Ordinary Least Squares - I

Lecture 7

Louis SIRUGUE

11/2021

In Part I we saw

- Different classes of R objects

```
class("numeric")
```

- Vectors

```
match(8, c(6, 1, 9, 5, 8, 4))
```

- If/else statements and loops

```
if (1 != 1) {print("a")}
```

- Functions and packages

```
library(tidyverse)
```

In Part I we saw

The pipe operator

```
tibble(letter = c("A", "B", "C"),
       figure = c(1, 2, 3)) %>%
    mutate(var = paste(letter, figure, sep = "-"))
```

```
## # A tibble: 3 x 3
##   letter figure var
##   <chr>   <dbl> <chr>
## 1 A           1 A-1
## 2 B           2 B-2
## 3 C           3 C-3
```

Chaining operations

```
you_can_use %>% View() %>%
  or() %>% head() %>%
  or() %>% whatEverYouWant() %>%
  but() %>% CheckRegularlyWhatYouDo()
```

In Part I we saw

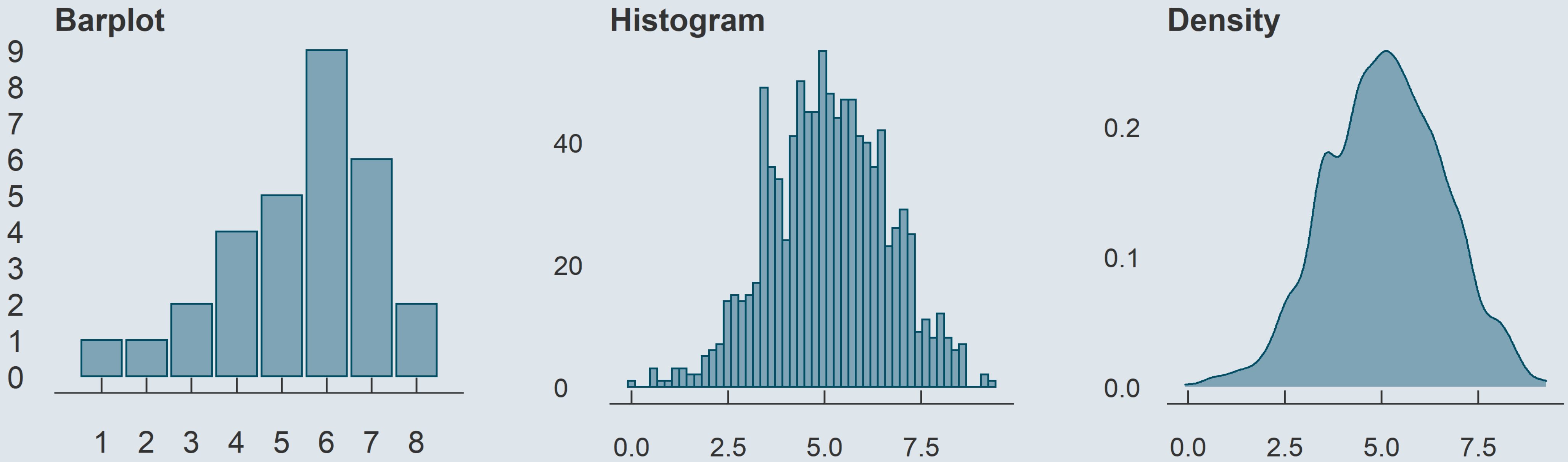
Important functions of the dplyr grammar

Function	Meaning
mutate()	Modify or create a variable
select()	Keep a subset of variables
filter()	Keep a subset of observations
arrange()	Sort the data
group_by()	Group the data
summarise()	Summarizes variables into 1 observation per group
bind_rows()	Append data
left/right/inner/full_join()	Merge data
pivot_longer/wider()	Reshape data

In Part I we saw

Distributions

- The **distribution** of a variable documents all its possible values and how frequent they are

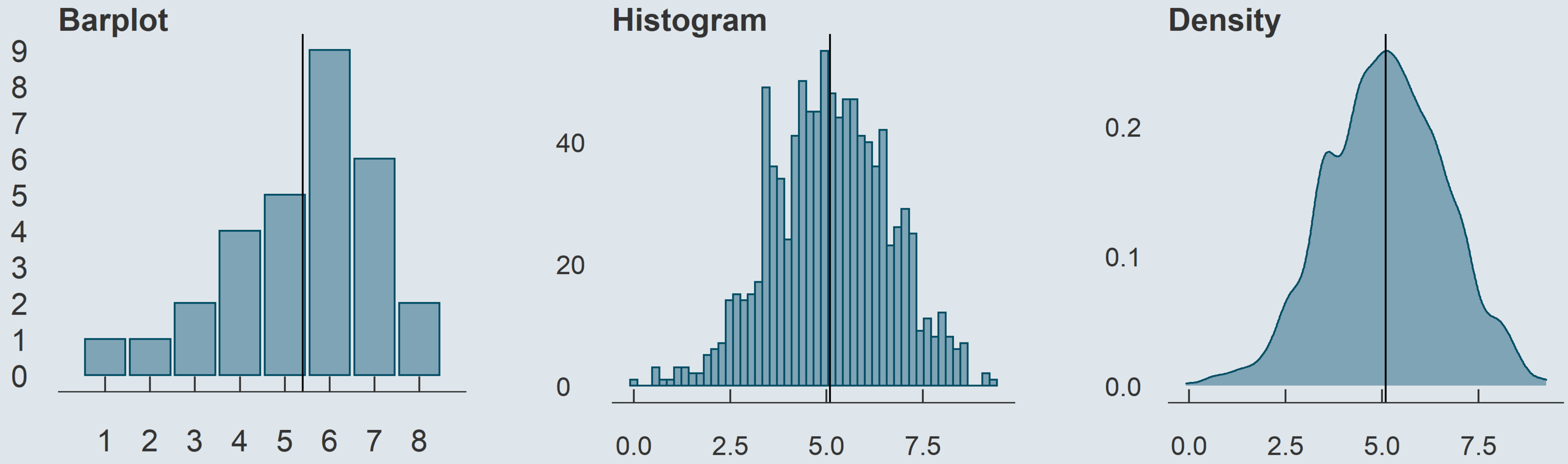


- We can describe a distribution with:

In Part I we saw

Distributions

- The **distribution** of a variable documents all its possible values and how frequent they are

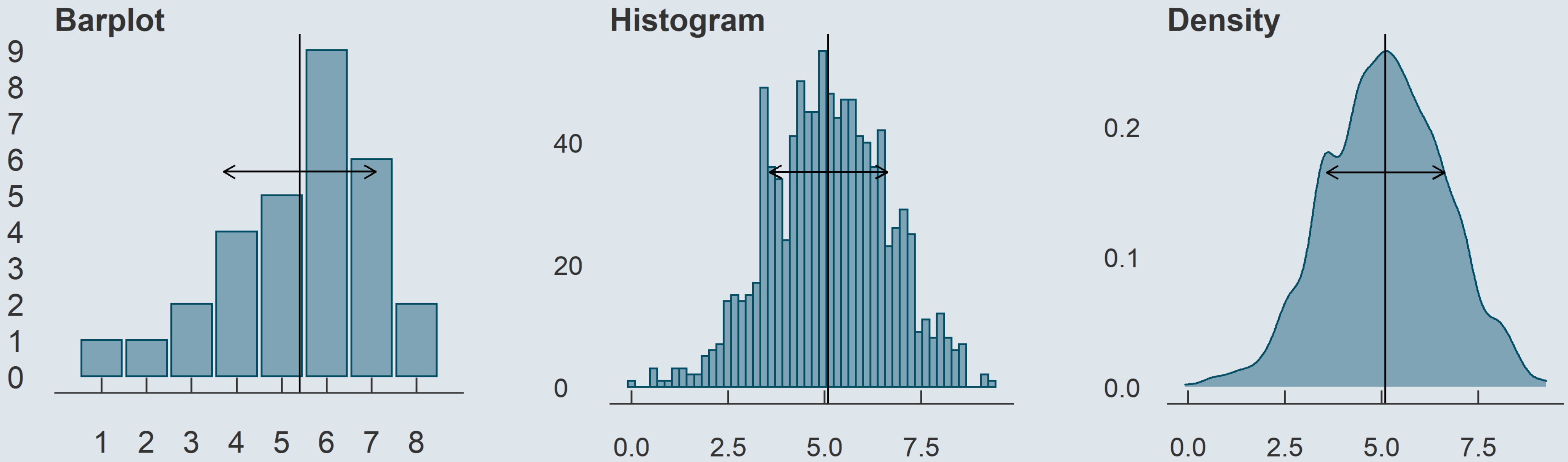


- We can describe a distribution with:
  - Its **central tendency**

In Part I we saw

Distributions

- The **distribution** of a variable documents all its possible values and how frequent they are



- We can describe a distribution with:
  - Its **central tendency**
  - And its **spread**

In Part I we saw

Central tendency

- The **mean** is the sum of all values divided by the number of observations

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The **median** is the value that divides the (sorted) distribution into two groups of equal size

$$\mathrm{Med}(x) = \begin{cases} x[\frac{N+1}{2}] & \text{if } N \text{ is odd} \\ \frac{x[\frac{N}{2}]+x[\frac{N}{2}+1]}{2} & \text{if } N \text{ is even} \end{cases}$$

Spread

- The **standard deviation** is square root of the average squared deviation from the mean

$$\mathrm{SD}(x) = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

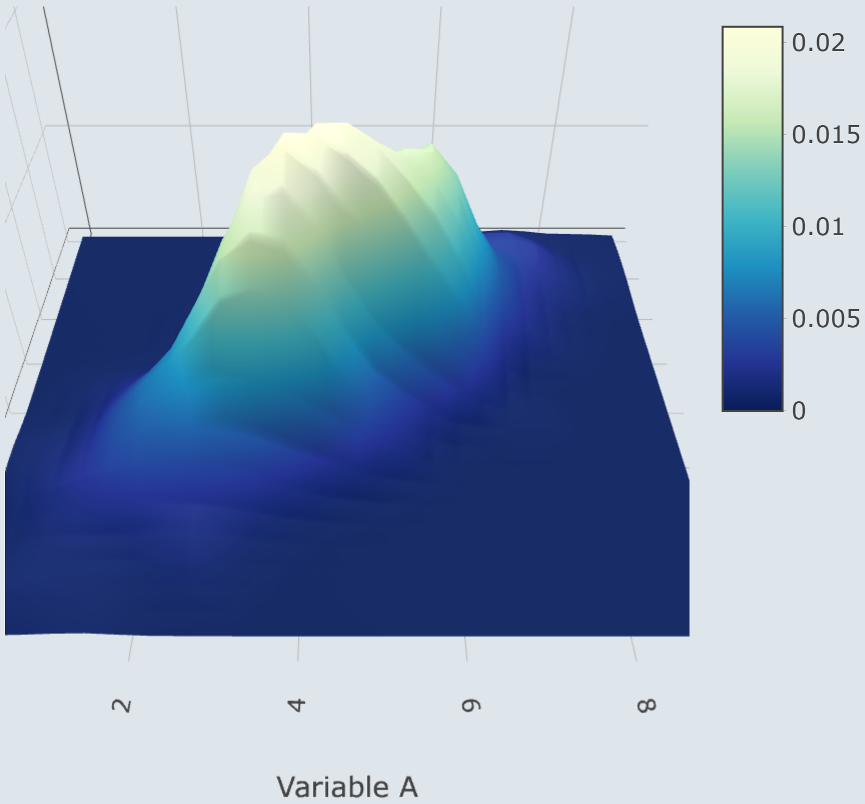
- The **interquartile range** is the difference between the maximum and the minimum value from the middle half of the distribution

$$\mathrm{IQR} = Q_3 - Q_1$$

In Part I we saw

Joint distribution

- The **joint distribution** shows the possible values and associated frequencies for two variable simultaneously



In Part I we saw

Joint distribution

→ ***When describing a joint distribution, we're interested in the relationship between the two variables***

- The **covariance** quantifies the joint deviation of two variables from their respective mean

$$\text{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- The **correlation** is the covariance of two variables divided by the product of their standard deviation

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\text{SD}(x) \times \text{SD}(y)}$$

In Part I we saw

Graphs with `ggplot()`

The 3 core components of the ggplot() function

Component	Contribution	Implementation
Data	Underlying values	ggplot(data, | data %>% ggplot(.,
Mapping	Axis assignment	aes(x = V1, y = V2, ...))
Geometry	Type of plot	+ geom_point() + geom_line() + ...

