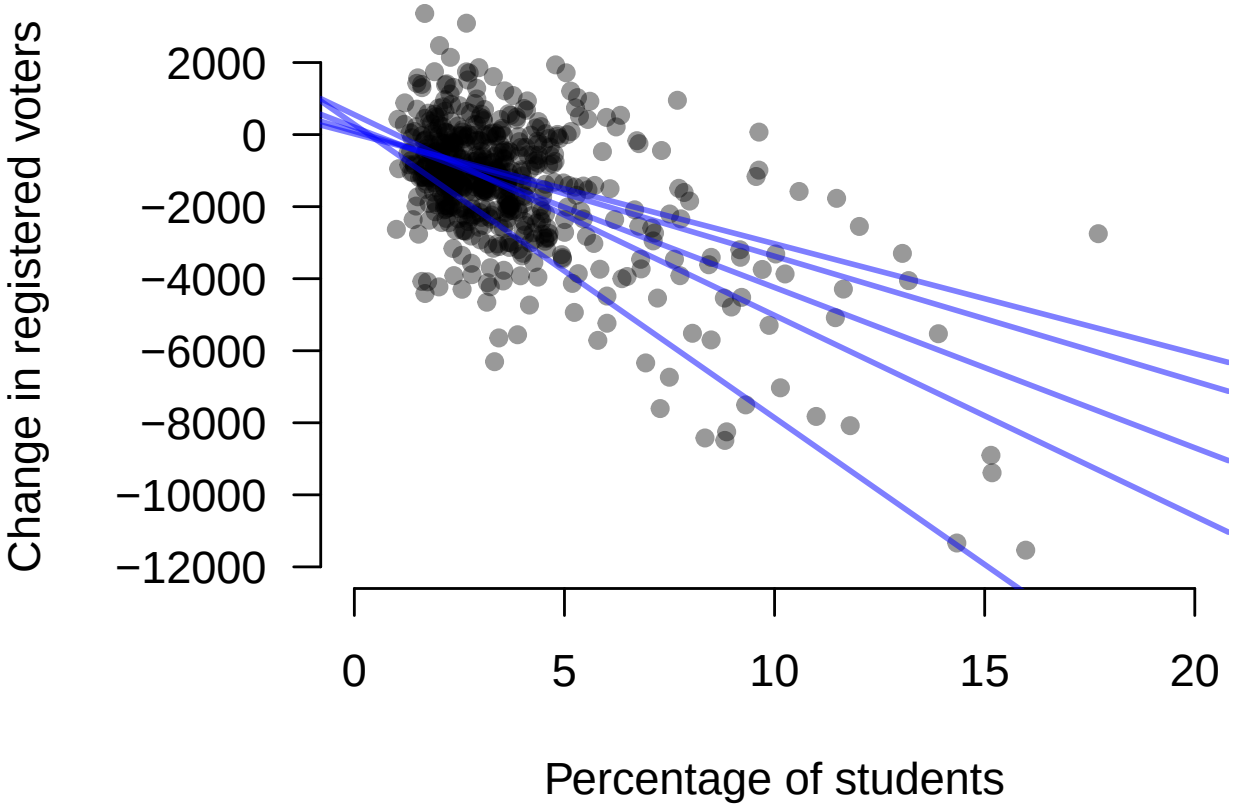
Estimating α and β

The main goal of the simple regression model is to estimate a line that "fits" the data. Which of these lines best "fits" our data?



Estimating α and β

The main goal of the simple regression model is to estimate a line that "fits" the data. Which of these lines best "fits" our data?



Ordinary Least Squares

- 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 most widely used approach to estimating the parameters of the linear regression model is the ordinary least squares (OLS) method.
- The OLS estimator chooses the regression coefficients so that the estimated line is as close as possible to the data

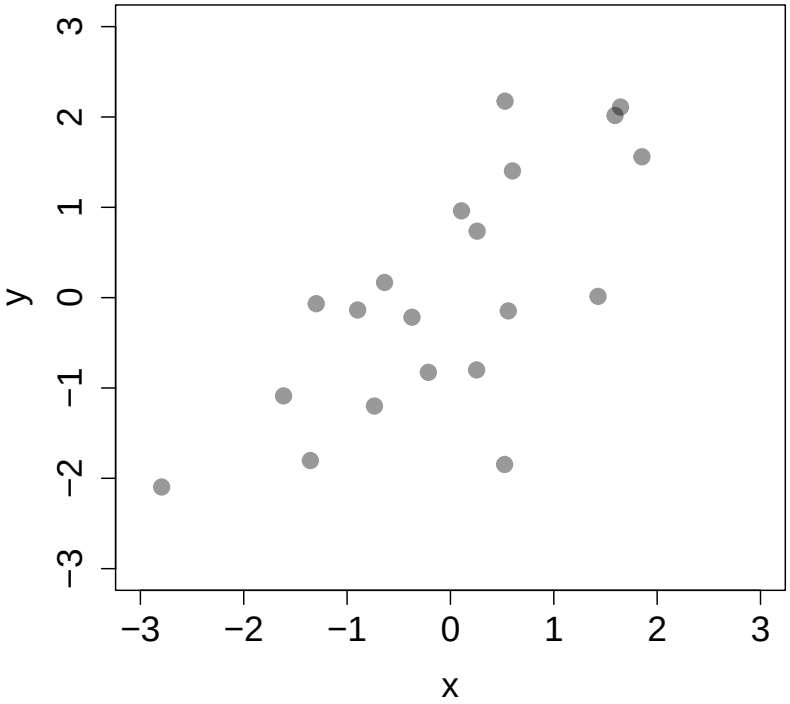
Ordinary Least Squares

- The most widely used approach to estimating the parameters of the linear regression model is the ordinary least squares (OLS) method.
- The OLS estimator chooses the regression coefficients so that the estimated line is as close as possible to the data
- Formally, from all possible α and β values, it chooses $\hat{\alpha}$ and $\hat{\beta}$ that minimize the sum of the squared residuals (SSR)

$$SSR = \sum_{i=1}^{n} \left[Y_i - (\hat{\alpha} + \hat{\beta} X_i) \right]^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

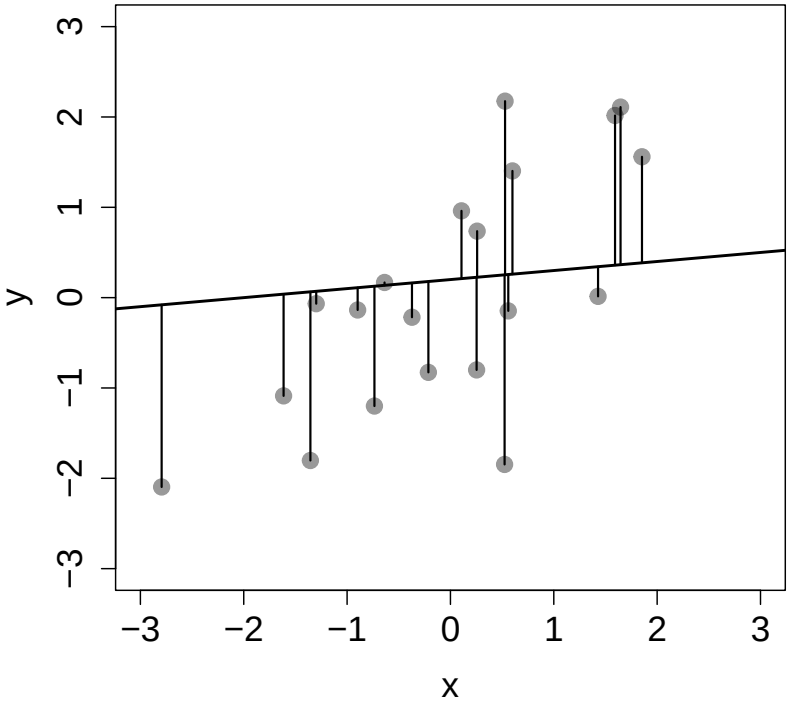
- OLS selects a line that makes the difference between the observed (Y_i) and fitted ($\hat{Y}_i$) values for each observation as small as possible

Ordinary Least Squares (intuition)



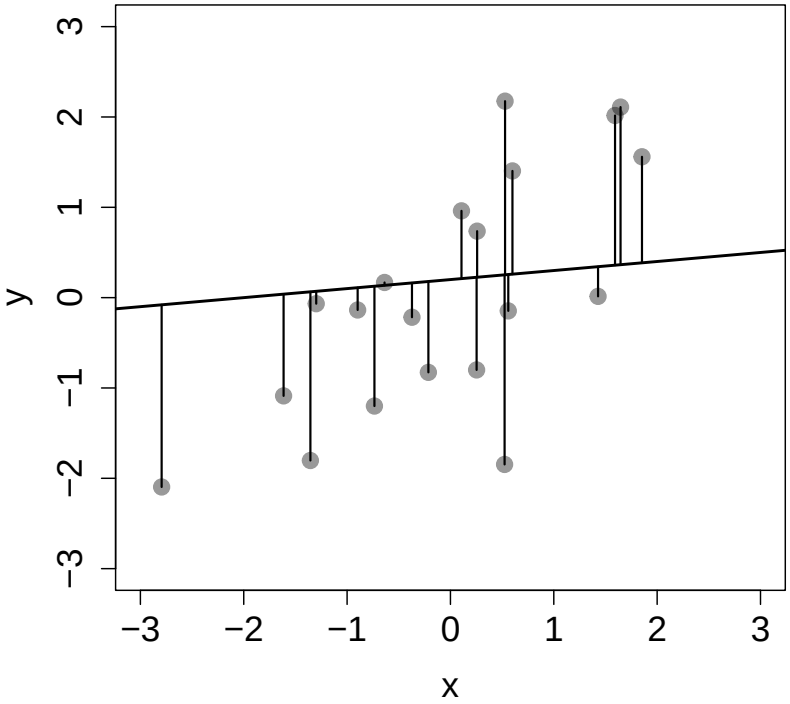
- Take some data

Ordinary Least Squares (intuition)