- Any other element should be added with a `+` sign

```
ggplot(data, aes(x = V1, y = V2)) +
  geom_point() + geom_line() +
  anything_else()
```

In Part I we saw

Main types of aesthetics

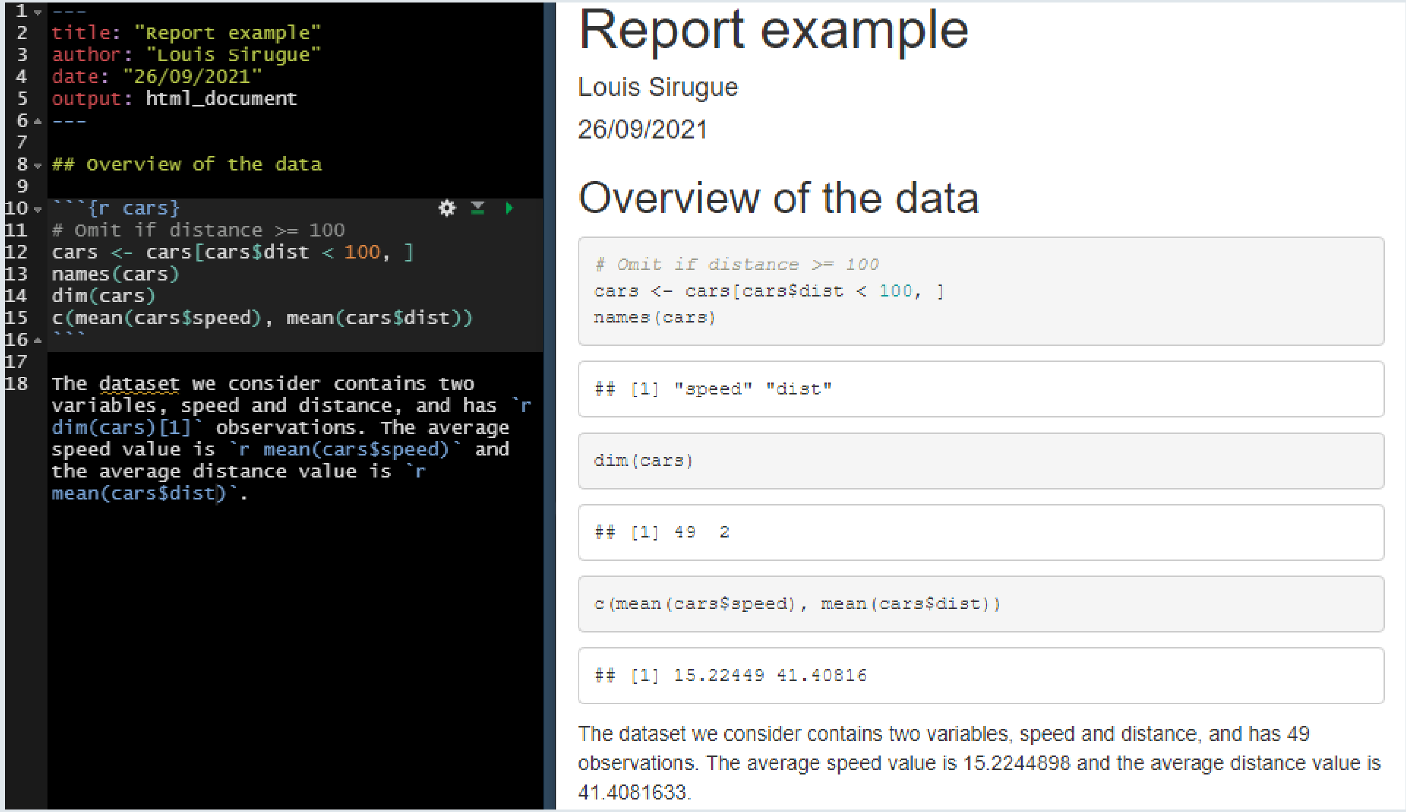
Argument	Meaning
alpha	opacity from 0 to 1
color	color of the geometry
fill	fill color of the geometry
size	size of the geometry
shape	shape for geometries like points
linetype	solid, dashed, dotted, etc.

- If specificed in the geometry it will apply uniformly to every all the geometry
- If assigned to a variable in aes it will vary with the variable according to a scale documented in legend

```
ggplot(data, aes(x = V1, y = V2, size = V3)) +
  geom_point(color = "steelblue", alpha = .6)
```

In Part I we saw

R Markdown: Three types of content



```
---
title: "Report example"
author: "Louis Sirugue"
date: "26/09/2021"
output: html_document
---

## Overview of the data

```{r cars}
# Omit if distance >= 100
cars <- cars[cars$dist < 100, ]
names(cars)
dim(cars)
c(mean(cars$speed), mean(cars$dist))
```

The dataset we consider contains two variables, speed and distance, and has `r dim(cars)[1]` observations. The average speed value is `r mean(cars$speed)` and the average distance value is `r mean(cars$dist)`.
```

Report example

Louis Sirugue

26/09/2021

Overview of the data

```
# Omit if distance >= 100
cars <- cars[cars$dist < 100, ]
names(cars)
```

```
## [1] "speed" "dist"
```

```
dim(cars)
```

```
## [1] 49  2
```

```
c(mean(cars$speed), mean(cars$dist))
```

```
## [1] 15.22449 41.40816
```

The dataset we consider contains two variables, speed and distance, and has 49 observations. The average speed value is 15.2244898 and the average distance value is 41.4081633.

YAML header

Code chunks

Text

In Part I we saw

R Markdown: Useful features

→ **Inline code** allows to include the output of some **R code within text areas** of your report

Syntax	Output
`` `paste("a", "b", sep = "-")` ``	`paste("a", "b", sep = "-")`
`` `r paste("a", "b", sep = "-")` ``	a-b

→ `kable()` for clean html tables and `datatable()` to navigate in large tables

```
kable(results_table)
datatable(results_table)
```

In Part I we saw

LaTeX for equations

- $\LaTeX$ is a convenient way to display mathematical symbols and to structure equations
  - The syntax is mainly based on backslashes and braces

→ What you type in the text area: `$x \neq \frac{\alpha \times \beta}{2}$`
→ What is rendered when knitting the document: $x \neq \frac{\alpha \times \beta}{2}$

- To include a LaTeX equation in R Markdown, you simply have to surround it with the `$` sign:

The mean formula with one `$` on each side

→ For inline equations

$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

The mean formula with two `$` on each side

→ For large/emphasized equations

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Today: *We start Econometrics!*

1. Univariate regressions

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

2. Inference

- 2.1. Standard error
- 2.2. Confidence interval
- 2.3. P-value

3. Multivariate regressions and lm()

- 3.1. Multivariate regressions
- 3.2. The `lm()` function

4. Wrap up!

Today: *We start Econometrics!*

1. Univariate regressions

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

1. Univariate regressions

1.1. Introduction to regressions

- Consider the following dataset

```r
ggcurve <- read.csv("ggcurve.csv")
kable(head(ggcurve, 5), "First 5 rows")
```

First 5 rows

country	ige	gini
Denmark	0.15	0.38
Norway	0.17	0.33
Finland	0.18	0.38
Canada	0.19	0.46
Australia	0.26	0.44

The data contains 2 variables at the country level:

1. **IGE:** Intergenerational elasticity, which captures the % average increase in child income for a 1% increase in parental income

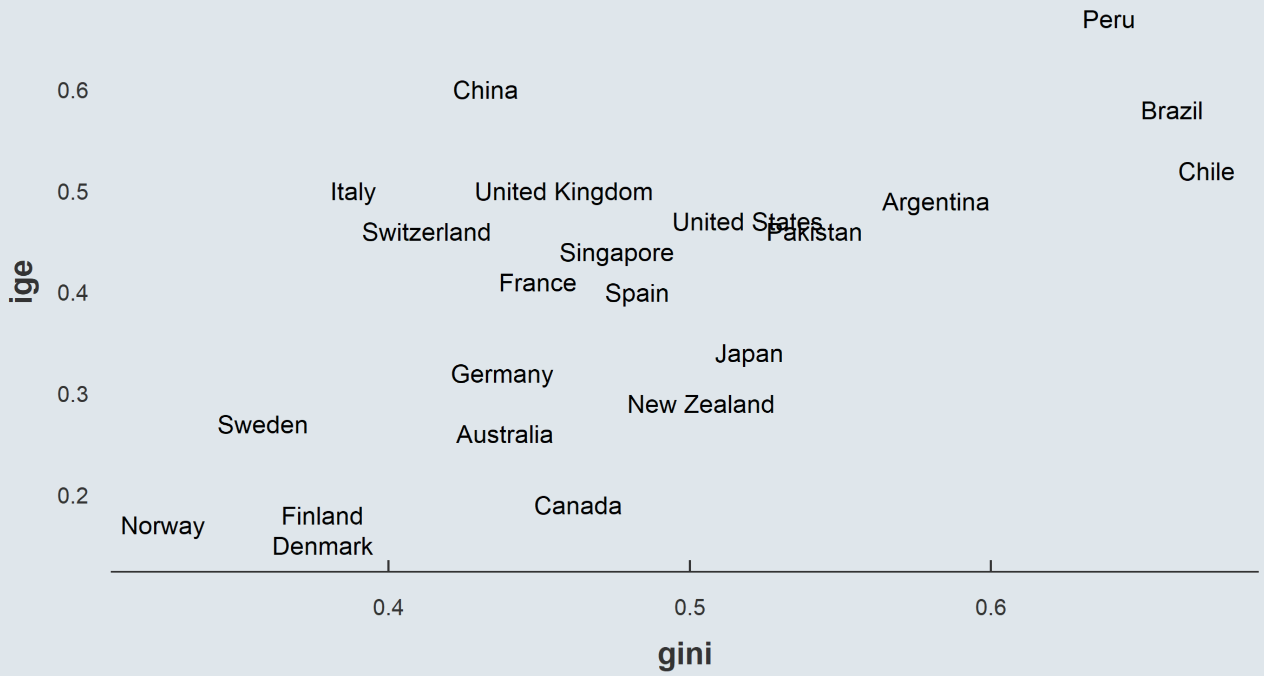
2. **Gini:** Gini index of income inequality from 0 (everybody has the same income) to 1 (a single individual has all the income)

1. Univariate regressions

1.1. Introduction to regressions

- To investigate the relationship between these two variables we can start with a scatterplot:

```
ggplot(ggcurve , aes(x = gini, y = ige, label = country)) + geom_text()
```



1. Univariate regressions

1.1. Introduction to regressions

- We see that the two variables are positively correlated with each other:
  - When one tends to be high relative to its mean, the other as well
  - When one tends to be low relative to its mean, the other as well

```
cor(ggcurve$gini, ggcurve$ige)
```

```
## [1] 0.6517277
```

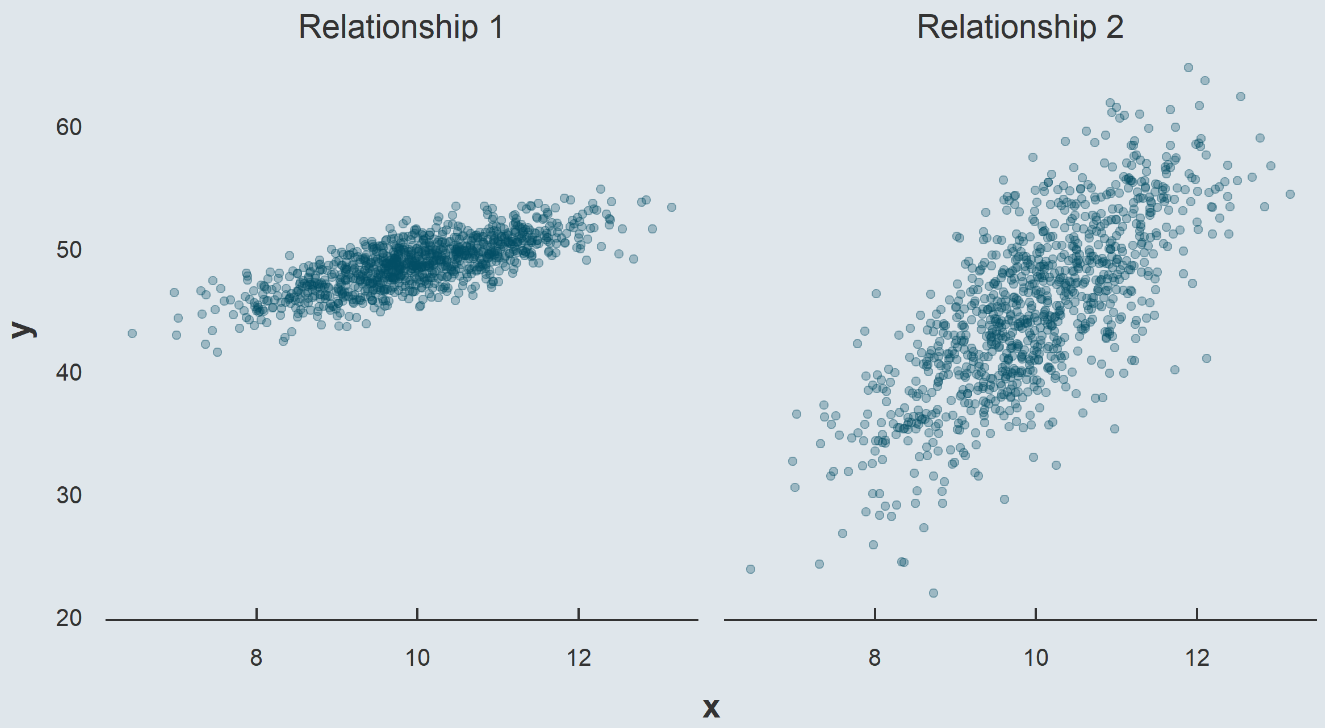
- The correlation coefficient is equal to .65,
  - Remember that the correlation can take values from -1 to 1
  - Here the correlation is indeed positive and fairly strong

→ ***But the correlation does not indicate whether or not a given change in x is associated with a large change in y***

1. Univariate regressions

1.1. Introduction to regressions

- Consider these two relationships :



→ One is less noisy but flatter

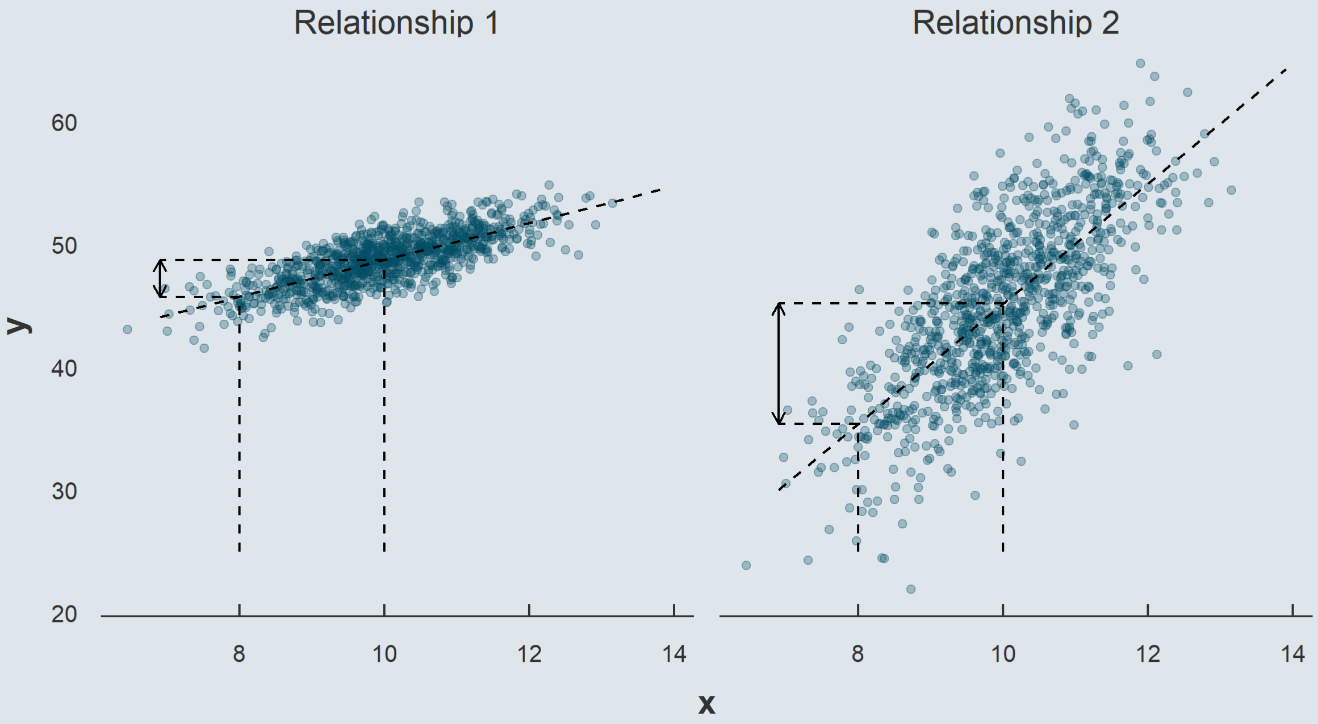
→ One is noisier but steeper

Both have a correlation of .75

1. Univariate regressions

1.1. Introduction to regressions

- Consider these two relationships :



But a given increase in x is not associated with a same increase in y!

1. Univariate regressions

1.1. Introduction to regressions

- Knowing that the income inequality is correlated to intergenerational mobility is one thing
- But should we expect a given increase in income inequality to be associated with a high or low change in intergenerational mobility?

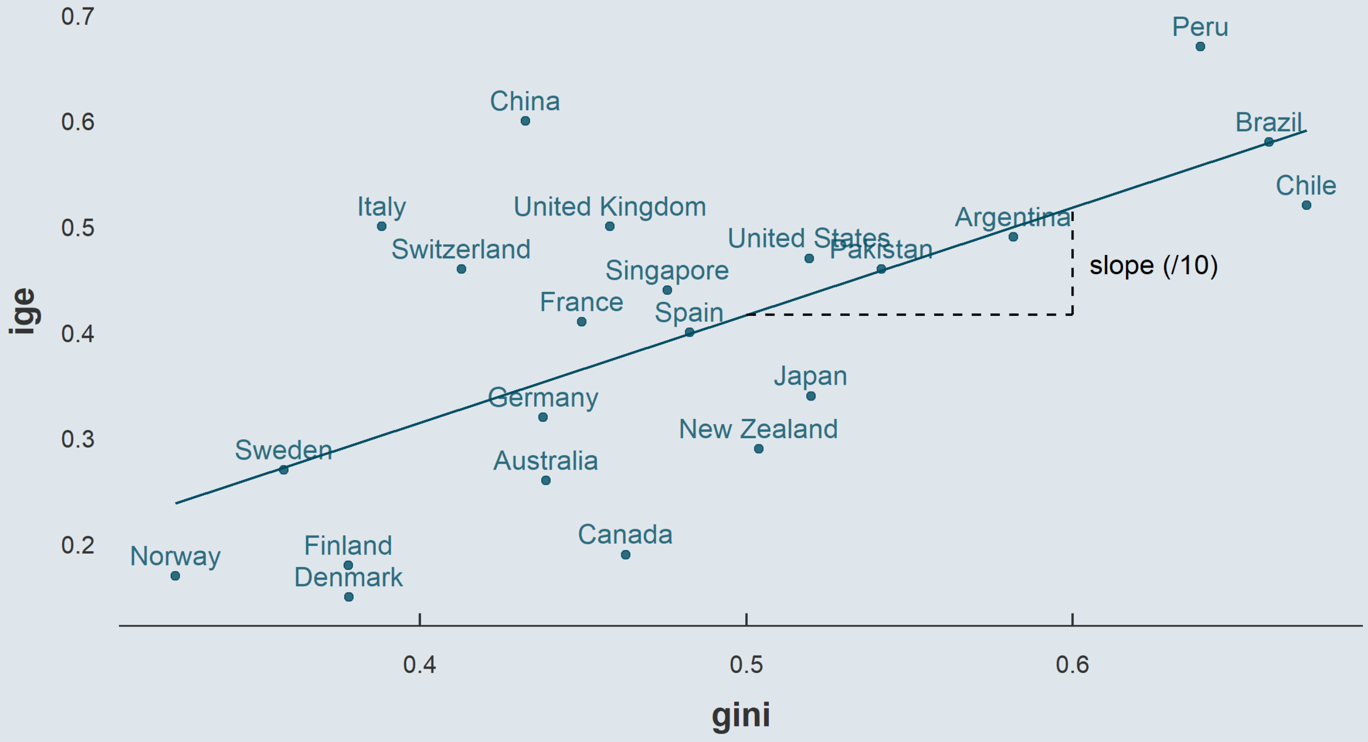
- It is usually the type of question we're interested in:
  - How much more should I expect to earn for an additional year of education?
  - By how many years would life expectancy be expected to decrease for a given increase in air pollution?
  - By how much would test scores increase for a given decrease in the number of students per teacher?
- And this is typically what is of interest for policymakers