- Take some data
- Plot a line through the points

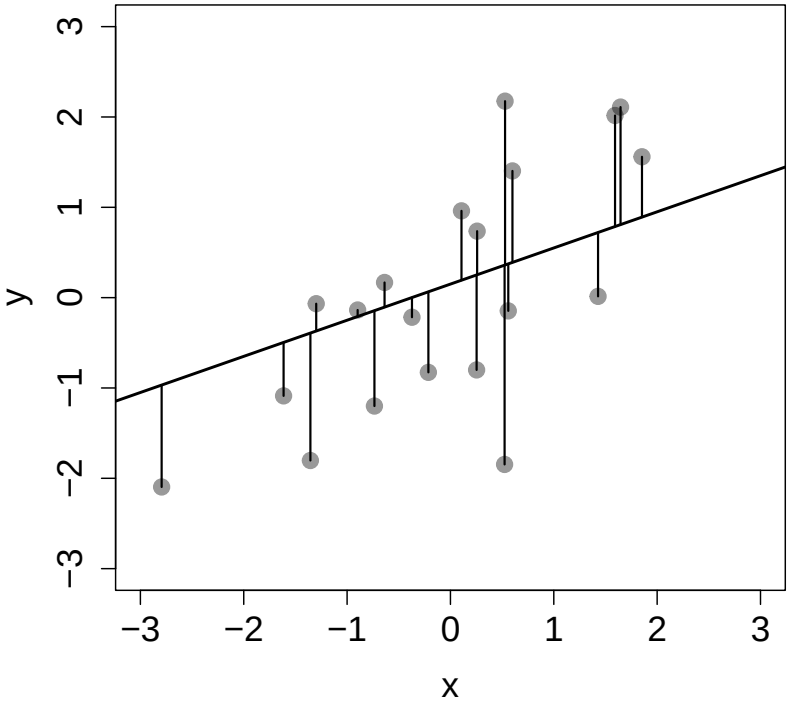
Ordinary Least Squares (intuition)



- Take some data
- Plot a line through the points
- For this line, the sum of the squared distances between Y_i and $\hat{Y}_i$:

$$\sum_{i=1}^{n} \left[Y_i - (\hat{\alpha} + \hat{\beta} X_i) \right]^2 = 30.54$$

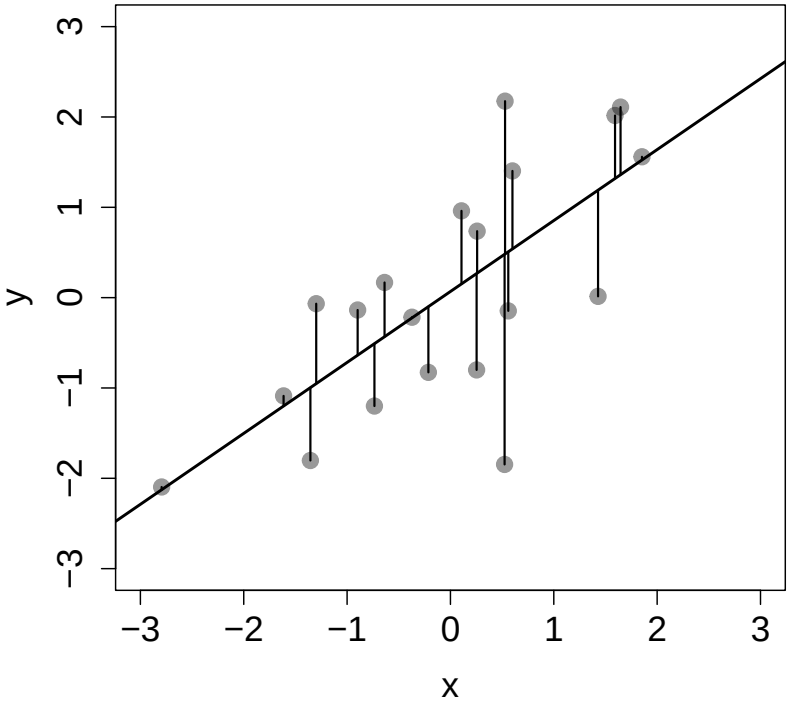
Ordinary Least Squares (intuition)



- Take some data
- Plot a line through the points
- For this line, the sum of the squared distances between Y_i and $\hat{Y}_i$:

$$\sum_{i=1}^{n} \left[Y_i - (\hat{\alpha} + \hat{\beta} X_i) \right]^2 = 21.28$$

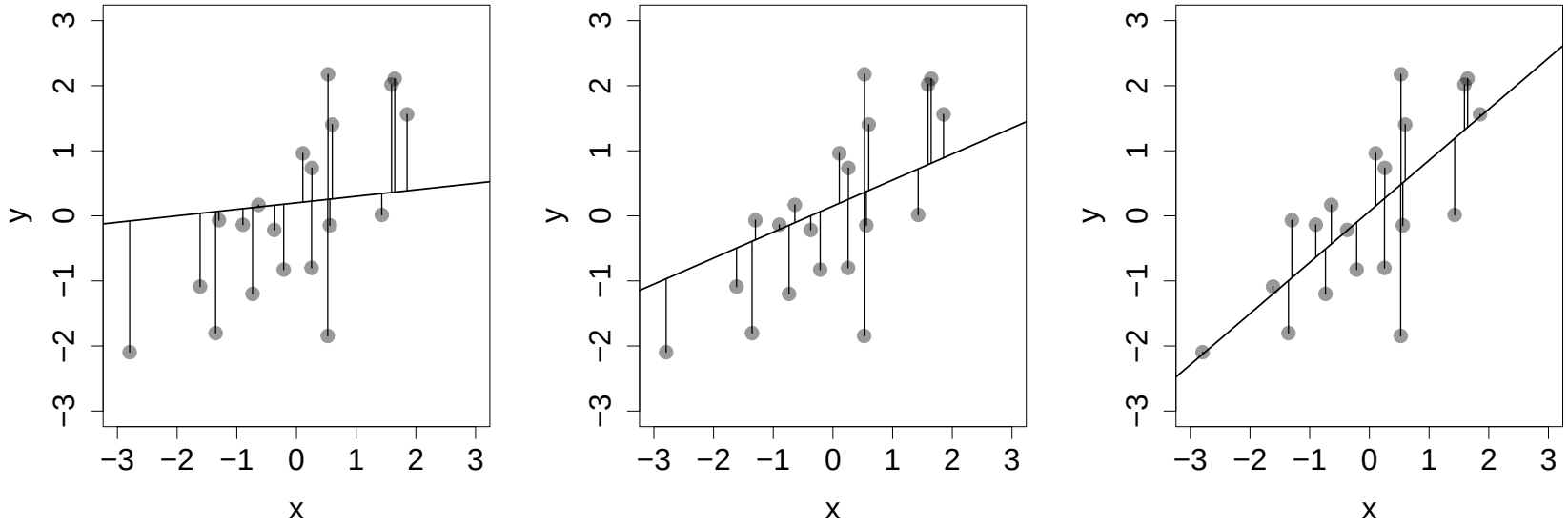
Ordinary Least Squares (intuition)



- Take some data
- Plot a line through the points
- For this line, the sum of the squared distances between Y_i and $\hat{Y}_i$:

$$\sum_{i=1}^{n} \left[Y_i - (\hat{\alpha} + \hat{\beta} X_i) \right]^2 = 16.95$$

Ordinary Least Squares (intuition)



$\rightarrow$ OLS selects the line that minimizes the sum of the squared distances between each point and the line

Ordinary Least Squares (formulae)

When we have only two variables, we can apply two straightforward formulae to recover the OLS estimates:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y.

Estimating OLS in R

Fortunately, R makes it easy to estimate the OLS model:

```
simple_ols_model <- lm(voters_change ~ students,
                       data = constituencies)
simple_ols_model
```

```
##
## Call:
## lm(formula = voters_change ~ students, data = constituencies)
##
## Coefficients:
## (Intercept)     students
##       205.1       -445.0
```

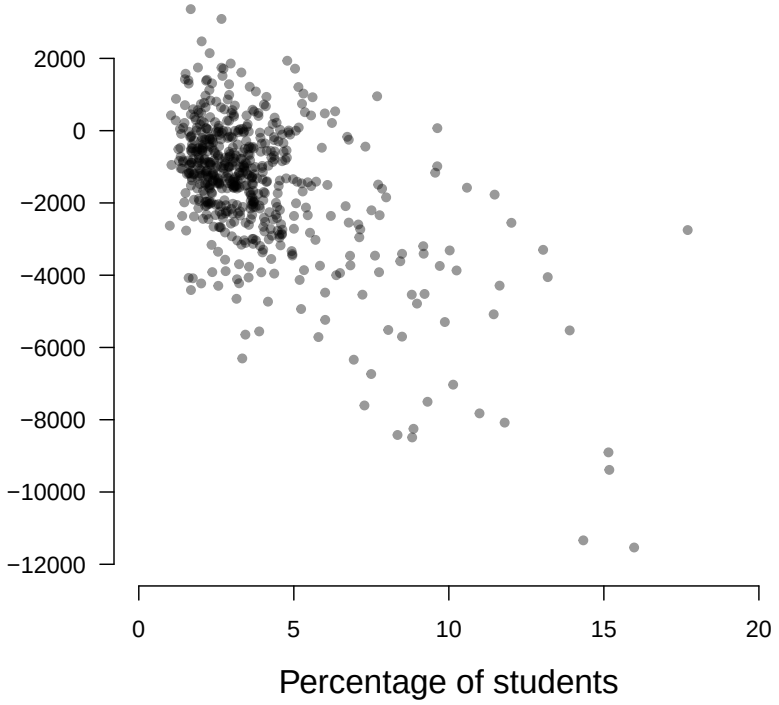
where `(Intercept)` $= \alpha$ and `students` $= \beta$

OLS estimates: vizualisation

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

$$\widehat{Voters}_i = \hat{\alpha} + \hat{\beta} \times Students_i$$

- $Voters$ is the change in registered voters
- $Students$ is the % of students

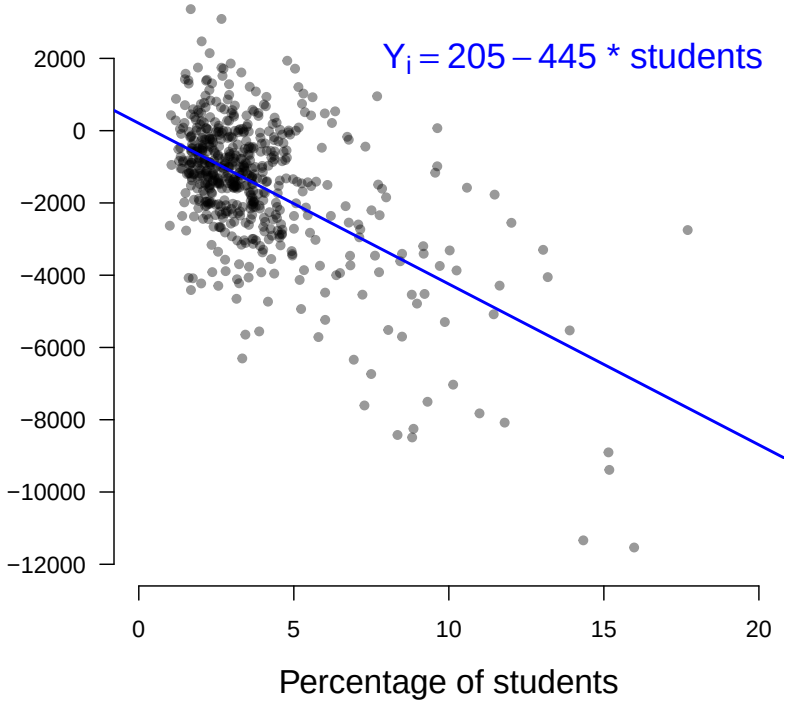


OLS estimates: vizualisation

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

$$\widehat{Voters}_i = 205 - 445 \times Students_i$$

- $Voters$ is the change in registered voters
- $Students$ is the % of students



OLS estimates: interpretation

What is the interpretation of $\widehat{\beta}$ = -445?

- Generic: A one-unit increase in X is associated with a $\widehat{\beta}$ 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 $\hat{\alpha} = 205.1$?

- Generic: $\hat{\alpha}$ is the average value of Y, when X is equal to 0
- Specific: For a hypothetical constituency with 0 students, the model predicts that the number of registered voters would increase by 205 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 calculate fitted values ($\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$) 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?

$$\hat{Y}_i = 205 - 445 * 10 = -4245$$

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

$$\hat{Y}_i = 205 - 445 * 20 = -8695$$

Fitted values in R

It is easy to calculate these fitted values in R:

```
predict(simple_ols_model, newdata = data.frame(students = 10))
```

```
##         1
## -4244.566
```

```
predict(simple_ols_model, newdata = data.frame(students = 20))
```

```
##         1
## -8694.281
```

- `predict` tells R that we would like to calculate fitted values
- the `newdata` argument is used to specify the values for which we would like to calculate fitted values

Regression and correlation

Last week we saw that the correlation coefficient is another way to summarise the relationship between two continuous variables.

What is the relationship between the correlation coefficient, ρ, and the regression coefficient β?

$$\hat{\beta} = \text{correlation of X and Y} \times \frac{\text{standard deviation of Y}}{\text{standard deviation of X}}$$

Implications:

- When the correlation is positive (negative), so is β
- If X increases by 1 standard deviation, Y increases by ρ standard deviations

Regression or correlation?

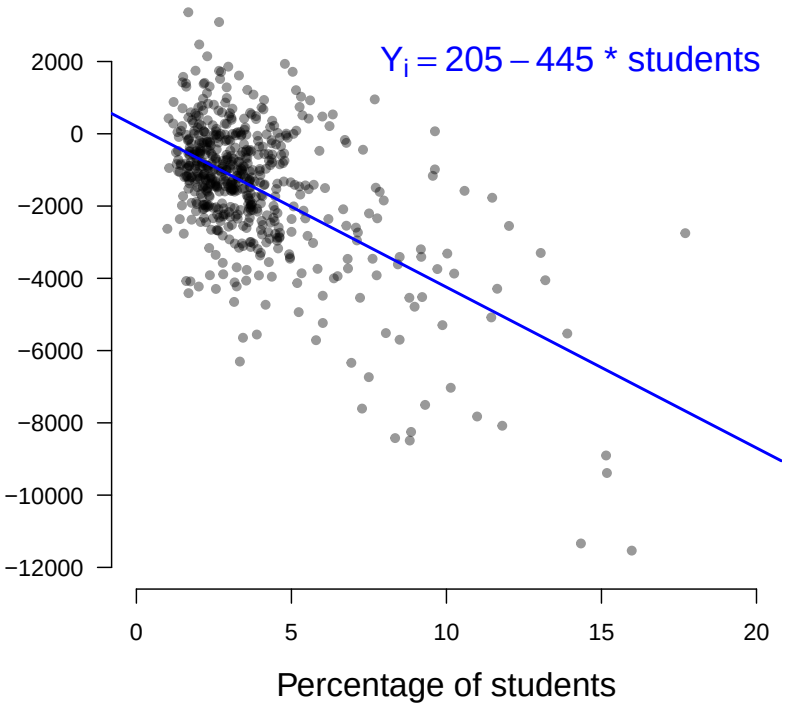
1. Regression is a better tool for making statements about the substantive magnitude of the relationship between variables
   - Correlation tells us whether X and Y are positively or negatively related, and something vague about the "strength" of the correlation
   - β tells you how many units Y changes when X increases by 1 unit
2. Regression is a more flexible approach
   - Not limited to associations between 2 variables – multiple variables can be included
   - Not limited to *linear* associations

We will spend much more time focussing on regression than correlation.

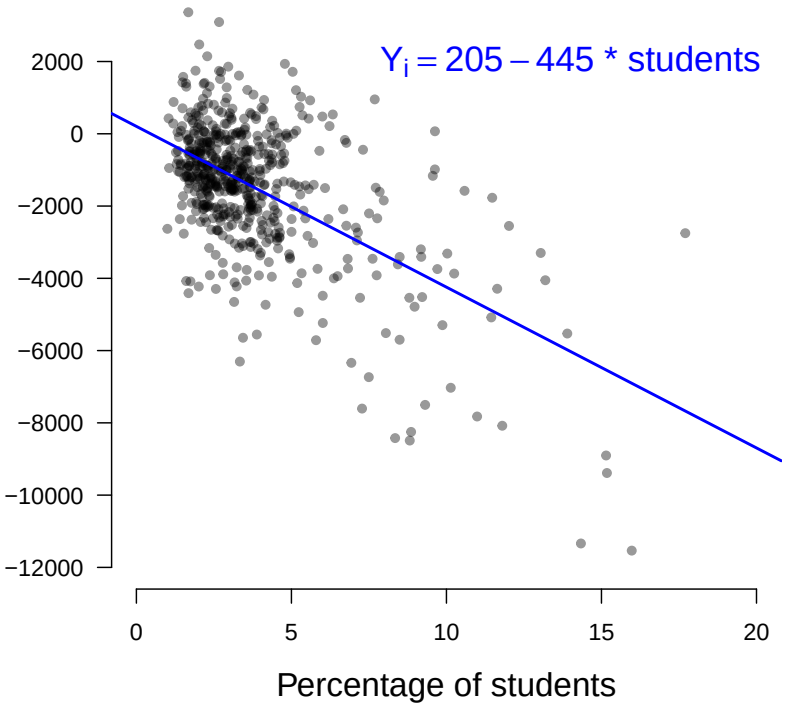
Measures of fit

How good is our model?

Is this a *perfect* model for our data?



How good is our model?

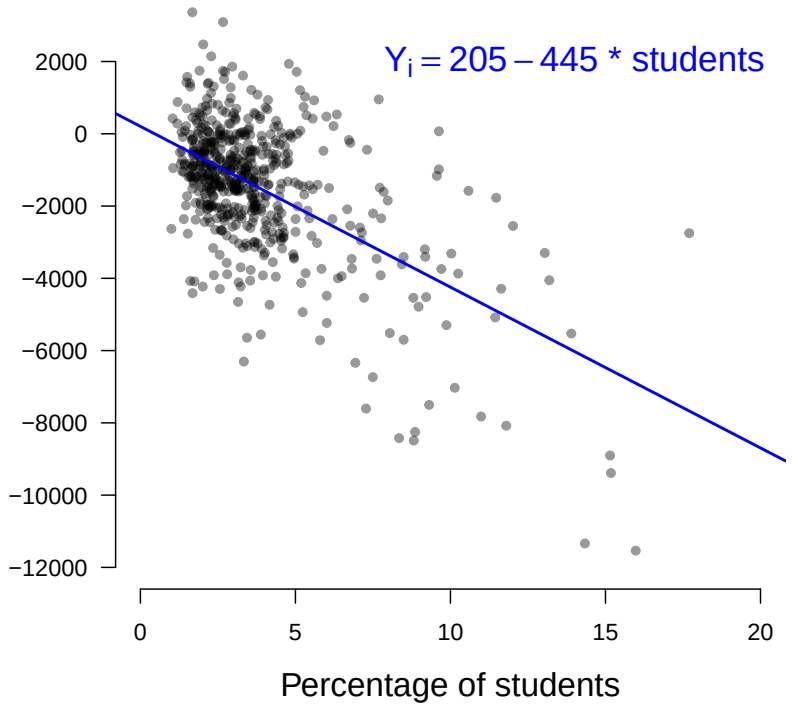


Is this a *perfect* model for our data?