→ ***But how to compute this expected change in y for a given change of x?***

1. Univariate regressions

1.2. Coefficients estimation

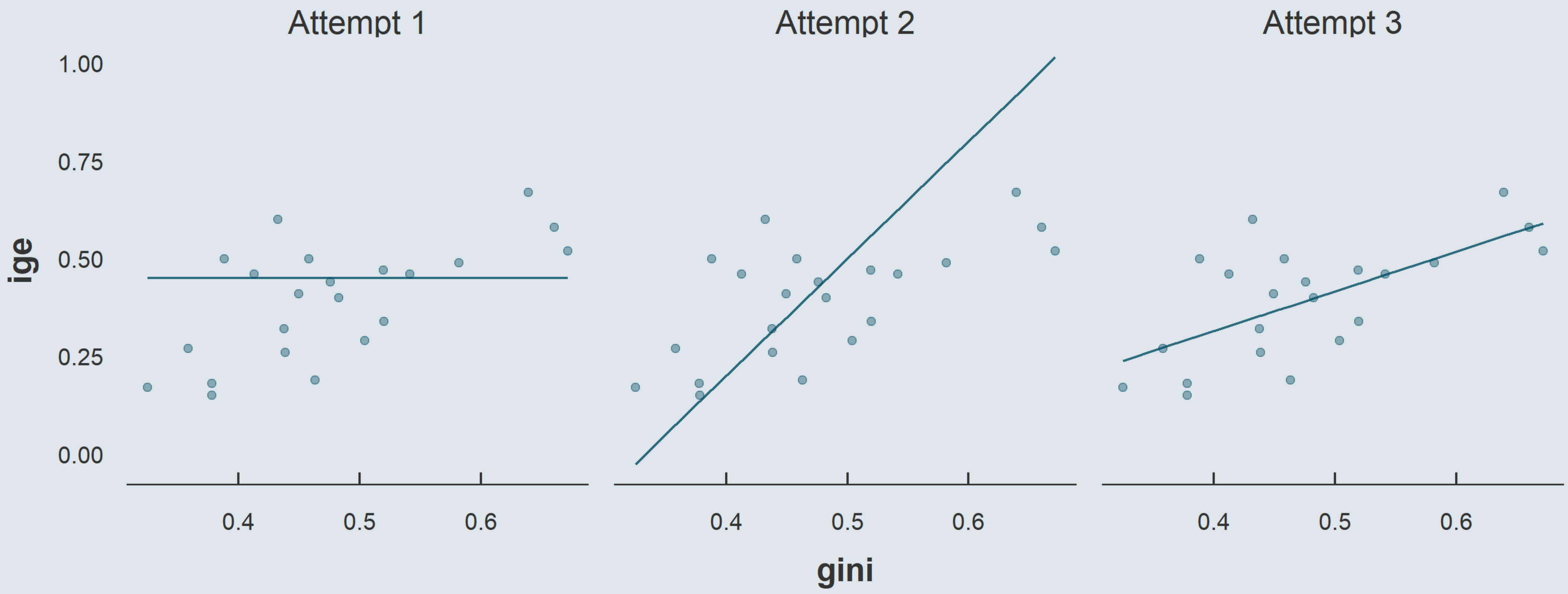
- The idea is to find the line that fits the data the best
  - Such that its slope can indicate how we expect y to change if we increase x by 1 unit



1. Univariate regressions

1.2. Coefficients estimation

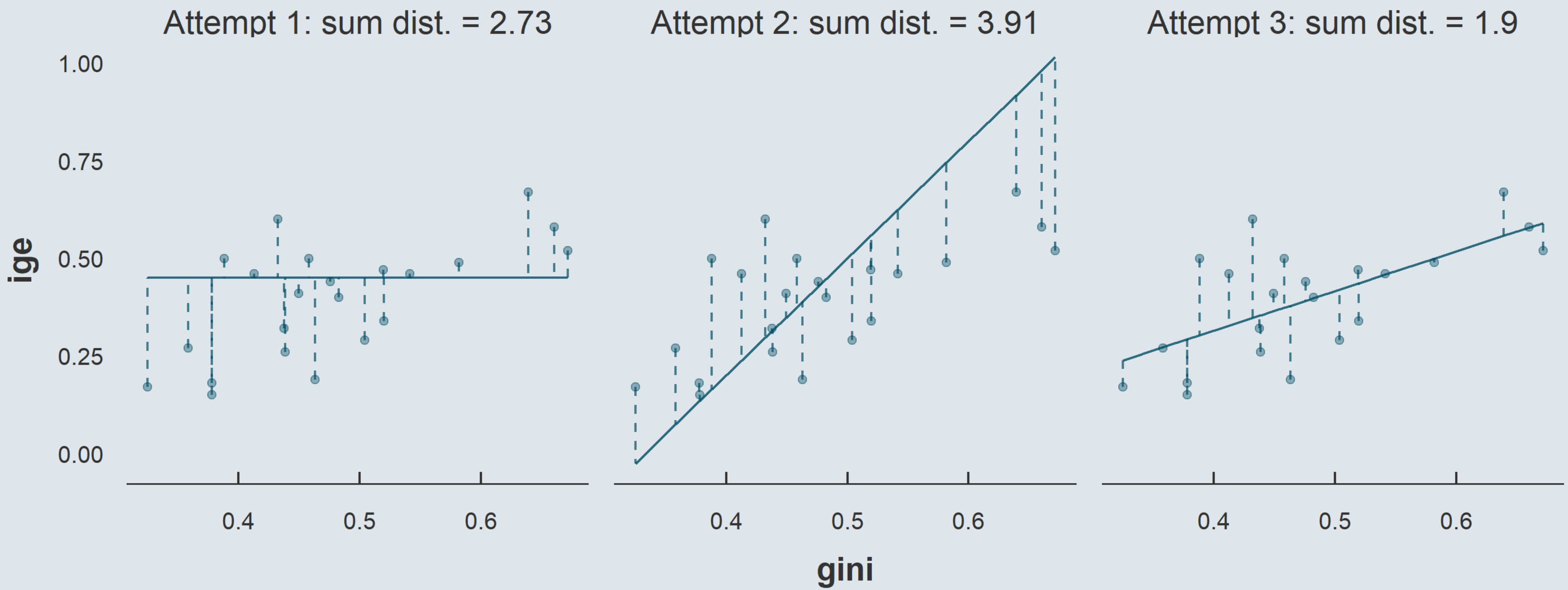
- But how do we find that line ?



1. Univariate regressions

1.2. Coefficients estimation

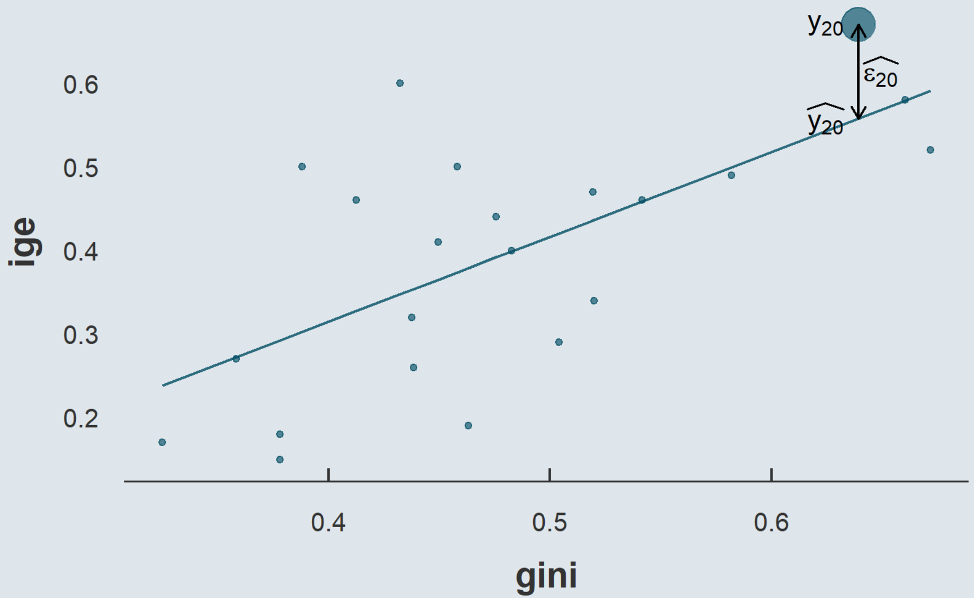
- We try to minimize the distance between each point and our line



1. Univariate regressions

1.2. Coefficients estimation

Take for instance the 20th observation: Peru



And consider the following notations:

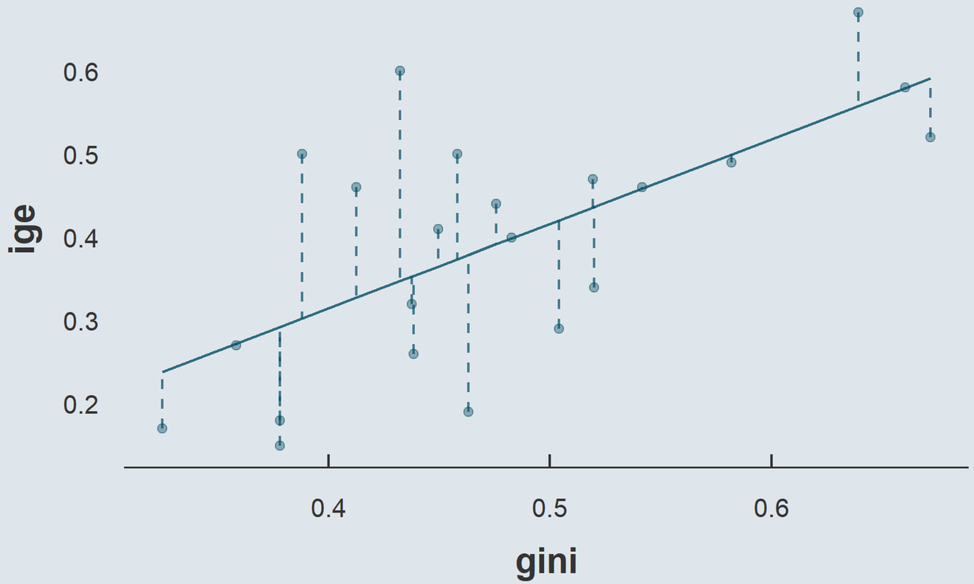
- We denote y_i the ige of the i^{th} country
- We denote x_i the gini of the i^{th} country
- We denote $\widehat{y_i}$ the value of the y coordinate of our line when $x = x_i$

→ The distance between the i^{th} y value and the line is thus $y_i - \widehat{y_i}$

- We label that distance $\widehat{\varepsilon_i}$

1. Univariate regressions

1.2. Coefficients estimation



- Because $\widehat{\varepsilon_i}$ is the value of the distance between a point y_i and its corresponding value on the line $\widehat{y_i}$ we can write:

$$y_i = \widehat{y_i} + \widehat{\varepsilon_i}$$

- And because $\widehat{y_i}$ is a straight line, it can be expressed as

$$\widehat{y_i} = \hat{\alpha} + \hat{\beta} x_i$$

- Where:
  - $\hat{\alpha}$ is the y-intercept
  - $\hat{\beta}$ is the slope

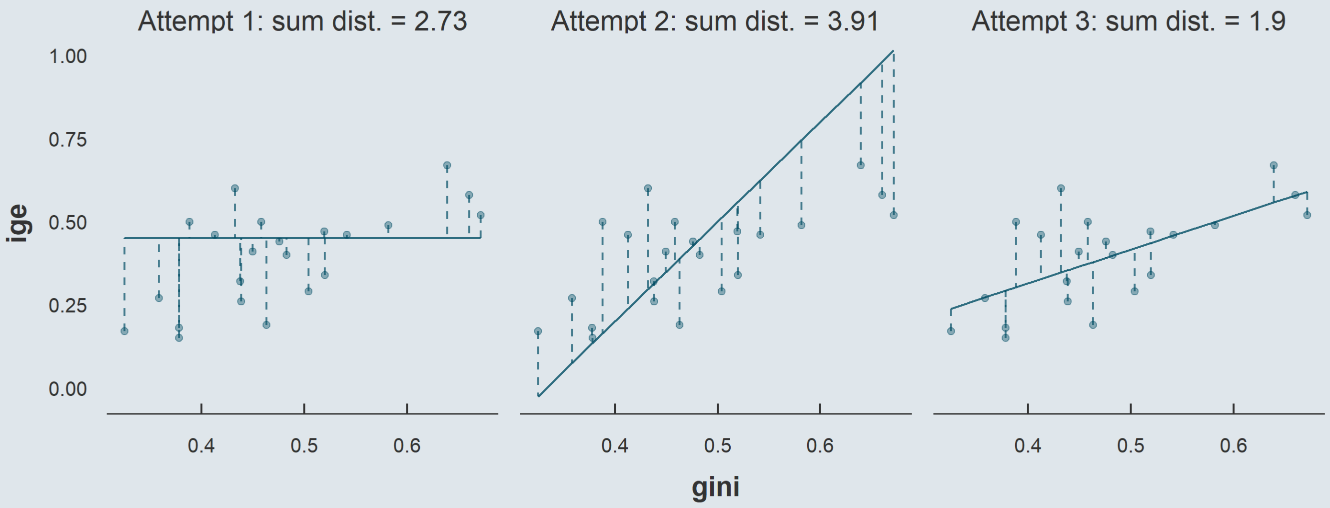
1. Univariate regressions

1.2. Coefficients estimation

- Combining these two definitions yields the equation:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \widehat{\varepsilon_i} \begin{cases} y_i = \widehat{y_i} + \widehat{\varepsilon_i} & \text{Definition of distance} \\ \widehat{y_i} = \hat{\alpha} + \hat{\beta} x_i & \text{Definition of the line} \end{cases}$$

- Depending on the values of $\hat{\alpha}$ and $\hat{\beta}$, the value of every $\widehat{\varepsilon_i}$ will change



Attempt 1: $\hat{\alpha}$ is too high and $\hat{\beta}$ is too low → $\widehat{\varepsilon_i}$ are large

Attempt 2: $\hat{\alpha}$ is too low and $\hat{\beta}$ is too high → $\widehat{\varepsilon_i}$ are large

Attempt 3: $\hat{\alpha}$ and $\hat{\beta}$ seem appropriate → $\widehat{\varepsilon_i}$ are low

1. Univariate regressions

1.2. Coefficients estimation

- We want to find the values of $\hat{\alpha}$ and $\hat{\beta}$ that minimize the overall distance between the points and the line

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} \widehat{\varepsilon_i}^2$$

  - Note that we square $\widehat{\varepsilon_i}$ to avoid that its positive and negative values compensate
  - This method is what we call **Ordinary Least Squares (OLS)**

- To solve this optimization problem, we need to express $\widehat{\varepsilon_i}$ it in terms of alpha $\hat{\alpha}$ and $\hat{\beta}$

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \widehat{\varepsilon_i}$$

$$\iff$$

$$\widehat{\varepsilon_i} = y_i - \hat{\alpha} - \hat{\beta}x_i$$

1. Univariate regressions

1.2. Coefficients estimation

- And our minimization problem writes

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

$$\begin{aligned}
\frac{\partial}{\partial \hat{\alpha}} = 0 \quad &\iff \quad -2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \\
\frac{\partial}{\partial \hat{\beta}} = 0 \quad &\iff \quad -2 x_i \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0
\end{aligned}$$

- Rearranging the first equation yields

$$\sum_{i=1}^{n} y_i - n\hat{\alpha} - \sum_{i=1}^{n} \hat{\beta} x_i = 0 \iff \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

1. Univariate regressions

1.2. Coefficients estimation

- Replacing $\hat{\alpha}$ in the second equation by its new expression writes

$$-2\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \iff -2\sum_{i=1}^{n}\left[y_i - (\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}x_i\right] = 0$$

- And by rearranging the terms we obtain

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Notice that multiplying the nominator and the denominator by $1/n$ yields:

$$\hat{\beta} = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \qquad ; \qquad \hat{\alpha} = \bar{y} - \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \times \bar{x}$$

Practice

1) Write a function that takes two variables `x` and `y` as inputs, computes the $\hat{\alpha}$ and $\hat{\beta}$ coefficients, and returns the two coefficients in a vector as the output

2) Import the data `ggcurve.csv` and use your function to compute by how much the IGE increases on expectation for a one unit increase in the Gini index

3) Plot your results (scatter plot + line)

Remember:

$$\hat{\beta} = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \qquad \hat{\alpha} = \bar{y} - \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)} \times \bar{x}$$

You've got 5 minutes!

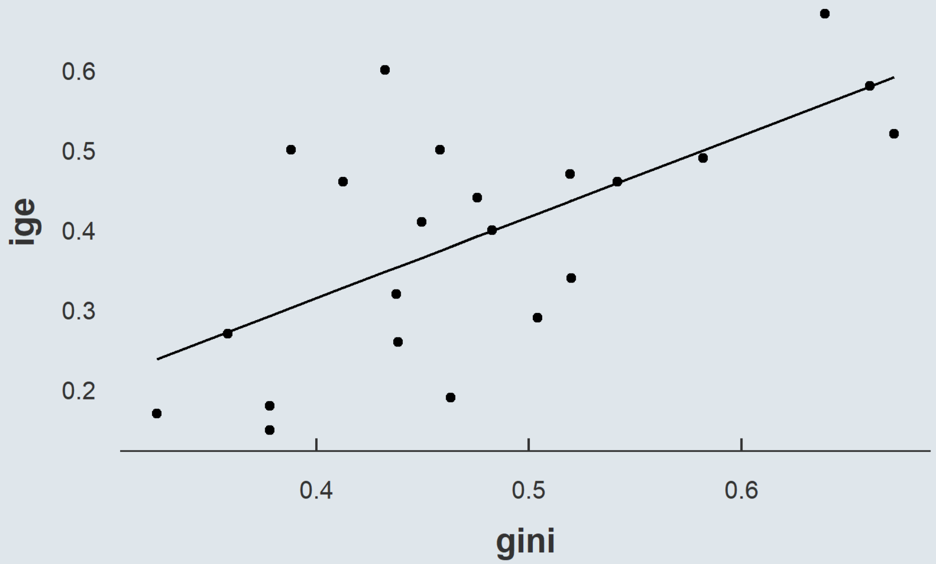
Solution

```r
# Define the function
alpha_beta <- function(x, y) {

  # Compute the beta coefficient
  beta <- cov(x, y)/var(x)

  # Compute the alpha coefficient
  alpha <- mean(y) - (beta * mean(x))

  # Return the two coefficients in a vector
  return(c(alpha, beta))

}


# Read the data
ggcurve <- read.csv("ggcurve.csv")

# Apply the fouction to the data using gini as
# the x variable and ige as the y variable
alpha_beta(ggcurve$gini, ggcurve$ige)
```

```
## [1] -0.09129311  1.01546204
```

Solution

```
# Store the coefficients
coefs <- alpha_beta(ggcurve$gini, ggcurve$ige)

ggcurve %>% # Compute the values on the line
  mutate(line = coefs[1] + gini * coefs[2]) %>%
  # Do the plot
  ggplot(., aes(x = gini)) + geom_point(aes(y = ige)) + geom_line(aes(y = line))
```



Vocabulary

- This equation we're working on is called a regression model

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \widehat{\varepsilon_i}$$

  - We say that we regress y on x to find the coefficients $\hat{\alpha}$ and $\hat{\beta}$ that characterize the regression line
  - We often call $\hat{\alpha}$ and $\hat{\beta}$ *parameters* of the regression because it is what we tune to fit our model to the data

- We also have different names for the x and y variables
  - y is called the *dependent* or *explained* variable
  - x is called the *independent* or *explanatory* variable

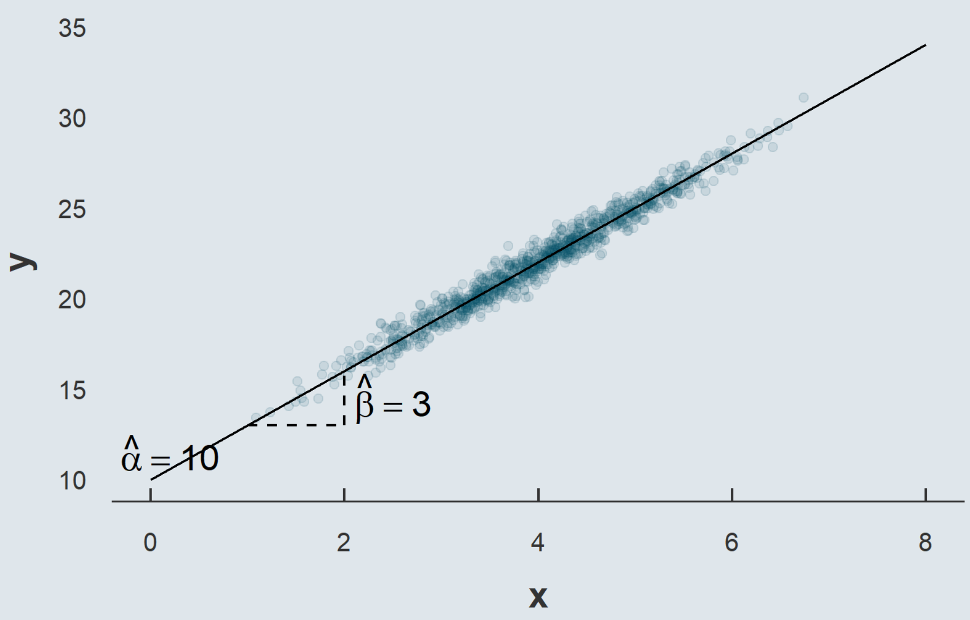
- We call $\widehat{\varepsilon_i}$ the residuals because it is what is left after we fitted the data the best we could