- No! All models are bad, but some are useful.

Does a large student population *cause* decreased electoral registration?

How good is our model?



Is this a *perfect* model for our data?

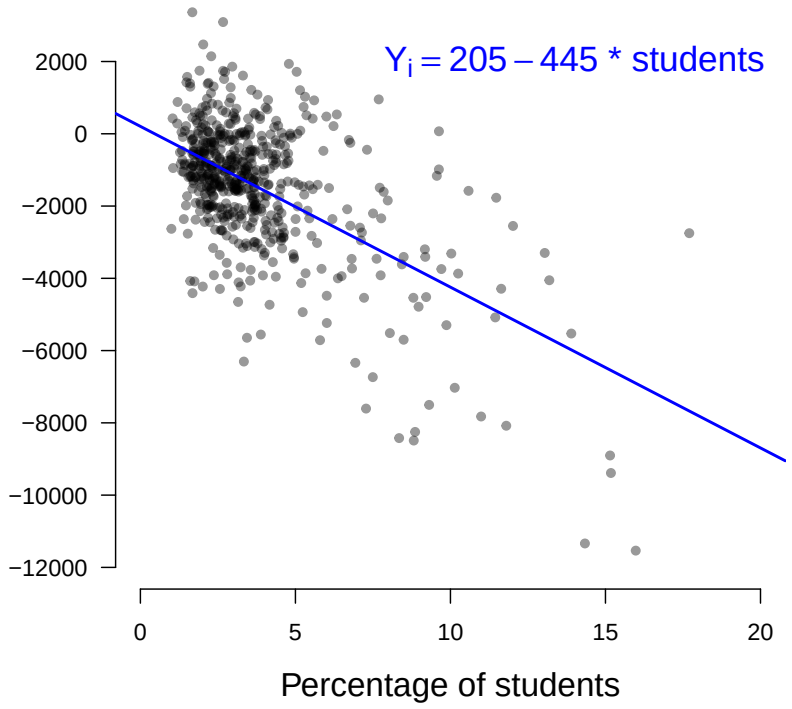
- No! All models are bad, but some are useful.

Does a large student population *cause* decreased electoral registration?

- No! Student-y areas may be different in many ways.

Is this a *good* model for our data?

How good is our model?



Is this a *perfect* model for our data?

- No! All models are bad, but some are useful.

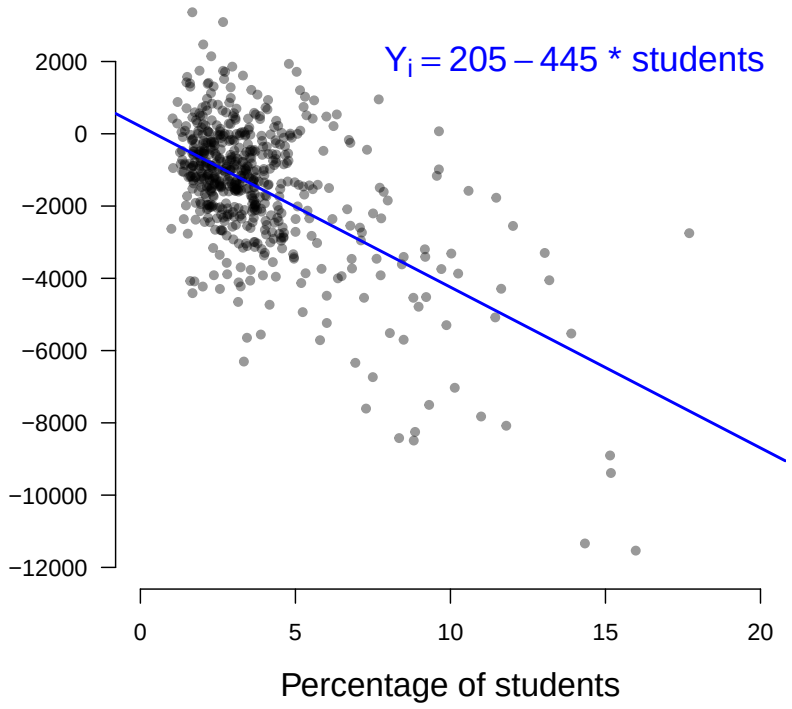
Does a large student population *cause* decreased electoral registration?

- No! Student-y areas may be different in many ways.

Is this a *good* model for our data?

- It depends! What do you want your model to do?

How good is our model?



Is this a *perfect* model for our data?

- No! All models are bad, but some are useful.

Does a large student population *cause* decreased electoral registration?

- No! Student-y areas may be different in many ways.

Is this a *good* model for our data?

- It depends! What do you want your model to do?

Measures of model fit help us to assess the degree to which our model approximates the real variation in our data.

R-squared

R^2 – The Coefficient of Determination

R^2 measures the proportion of the variation in Y_i that is explained by X_i. It varies between between 0 and 1 and can be used to describe how much of the variation in our dependent variable is "explained" by our independent variable.

- If X explains all the variation in Y, then $R^2 = 1$
- If X explains none of the variation in Y, then $R^2 = 0$
- You do not need to know how to *calculate* R^2, but you do need to know how to *interpret* it!

R-squared

R^2 starts from the identity

$$Y_i = \hat{Y}_i + \hat{\epsilon}_i$$

where

- Y_i is the observed value of Y for observation i
- $\hat{Y}_i$ is the fitted value of Y for observation i
- $\hat{\epsilon}_i$ is the residual for observation i ($\hat{\epsilon}_i \equiv Y_i - \hat{Y}_i$)

R-squared

Imagine that we were to use a really dumb "model" to predict Y for each value in our data:

$$\hat{Y}_{i(\text{dumb})} = \bar{Y}$$

We could assess the accuracy of these "predictions" by calculating the distance between the predicted values and the observed values:

$$\text{TSS (Total Sum of Squares)} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(\text{dumb})})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

The TSS is therefore the sum of the squared distances between each observation and the mean.

R-squared

We can then compare the predictions from this dumb model, to the predictions (fitted values) from our regression model:

$$\hat{Y}_{i(\text{ols})} = \hat{\alpha} + \hat{\beta} X_i$$

Again, let's calculate the accuracy by summing the distances between the predicted and observed values (i.e. the residuals):

$$\text{SSR (Sum of Squared Residuals)} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(\text{ols})})^2$$

If our regression model is doing a good job, we should make fewer or smaller prediction errors than when using the dumb model.

R-squared

The R^2 is a statistic that summarises how much better the predictions from our regression model are relative to a baseline model where we just use the mean value of Y as a prediction for all observations (i.e. the dumb model)

Definition:
The R^2 is defined as

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

where

- TSS (Total sum of squares) equals $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- SSR (Sum squared residuals) equals $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

R-squared

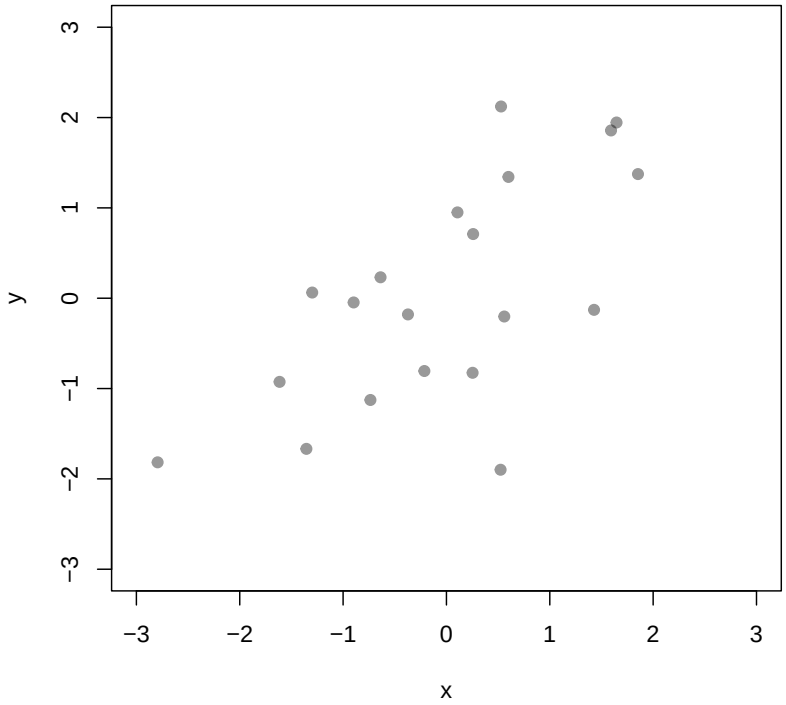
The R^2 is a statistic that summarises how much better the predictions from our regression model are relative to a baseline model where we just use the mean value of Y as a prediction for all observations (i.e. the dumb model)

Intuition:

- R^2 varies between 0 and 1
- When the residuals (prediction errors) from our model are large (SSR is large), R^2 is closer to 0
- When the residuals (prediction errors) from our model are small (SSR is small), R^2 is closer to 1

R-squared

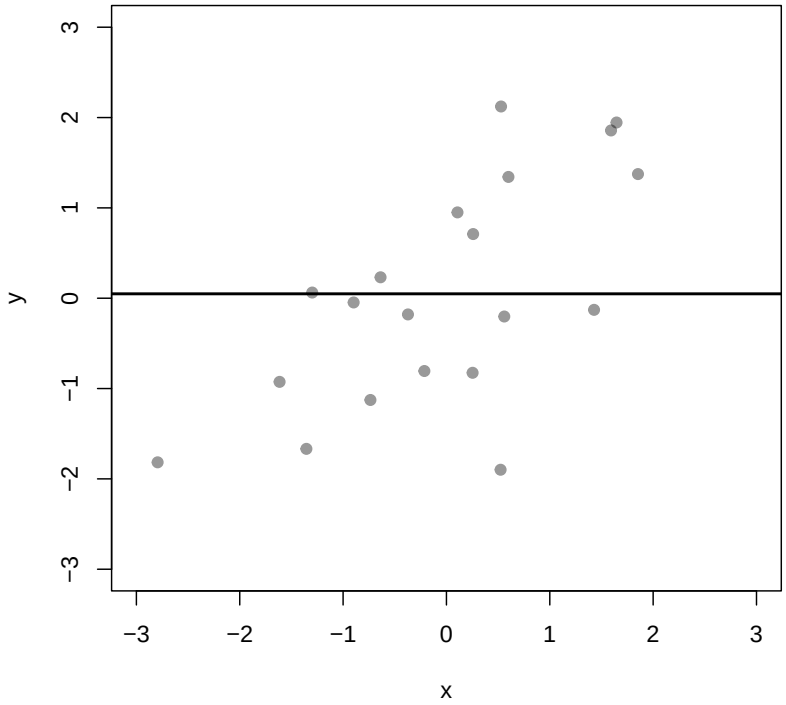
$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$



What is the total squared prediction error using the mean?

R-squared

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

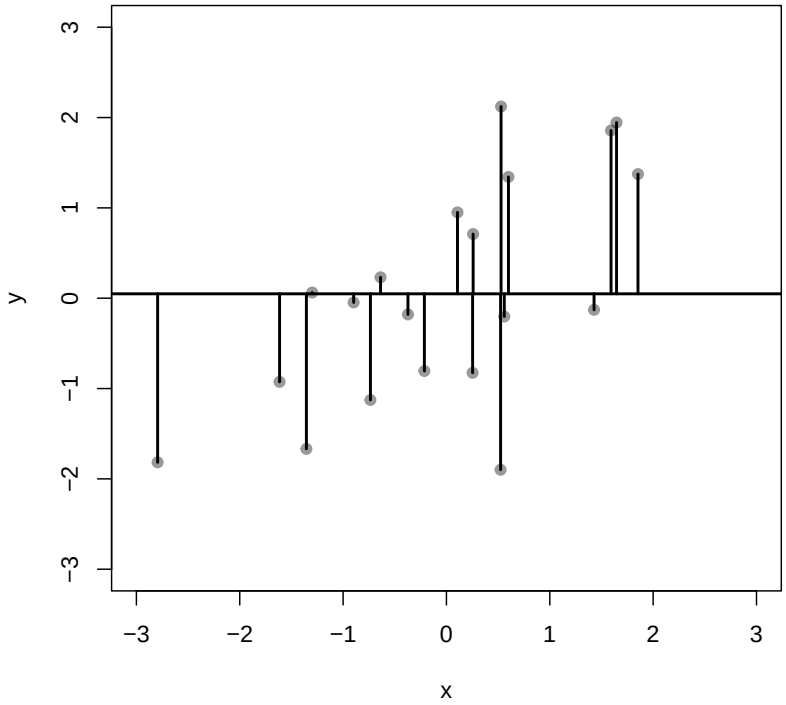


What is the total squared prediction error using the mean?

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

R-squared

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

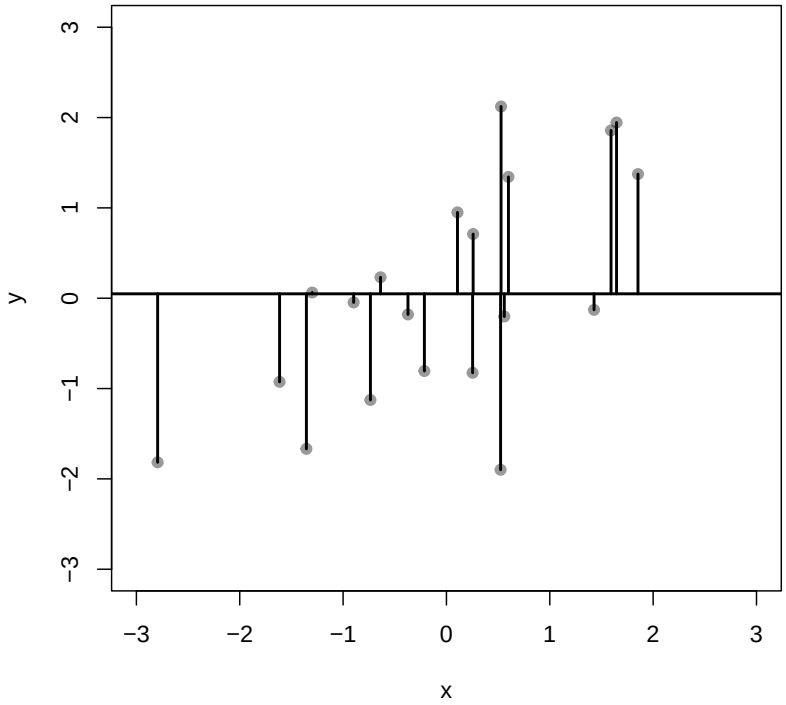


What is the total squared prediction error using the mean?

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

R-squared

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

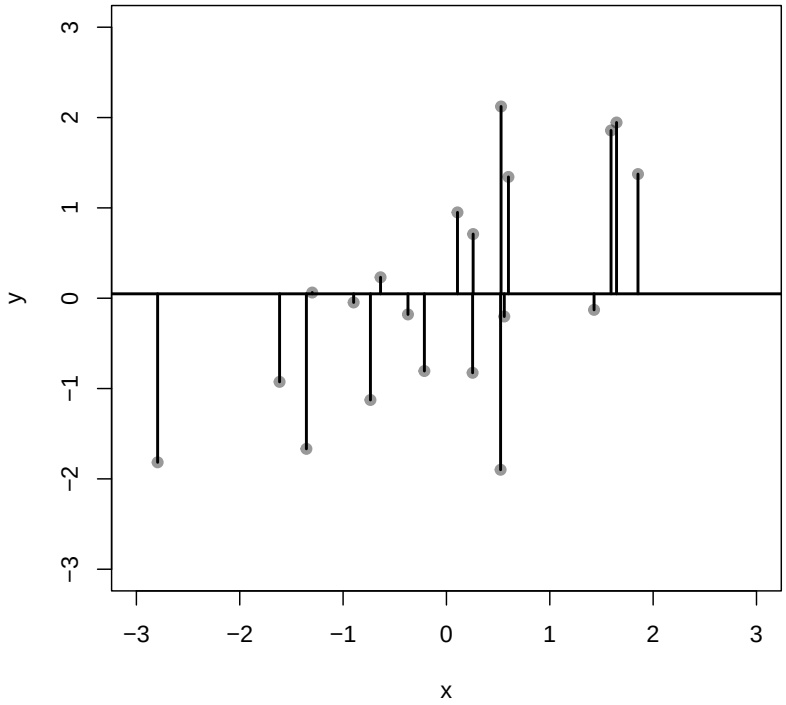


What is the total squared prediction error using the mean?

- TSS $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- TSS $= 30$

R-squared

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



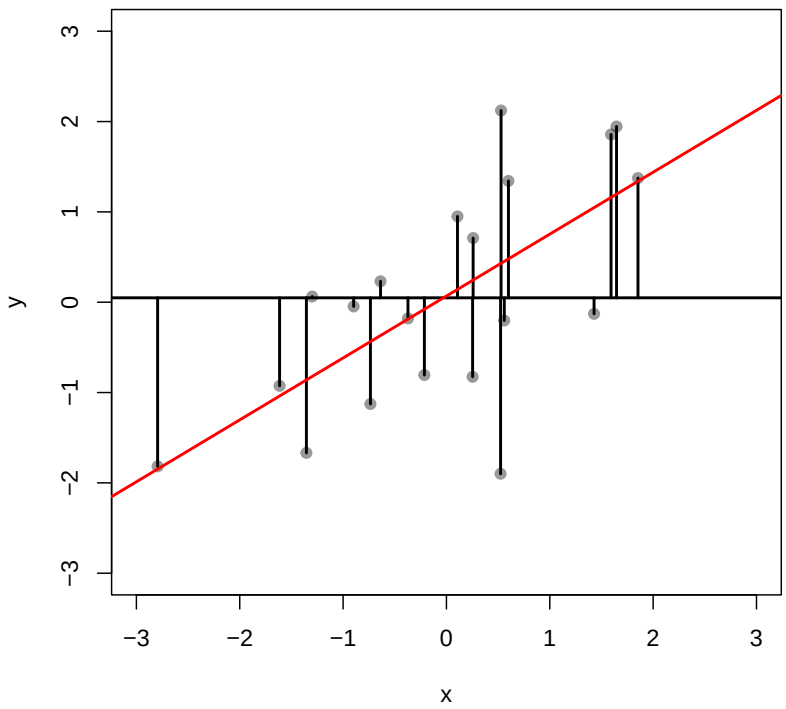
What is the total squared prediction error using the mean?

- TSS $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- TSS $= 30$

What is the total squared prediction error using the regression line?

R-squared

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



What is the total squared prediction error using the mean?