- And $\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i$, i.e., the value on the regression line for a given x_i are called the fitted values

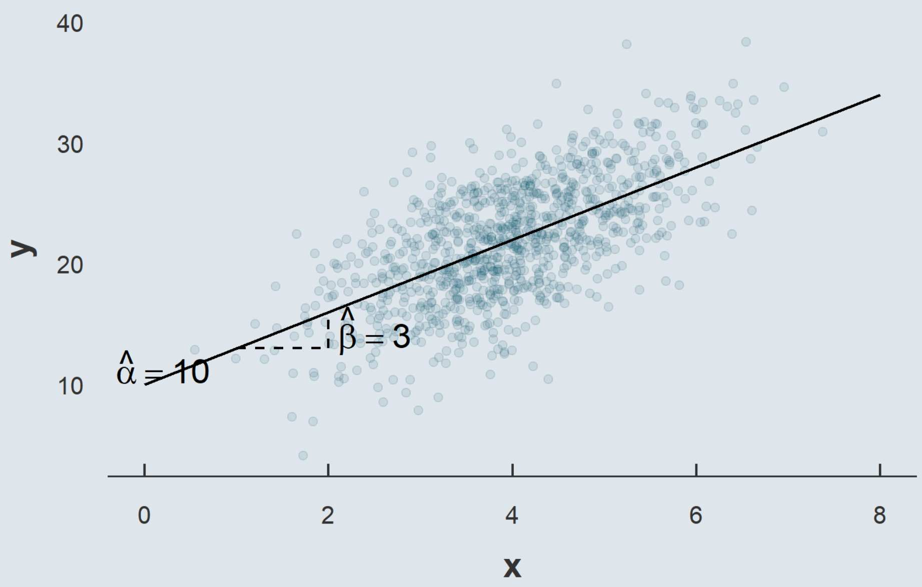
1. Univariate regressions

1.3. Regression fit

- Now we know how to compute the expected change in y for a one unit increase in x by fitting the best straight line we can



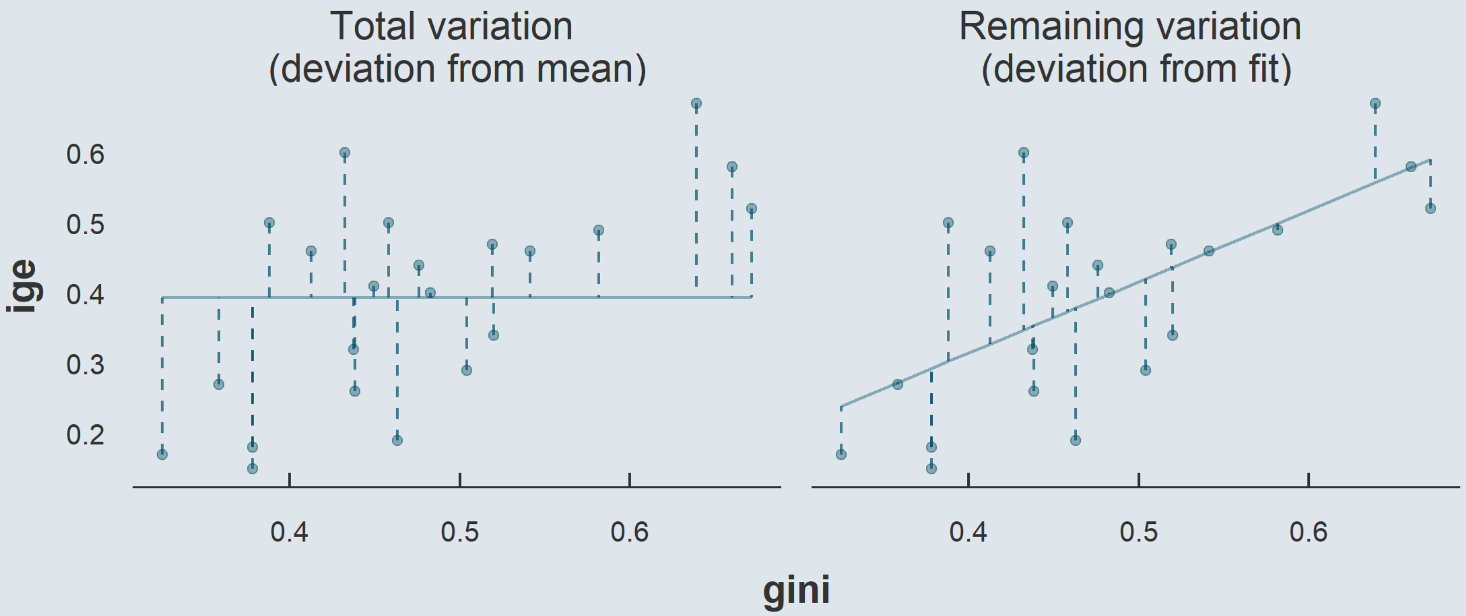
- But even the best line may fail to explain a lot a the variation in y, we need to evaluate the extent to which our model does explain the y variations



1. Univariate regressions

1.3. Regression fit

- We can have an idea of the extent to which our linear model explains the variations in y using
  - The total variation of the y variable (its variance $\sum_{i=1}^{n}(y_i - \bar{y})^2$)
  - The remaining variation of the y variable once its modeled (the sum of squared residuals $\sum_{i=1}^{n}\hat{\varepsilon_i}^2$)



1. Univariate regressions

1.3. Regression fit

- We can then obtain a proper formula from the following reasoning

$$\text{Total variation} = \text{Explained variation} + \text{Remaining variation}$$

$$\frac{\text{Explained variation}}{\text{Total variation}} = 1 - \frac{\text{Remaining variation}}{\text{Total variation}}$$

$$\frac{\text{Explained variation}}{\text{Total variation}} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \equiv \mathrm{R}^2$$

- Because all the terms are sums of squares, we usually talk about *Total Sum of Squares* (TSS), *Explained Sum of Squares* (ESS) and *Residual Sum of Squares* (RSS)

- Because the total sum of squares is the variance of y, the R^2 (also called *coefficient of determination*) can be interpreted as the share of the variance of y explained by the model

Overview

1. Univariate regressions ✓

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

2. Inference

- 2.1. Standard error
- 2.2. Confidence interval
- 2.3. P-value

3. Multivariate regressions and lm()

- 3.1. Multivariate regressions
- 3.2. The `lm()` function

4. Wrap up!

Overview

1. Univariate regressions ✓

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

2. Inference

- 2.1. Standard error
- 2.2. Confidence interval
- 2.3. P-value

2. Inference

2.1. Standard error

- In practice we estimate the parameters of a regression on a given sample of the population of interest

- In our case we have computed $\hat{\alpha}$ and $\hat{\beta}$ using 22 countries
  - But what if instead of having Japan and Canada we had Austria and Latvia?

- The $\hat{\beta}$ from our sample is actually an estimation of the unobserved β of the underlying population

→ ***To make inference possible we would like to know how reliable $\hat{\beta}$ is, how confident we are in its estimation***

2. Inference

2.1. Standard error

- To get an idea of the precision of β, we can estimate its *standard error*
  - We won't go through the theoretical computations together, but let's have a look at the formula

$$\text{se}(\hat{\beta}) = \sqrt{\widehat{\text{Var}(\hat{\beta})}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2}{(n - \#\text{parameters}) \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Notice that the variance, and thus the standard error of our estimate:
  - Decreases as our sample gets bigger
  - Gets larger if the points are further away from the regression line on average for a given variance of x

And keep in mind that while the ***standard deviation*** *measures the amount of variability, or dispersion, from the individual data values to the mean, the* ***standard error*** *measures how far an estimate from a given sample is likely to be from the true parameter of interest*

Practice

1) Compute the standard error of our $\hat{\beta}$ coefficient estimate

Remember:

$$\text{se}(\hat{\beta}) = \sqrt{\widehat{\text{Var}(\hat{\beta})}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon_i}^2}{(n - \#\text{parameters}) \sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

You've got 5 minutes!

Solution

```r
# Store the alpha and beta parameters
coefs <- alpha_beta(ggcurve$gini, ggcurve$ige)

data <- ggcurve %>%

  # Rename x and y for convenience
  rename(x = gini, y = ige) %>%

  # Compute what we need
  mutate(yhat = coefs[1] + (x * coefs[2]),
         e2 = (y - yhat)^2,
         x_xbar2 = (x - mean(x))^2)

# Compute numerator and denominator
num <- sum(data$e2)
den <- (nrow(data) - 2) * sum(data$x_xbar2)

se_beta <- sqrt(num/den) # Square root
se_beta
```

```
## [1] 0.2642477
```

$$\text{se}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon_i}^2}{(n - \#\text{parameters}) \sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

2. Inference

2.2. Confidence intervals

- The magnitude of the standard error gives an indication of the precision of our estimate:
  - The larger the estimate relative to its standard error, the more precise the estimate

- But standard errors are not easily interpretable by themselves
  - A more direct way to get a sense of the precision for inference is to construct a confidence interval

→ Instead of saying that our estimation $\hat{\beta}$ is equal to 1.02, we would like to say that we are 95% sure that the actual β lies between two given values

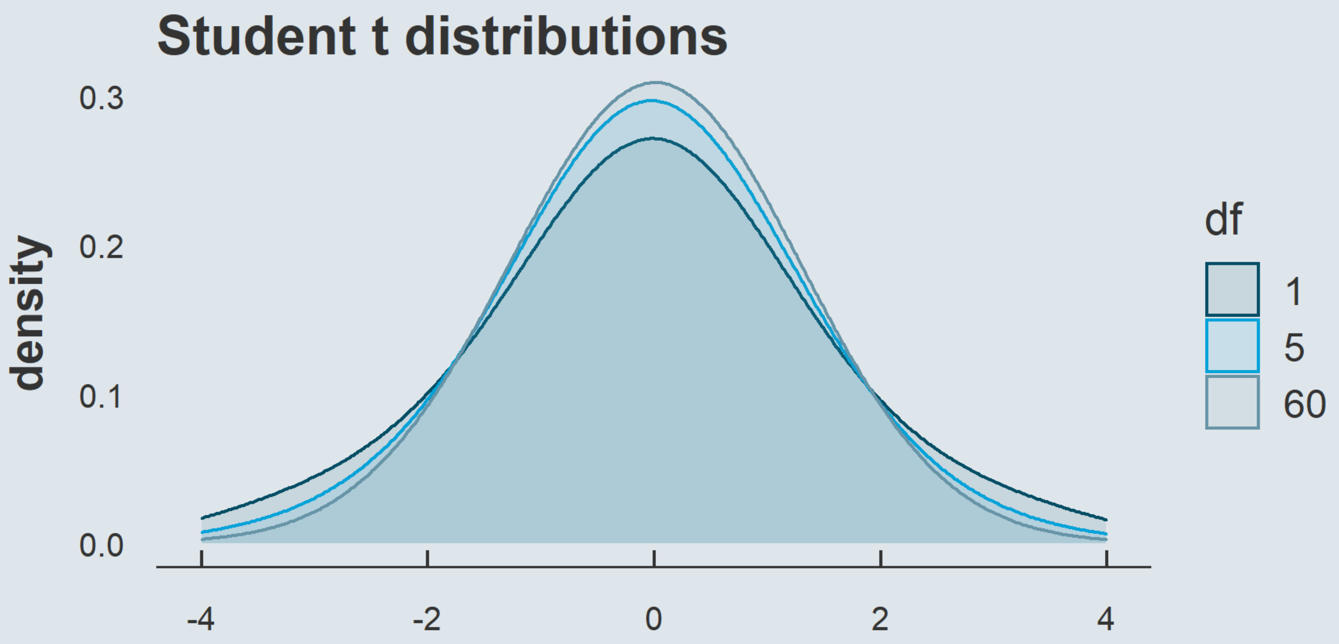
- To obtain a confidence interval we can use the fact that under specific conditions (that you're gonna see next year) it is possible to derive how this object is distributed:

$$\hat{t} \equiv \frac{\hat{\beta} - \beta}{\text{se}(\hat{\beta})}$$

2. Inference

2.2. Confidence intervals

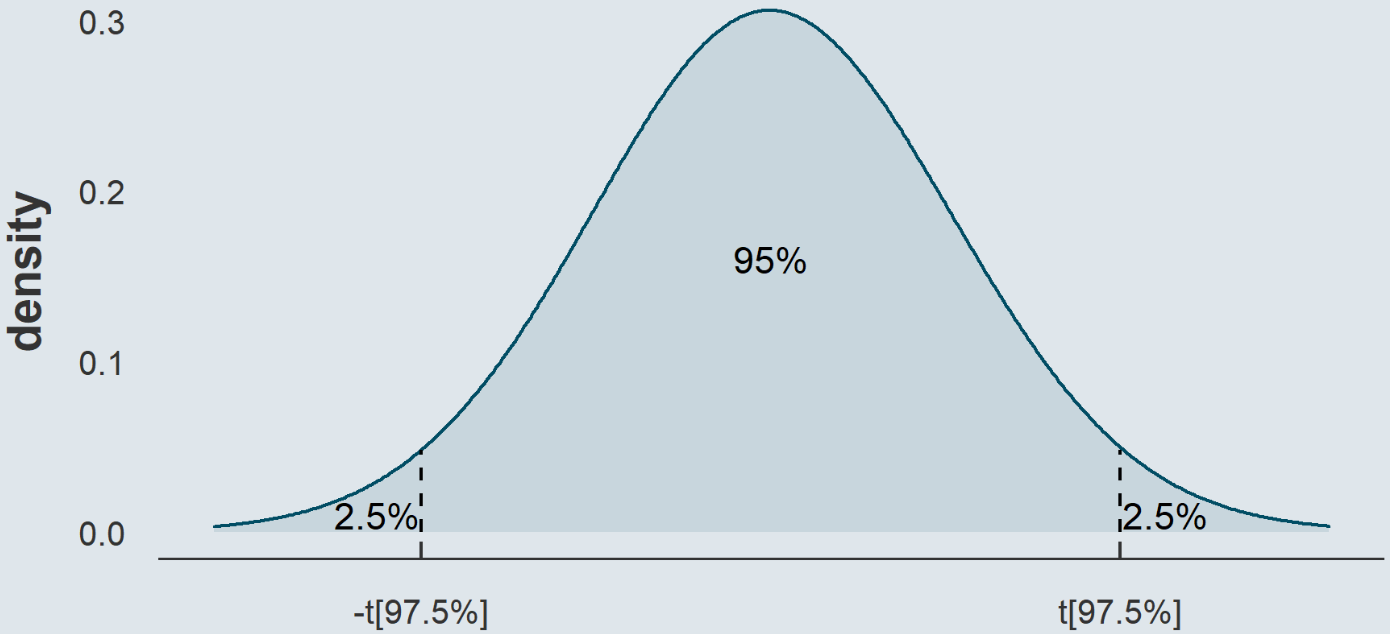
- Theory shows that $\hat{t} \equiv \frac{\hat{\beta}-\beta}{\mathrm{se}(\hat{\beta})}$ follows a Student t distribution whose number of degrees of freedom is equal to n (in our case 22 countries) minus the number of parameters estimated in the model (in our case 2: α and β)



2. Inference

2.2. Confidence intervals

- Denote $t_{97.5\%}$ the value such that 97.5% of the distribution is below that value
  - Then 95% of the distribution lies between $-t_{97.5\%}$ and $t_{97.5\%}$



2. Inference

2.2. Confidence intervals

- Because we know that $\hat{t} \equiv \frac{\hat{\beta}-\beta}{\mathrm{se}(\hat{\beta})}$ follows this distribution, we know that it has a 95% chance to fall within the two values $-t_{97.5\%}$ and $t_{97.5\%}$

$$\Pr\left[-t_{97.5\%} \leq \frac{\hat{\beta}-\beta}{\mathrm{se}(\hat{\beta})} \leq t_{97.5\%}\right] = 95\%$$

- Rearranging the terms yields:

$$\Pr\left[\hat{\beta} - t_{97.5\%} \times \mathrm{se}(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{97.5\%} \times \mathrm{se}(\hat{\beta})\right] = 95\%$$

- Thus, we can say that there is a 95% chance for β to be within

$$\hat{\beta} \pm t_{97.5\%} \times \mathrm{se}(\hat{\beta})$$

2. Inference

2.2. Confidence intervals

- We already know $\hat{\beta}$ and $\text{se}(\hat{\beta})$, but we can't compute $t_{97.5\%}$ manually

- To know the value of $t_{97.5\%}$ we need to rely on the qt() function
  - The first argument is the share of the distribution that should be below the value we're looking for (97.5%)
  - The second argument is the number of degrees of freedom of our model (the number of observations minus the number of parameters)

```
qt(.975, 20)
```

```
## [1] 2.085963
```

- We can then compute the 95% confidence interval of the true β from our previous computations:

```
c(coefs[2] - (qt(.975, 20) * se_beta), coefs[2] + (qt(.975, 20) * se_beta))
```

```
## [1] 0.4642511 1.5666730
```

2. Inference

2.3. P-value

- ***Confidence intervals*** are very effective to get a sense of the precision of our estimates and of the **range of values the true parameters could reasonably take**

- But the ***p-value*** is what we tend to ultimately focus on, it is the **% chance that the our estimation of the true parameter is different from 0 just coincidentally**

- **Confidence intervals and p-values are tightly linked**
  - If there is a 4% chance that a parameter equal to 2 is different from 0, I know that the 95% confidence interval will start above 0 but quite close, and stop a bit before 4
  - If a 95% confidence interval is bounded by 4 and 5, I know the the p-value will be way below 5%

- But these two indicators are **complementary** to easily get the full picture:
  - With a p-value we can easily know how sure we are that the parameter is different from 0, but it is difficult to get a sense of the set of values the parameters can reasonably take
  - With the confidence interval it is the opposite

2. Inference

2.3. P-value

- **Computation:** The principle is the same as for standard errors but the reasoning is reversed
  - For *confidence intervals*: we want to know among which values the parameter has a given percentage chance to fall into
  - For *p-value*: we want to know with which percentage chance 0 is out of the set of values that the parameter could reasonably take

- **Vocabulary:** We talk about *significance level*
  - When $\text{P-value} \leq .05$, we say that the estimate is significant(ly different from 0) at the 5% level
  - When the p-value is greater than a given threshold of acceptability, we say that the estimate is not significant

- **In practice:** Usually in Economics we use the 5% threshold
  - But this is arbitrary, in other fields the benchmark p-value is different
  - With this threshold we're wrong once in 20 times

Overview

1. Univariate regressions ✓

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

2. Inference ✓

- 2.1. Standard error
- 2.2. Confidence interval
- 2.3. P-value

3. Multivariate regressions and lm()

- 3.1. Multivariate regressions
- 3.2. The `lm()` function

4. Wrap up!

Overview

1. Univariate regressions ✓

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

2. Inference ✓

- 2.1. Standard error
- 2.2. Confidence interval
- 2.3. P-value

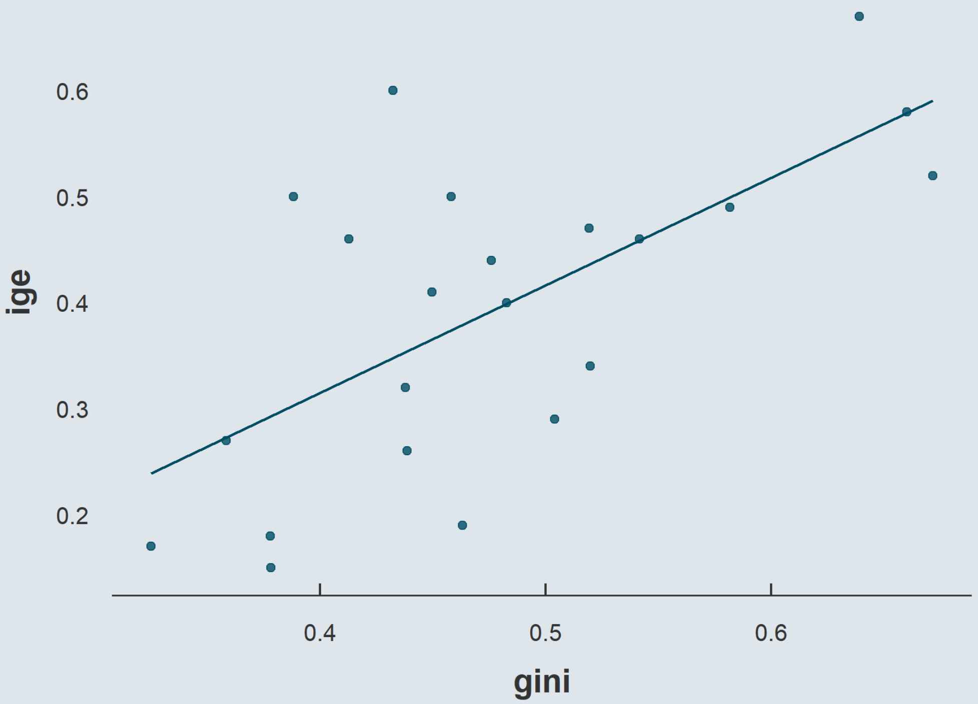
3. Multivariate regressions and lm()

- 3.1. Multivariate regressions
- 3.2. The `lm()` function

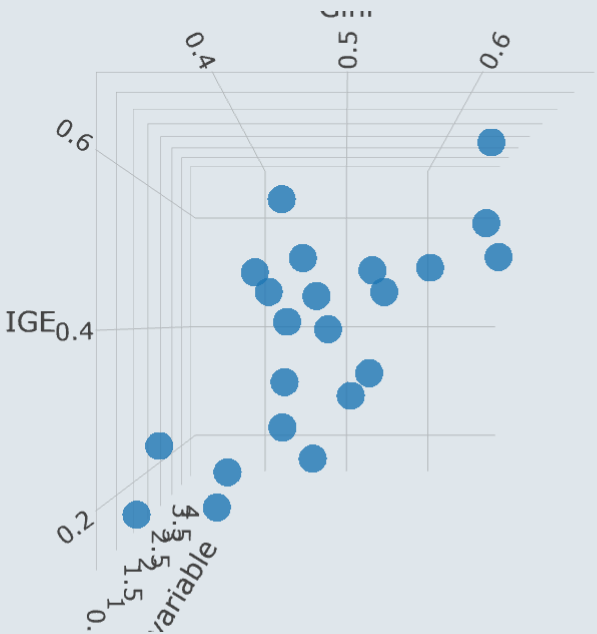
3. Multivariate regressions and lm()