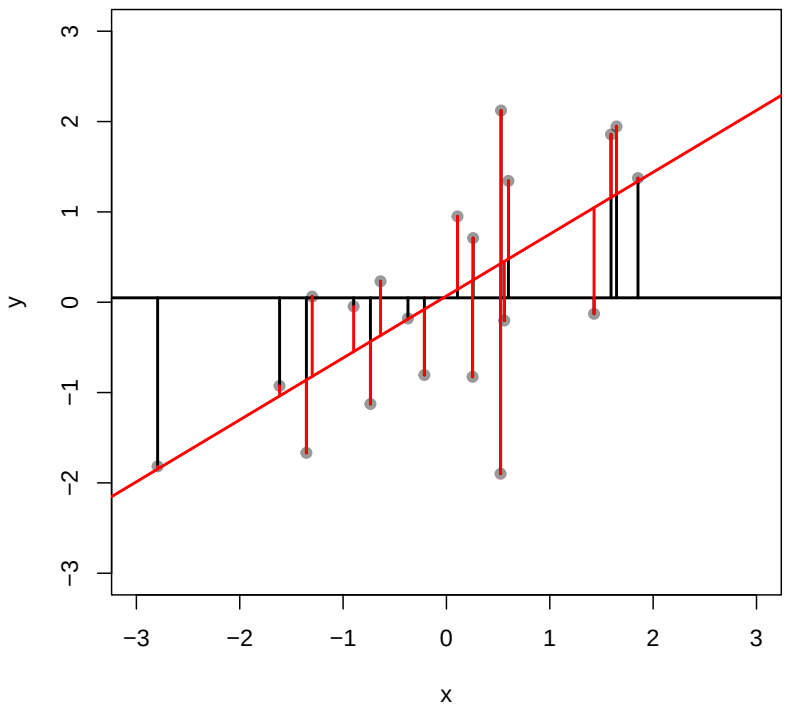
- TSS $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- TSS $= 30$

What is the total squared prediction error using the regression line?

- SSR $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

R-squared

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



What is the total squared prediction error using the mean?

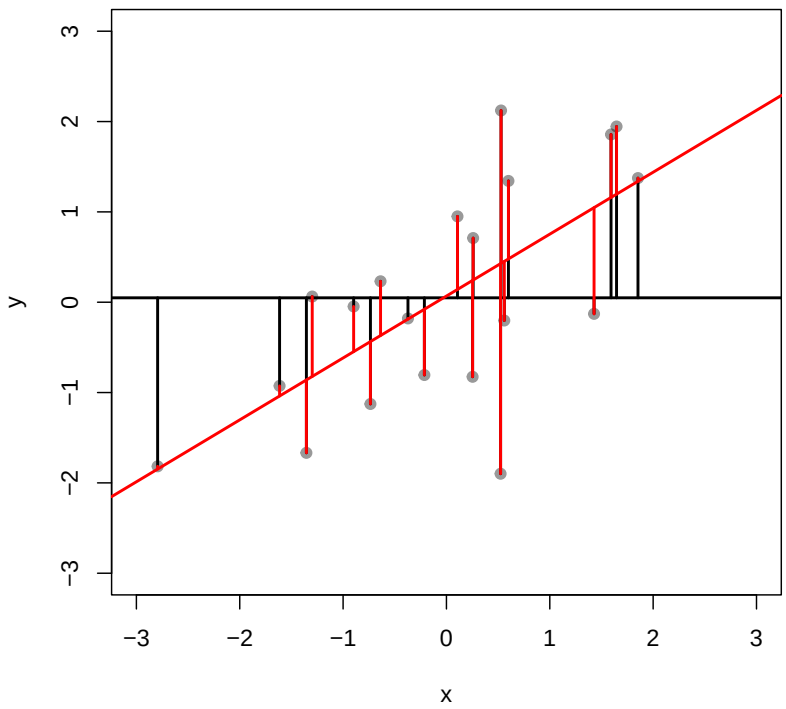
- TSS $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- TSS $= 30$

What is the total squared prediction error using the regression line?

- SSR $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

R-squared

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



What is the total squared prediction error using the mean?

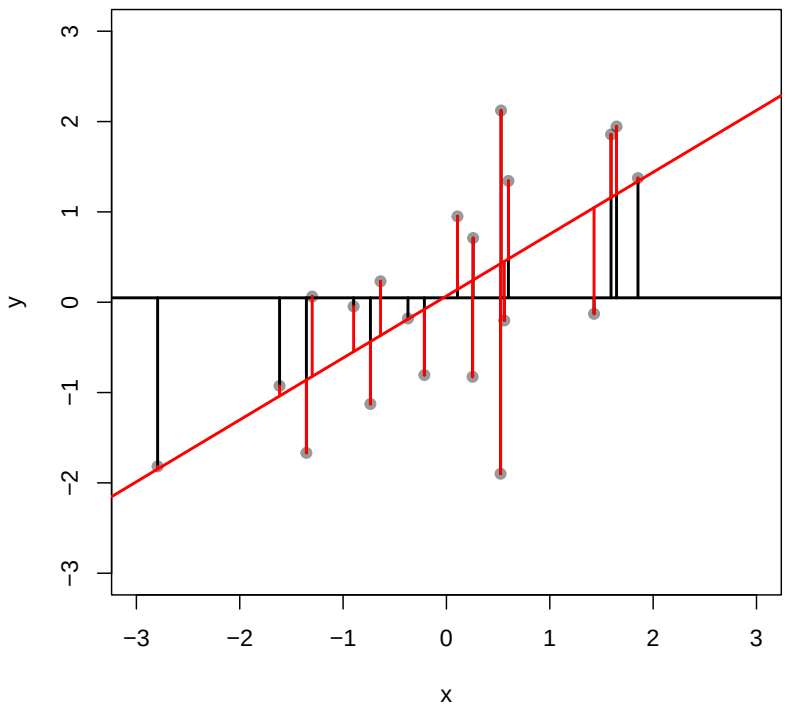
- TSS $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- TSS $= 30$

What is the total squared prediction error using the regression line?

- SSR $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- SSR $= 17$

R-squared

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



What is the total squared prediction error using the mean?

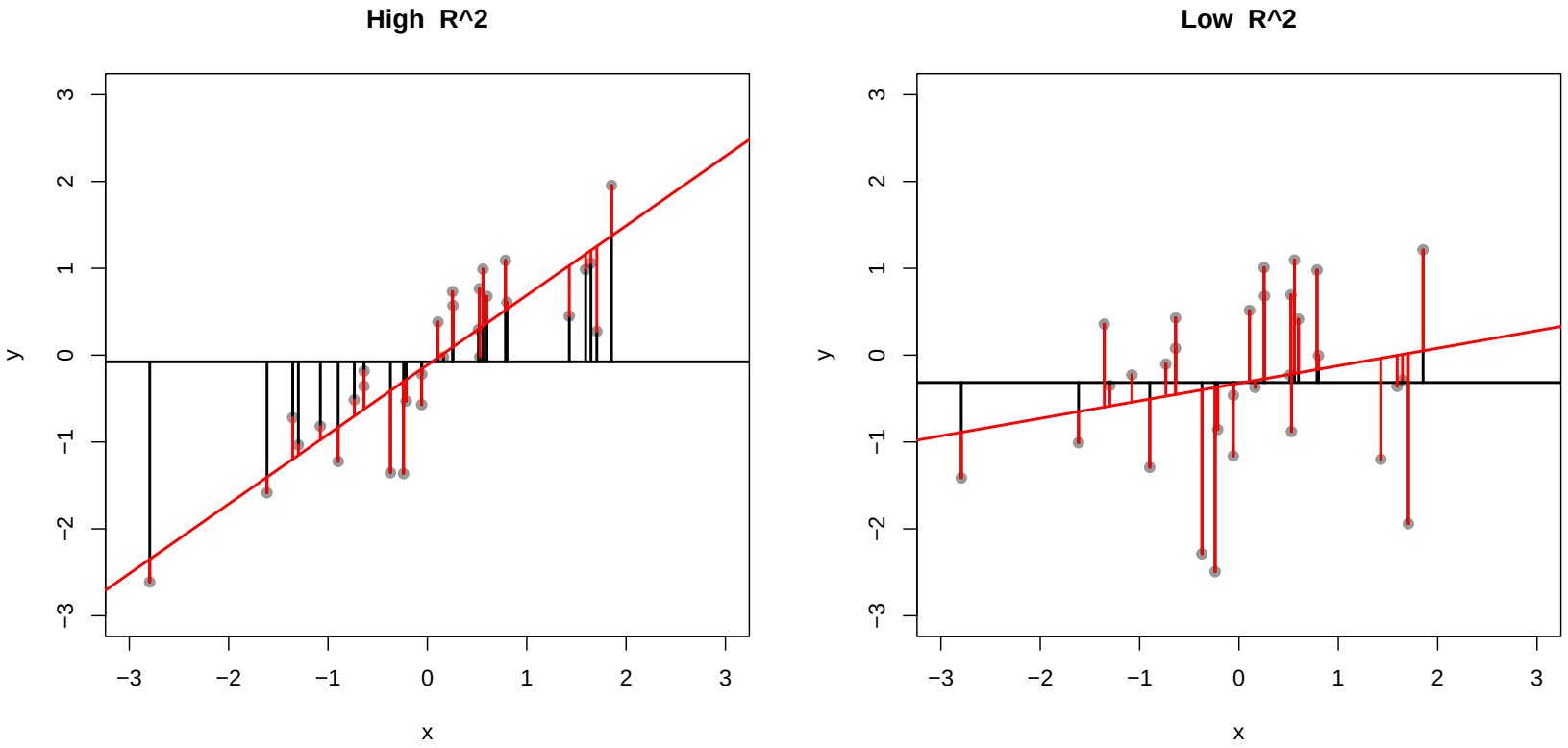
- TSS $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- TSS $= 30$

What is the total squared prediction error using the regression line?

- SSR $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- SSR $= 17$

$$R^2 = \frac{30 - 17}{30} = 0.44$$

R-squared



How useful is R^2?

What does R^2 tell us?

- Large values $\rightarrow$ independent variable is good at predicting Y
- Small values $\rightarrow$ independent variable is poor at predicting Y

How useful is R^2?

What does R^2 tell us?

- Large values $\rightarrow$ independent variable is good at predicting Y
- Small values $\rightarrow$ independent variable is poor at predicting Y

What does R^2 not tell us?