3.1. Multivariate regressions

- So far we fit a line in a relationship between two variables

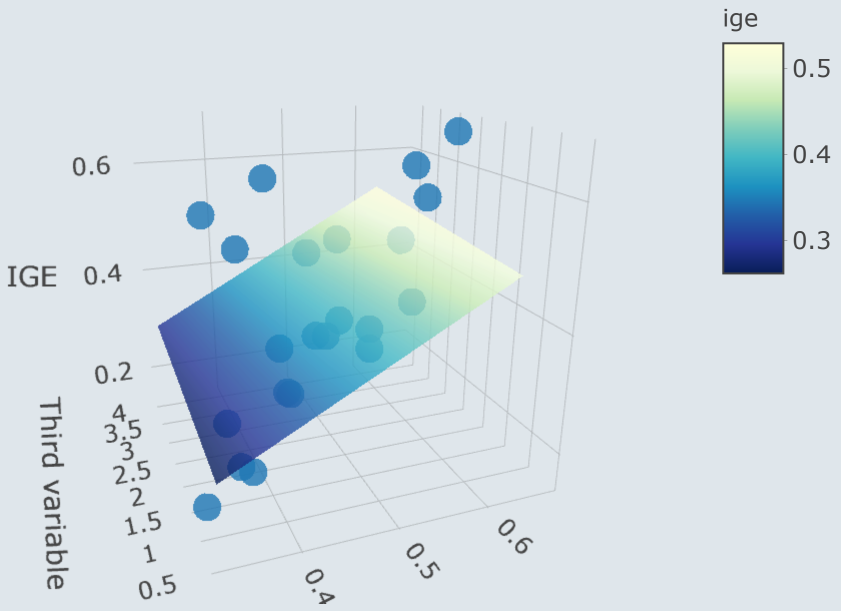


- What should we do if we want to account for a third variable? *(pivot the plot)*



3. Multivariate regressions and lm()

3.1. Multivariate regressions



→ We can fit a plane characterized by the parameters $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ from the multivariate regression estimation:

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \hat{\varepsilon}_i$$

- $\hat{\alpha}$ is the expected value of y when both x_1 and x_2 equal 0
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are the slopes of the plane along the x_1 and x_2 axes

3. Multivariate regressions and lm()

3.1. Multivariate regressions

- We can follow this reasoning and add more dimensions
  - Adding a third independent variable and fit a hyperplane in $\mathbb{R}^4$, and so on
  - But we can't have more variables than we have observations for the parameters to be identified

- Note that the plane that best fits the data does not necessarily has the slopes of the lines that best fit the separate 2D scatterplots
  - Imagine regressing earnings on sex and then adding occupation to the model
  - This may change the initial $\hat{\beta}_1$ and $\text{se}(\hat{\beta}_1)$ because part of the relationship between sex and earnings can be explained by occupation segregation

- If you add an x_2 to your model, you estimate how y is expected to change for a given increase in x_1 by taking into account the fact that x_2 may play a role in the relationship of interest between y and x_1
  - In that case we say that we *control* for x_2, that x_2 is a *control variable*

3. lm() and multivariate regressions

3.2. The `lm()` function

- In R there is a function that computes everything we saw today
  - The `lm()` function for **l**inear **m**odel
  - You have to indicate your regression model in the `formula` argument, and to specify the data

```
lm(formula = ige ~ gini, data = ggcurve)
```

```
##
## Call:
## lm(formula = ige ~ gini, data = ggcurve)
##
## Coefficients:
## (Intercept)         gini
##    -0.09129      1.01546
```

- We recover the $\hat{\alpha}$ and $\hat{\beta}$ estimates we computed manually

- To get more information on the model we can use the summary() function

3. lm() and multivariate regressions

```
summary(lm(ige ~ gini, ggcurve))
```

```
##
## Call:
## lm(formula = ige ~ gini, data = ggcurve)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.188991 -0.088238 -0.000855  0.047284  0.252310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.09129    0.12870  -0.709  0.48631
## gini         1.01546    0.26425   3.843  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1159 on 20 degrees of freedom
## Multiple R-squared:  0.4247,    Adjusted R-squared:  0.396
## F-statistic: 14.77 on 1 and 20 DF,  p-value: 0.001016
```

- It gives information on:
  - The distribution of $\hat{\varepsilon}_i$
  - The estimated parameters of our model (estimate, standard error, t-value = estimate ÷ standard error, p-value)
  - It puts symbols next to p-values when they exceed a given threshold
  - And general information about the model, R squared, degrees of freedom, etc.

3. lm() and multivariate regressions

3.2. The `lm()` function

- To get specific components from the lm summary, you can use the `$` and `[]` subsetting symbols:

```
summary(lm(ige ~ gini, ggcurve))$r.squared
```

```
## [1] 0.424749
```

```
summary(lm(ige ~ gini, ggcurve))$coefficients
```

```
##                Estimate Std. Error    t value    Pr(>|t|)
## (Intercept) -0.09129311  0.1287045 -0.7093234 0.486311455
## gini         1.01546204  0.2642477  3.8428420 0.001015706
```

```
summary(lm(ige ~ gini, ggcurve))$coefficients[2, "Estimate"]
```

```
## [1] 1.015462
```

Overview

1. Univariate regressions ✓

- 1.1. Introduction to regressions
- 1.2. Coefficients estimation
- 1.3. Regression fit

2. Inference ✓

- 2.1. Standard error
- 2.2. Confidence interval
- 2.3. P-value

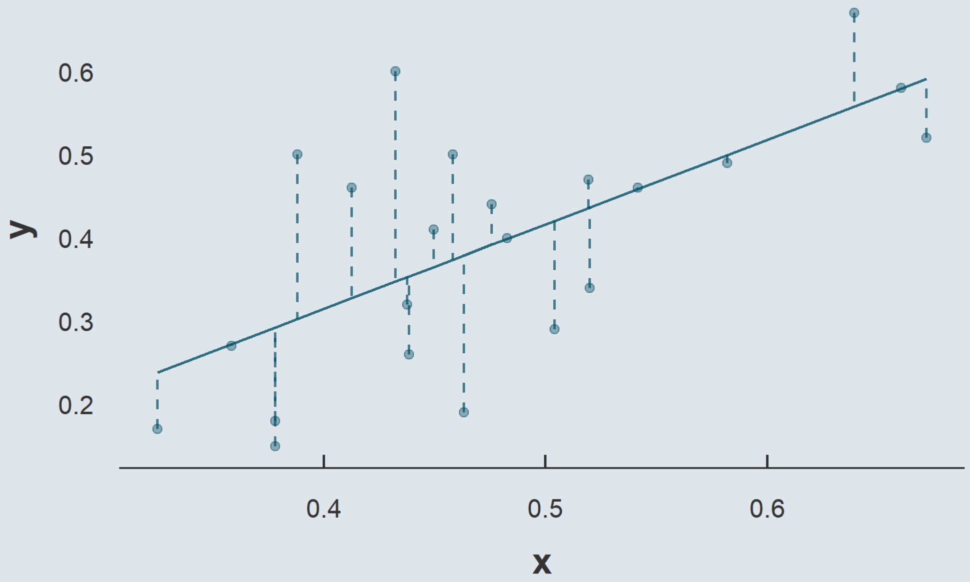
3. Multivariate regressions and lm() ✓

- 3.1. Multivariate regressions
- 3.2. The `lm()` function

4. Wrap up!

4. Wrap up!

1) The regression line minimizes the distance between the line and the data points



- This can be expressed with the regression equation

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{\varepsilon}_i$$

- Where $\hat{\alpha}$ is the intercept and $\hat{\beta}$ the slope of the line $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, and $\hat{\varepsilon}_i$ the distances between the points and the line

$$\hat{\beta} = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \times \bar{x}$$

4. Wrap up!

2) The estimated coefficient is not enough to draw any conclusion

- In practice we estimate the parameters of a regression on a given sample of the population of interest
- The $\hat{\beta}$ from our sample is actually an estimation of the unobserved true β of the underlying population

→ To make inference possible we would like to know how reliable $\hat{\beta}$ is, how confident we are in its estimation

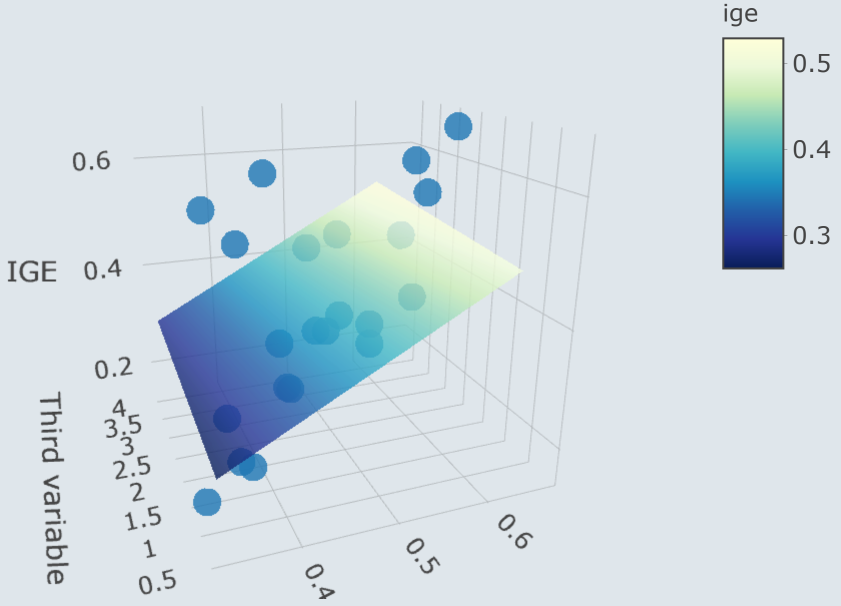
X% confidence interval: Range of values in which we are X% sure that the true value we want to estimate will fall

P-value: Probability that our estimate is different from 0 just by chance

→ A coefficient is significant at the 95% confidence level (5% significance level) if 0 is outside its 95% confidence interval $\Leftrightarrow$ if the p-value is smaller than 5%

4. Wrap up!

3) There can be more than 1 independent variable in a regression model



4) And regression models can be estimated with the `lm()` R function

```
model <- summary(lm(y ~ x1 + x2 + x3))
model$coefficients[, c(1, 2, 4)]
```

```
##              Estimate Std. Error Pr(>|t|)
## (Intercept)    3.0327     1.5823   0.0556
## x1             5.1801     1.6243   0.0015
## x2            -6.8980     1.5251   0.0000
## x3            -0.9861     1.5695   0.5300
```

```
model$r.squared
```

```
## [1] 0.03010538
```