- Large R^2 does not imply a causal relationship
- Low R^2 does not necessarily imply a useless regression

R-squared: example

```r
## We can find out more detail about our estimated model using "summary"
summary(simple_ols_model)
```

```
...
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   205.15     119.46   1.717   0.0865 .
## 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
...
```

```r
summary(simple_ols_model)$r.squared
```

```
## [1] 0.3225678
```

The % of students in a constituency explains 32% of the variation in the change in the number of registered voters.

Regression and the difference in means

Difference in means recap

When we spoke about causality, our main quantity of interest was the **average treatment effect**.

We **estimated** the ATE using the **difference-in-means** between two groups:

$$\text{Difference-in-means} = \bar{Y}_{X=1} - \bar{Y}_{X=0}$$

Difference in means in R

Let's imagine that we want to know how the numbers on the electoral register changed between urban and rural areas.

We can calculate this in R:

```
urban_change <- mean(constituencies$voters_change[constituencies$urban == 1])
urban_change
```

```
## [1] -2013.686
```

```
rural_change <- mean(constituencies$voters_change[constituencies$urban == 0])
rural_change
```

```
## [1] -964.8212
```

```
urban_change - rural_change
```

```
## [1] -1048.865
```

This suggests that, on average, urban constituencies saw greater decreases in registration than rural constituencies.

Linear regression with binary X variable

We motivated linear regression as a way of quantifying the relationship between two continuous variables, X and Y

Linear regression is in fact far more flexible

- Y should always be (approximately) continuous
- X can have essentially *any* level of measurement

When X is a binary or dummy variable, the estimated $\hat{\beta}$ will be equivalent to the difference-in-means estimate

Binary or "Dummy" Variables

Dummy variables are binary indicators that $= 1$ if an observation has a specific trait and $= 0$ otherwise.

Example: X_{male}, X_{labour}, X_{urban}

Linear regression with binary X variable

Consider a linear regression model with a binary X variable:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

What is the interpretation of α in this model?

Linear regression with binary X variable

Consider a linear regression model with a binary X variable:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

What is the interpretation of α in this model?

- α is the average value of Y when $X = 0$
- α is the average change in voter registration for *rural* constituencies

What is the interpretation of β in this model?

Linear regression with binary X variable

Consider a linear regression model with a binary X variable:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

What is the interpretation of α in this model?

- α is the average value of Y when $X = 0$
- α is the average change in voter registration for *rural* constituencies

What is the interpretation of β in this model?

- β is the average change in Y when X increases by one-unit
- What is a one-unit change in "urban"? Going from rural to urban!
- β is the average change in voter registration between urban and rural constituencies

Linear regression with binary X variable

Consider a linear regression model with a binary X variable:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

What is the interpretation of α in this model?

- α is the average value of Y when $X = 0$
- α is the average change in voter registration for *rural* constituencies

What is the interpretation of β in this model?

- β is the average change in Y when X increases by one-unit
- What is a one-unit change in "urban"? Going from rural to urban!
- β is the average change in voter registration between urban and rural constituencies

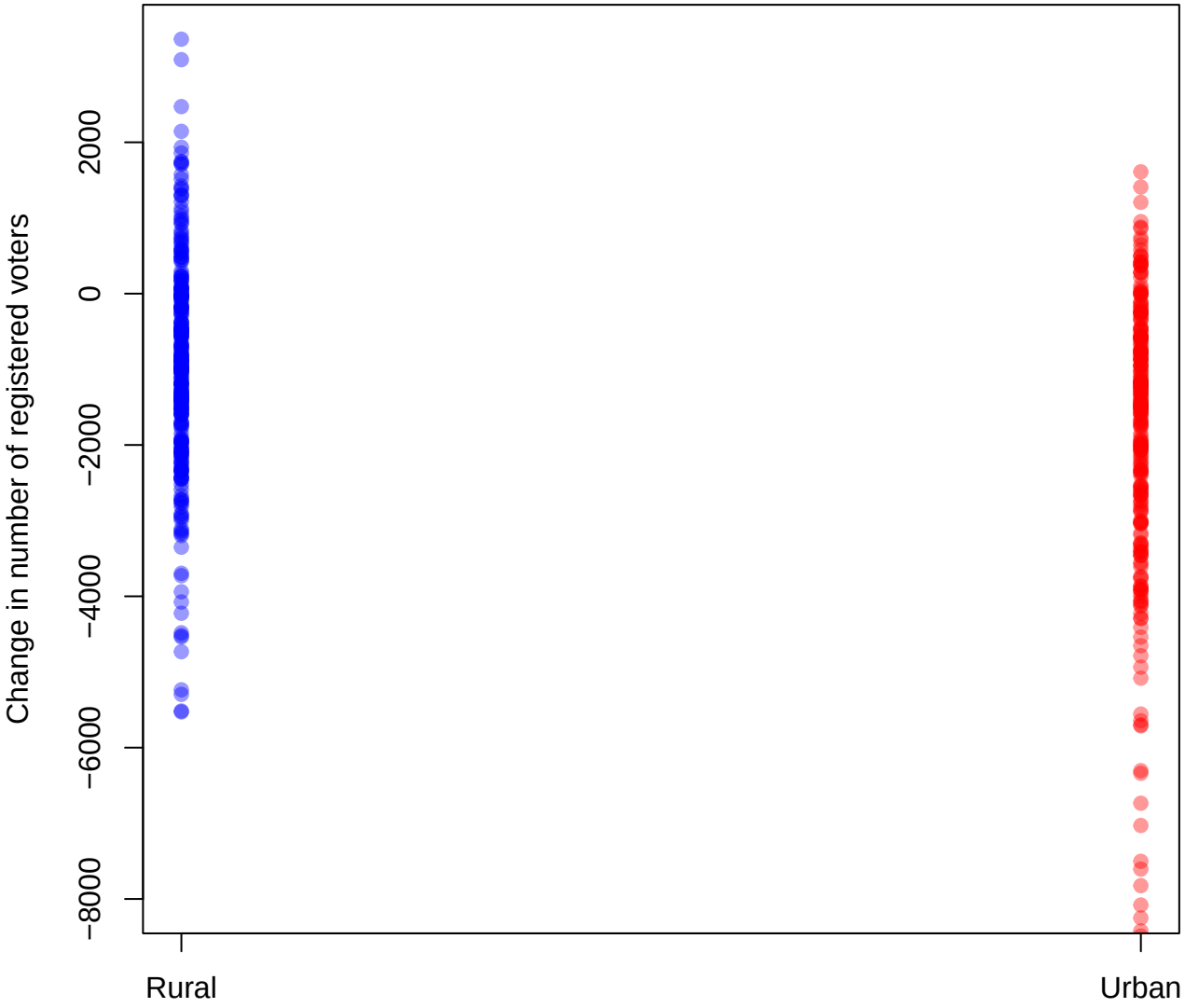
$\rightarrow$ β is the same thing as the difference in means!

Linear regression with binary X variable

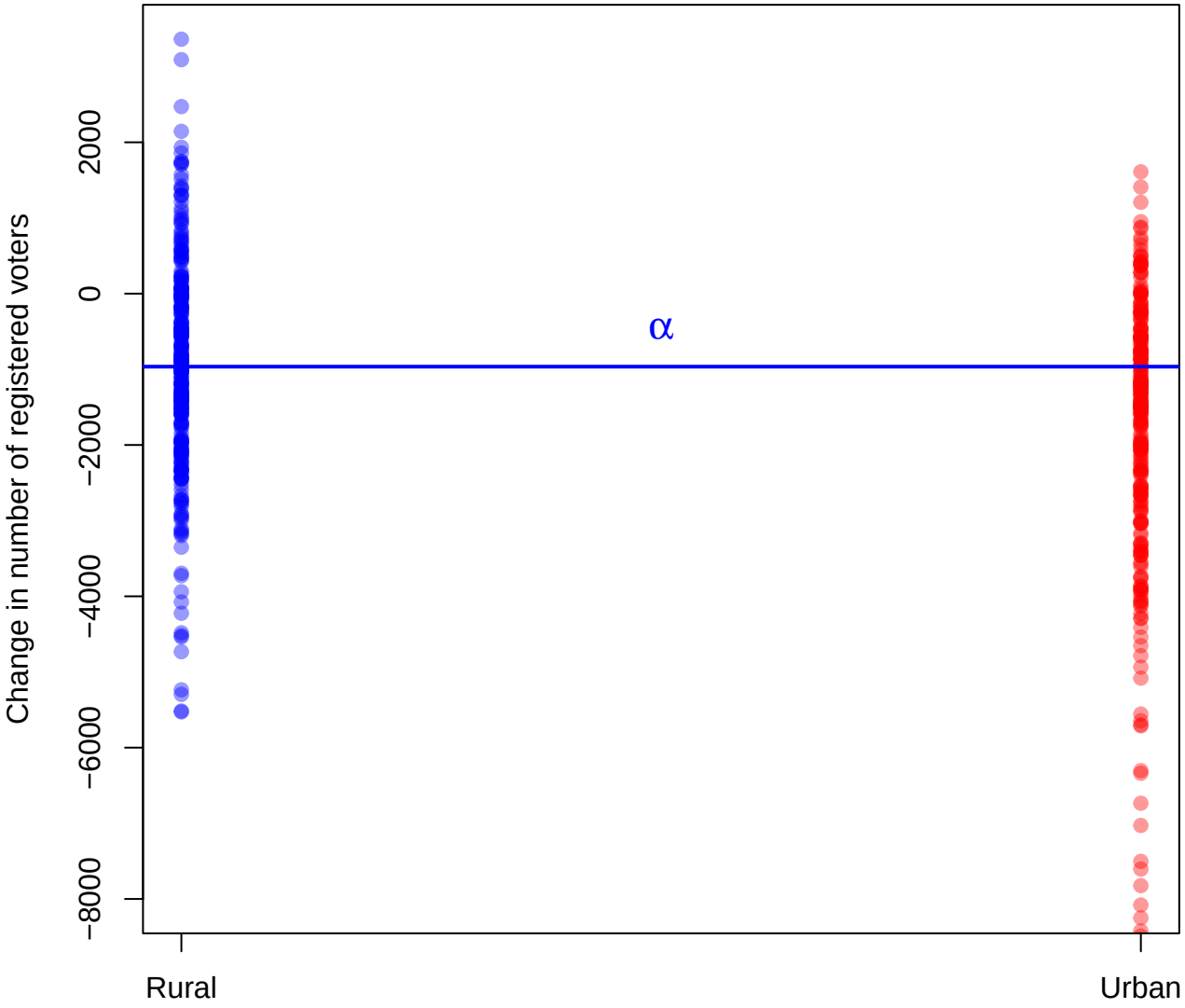
In general, when we have a linear model with a binary X:

- α is the average value of Y when X is equal to zero
- β is the average difference in Y for observations where $X = 0$ and $X = 1$
- A "one-unit" change in X means moving from one group to another

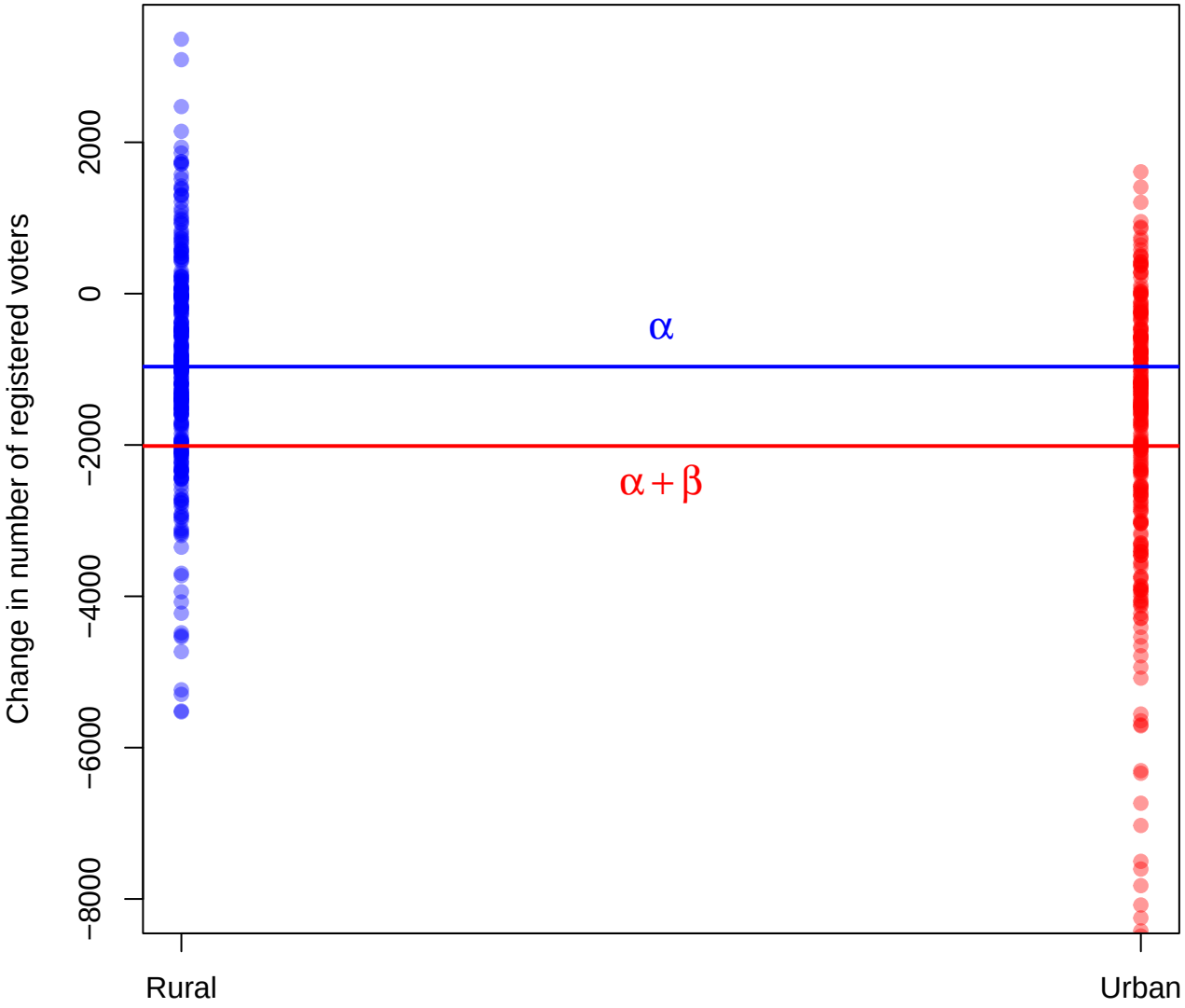
Linear regression with binary X variable



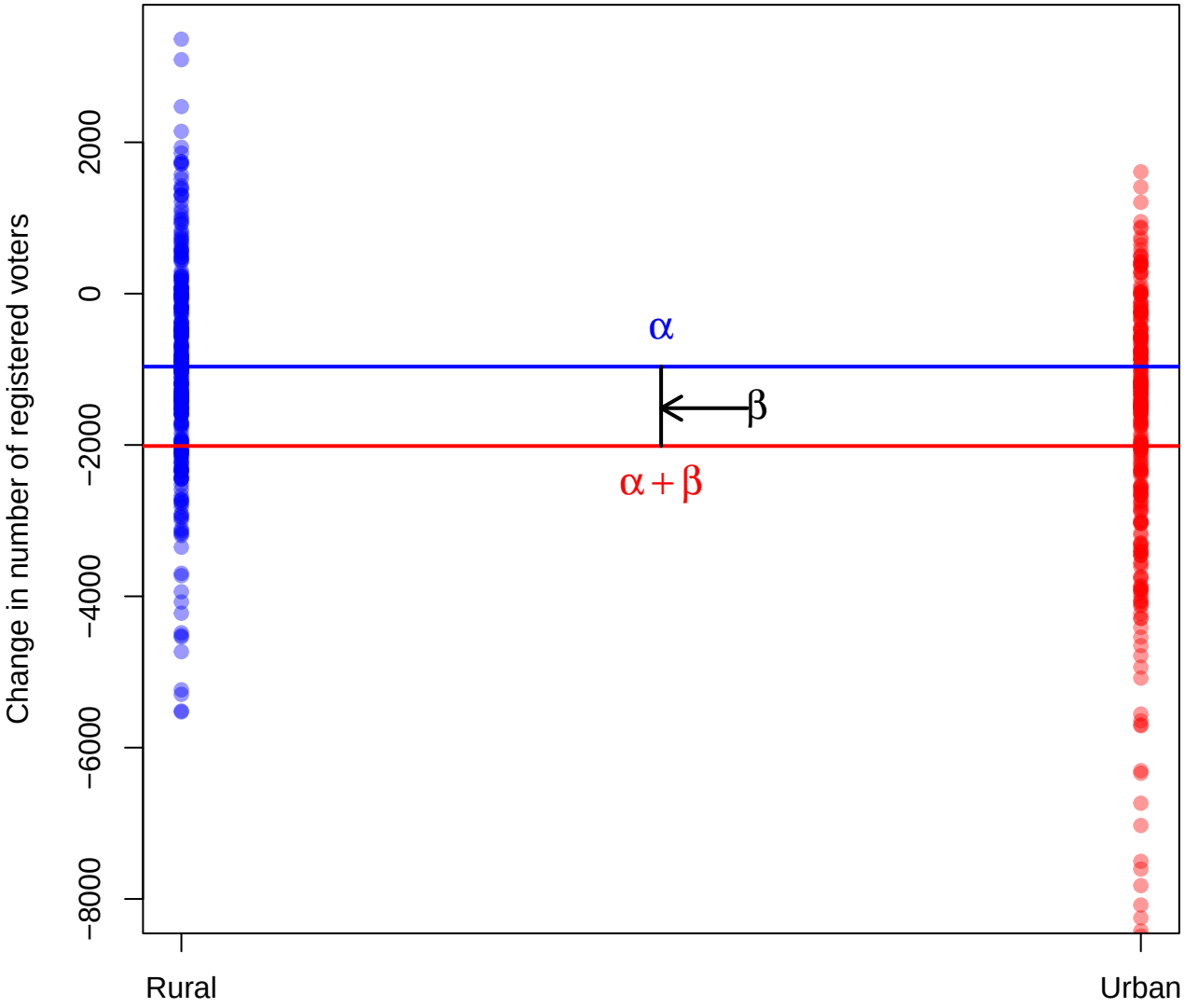
Linear regression with binary X variable



Linear regression with binary X variable



Linear regression with binary X variable



Linear regression with binary X variable

```r
urban_change
```

```
## [1] -2013.686
```

```r
rural_change
```

```
## [1] -964.8212
```

```r
urban_change - rural_change
```

```
## [1] -1048.865
```

```r
urban_ols <- lm(voters_change ~ urban,
               data = constituencies)
urban_ols
```

```
...
## Coefficients:
## (Intercept)        urban
##      -964.8      -1048.9
...
```

$\hat{\alpha}$ is the same as `rural_change`

- Registration decreased by 965 on average in rural areas

$\hat{\beta}$ is the same as `urban_change` - `rural_change`

- Registration decreased by 1049 more, on average, in urban rather than rural areas

Conclusion

What have we covered?

- Models are abstractions that allow us to characterise structure and describe general patterns
- Regression modelling is a tool for describing the relationships between variables
- Regression is useful, because we can use the estimates to describe the *substantive magnitude* of these relationships
- Regression is very flexible, and we are able to model our outcome as a function of different types of explanatory variables

Seminar

In seminars this week, you will learn about ...

1. ... fitting regressions using the `lm()` function.
2. ... calculating fitted values `predict()` function.
3. ... interpreting regression coefficients.
4. ... how to export and save plots from R.