Univariate regressions

Lecture 8

Louis SIRUGUE

CPES 2 - Fall 2022

In Part I we saw

- **Different classes of R objects**

```r
class("numeric") # What would be the output and why?
```

- **Vectors**

```r
match(8, c(6, 1, 9, 5, 8, 4)) # What would be the output and why?
```

- **Functions**

```r
age_from_ssn <- function(ssn) {
  return(2022 - (as.numeric(substr(ssn, 2, 3)) + 1900))
}
```

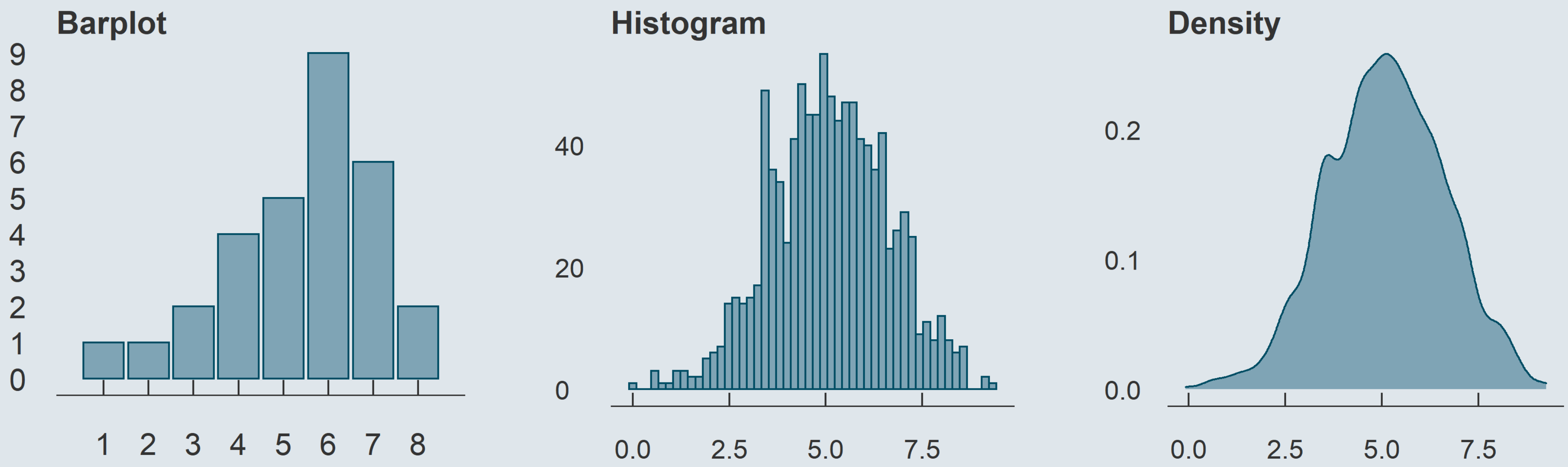
- **Packages**

```r
library(tidyverse)
```

In Part I we saw

Distributions

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

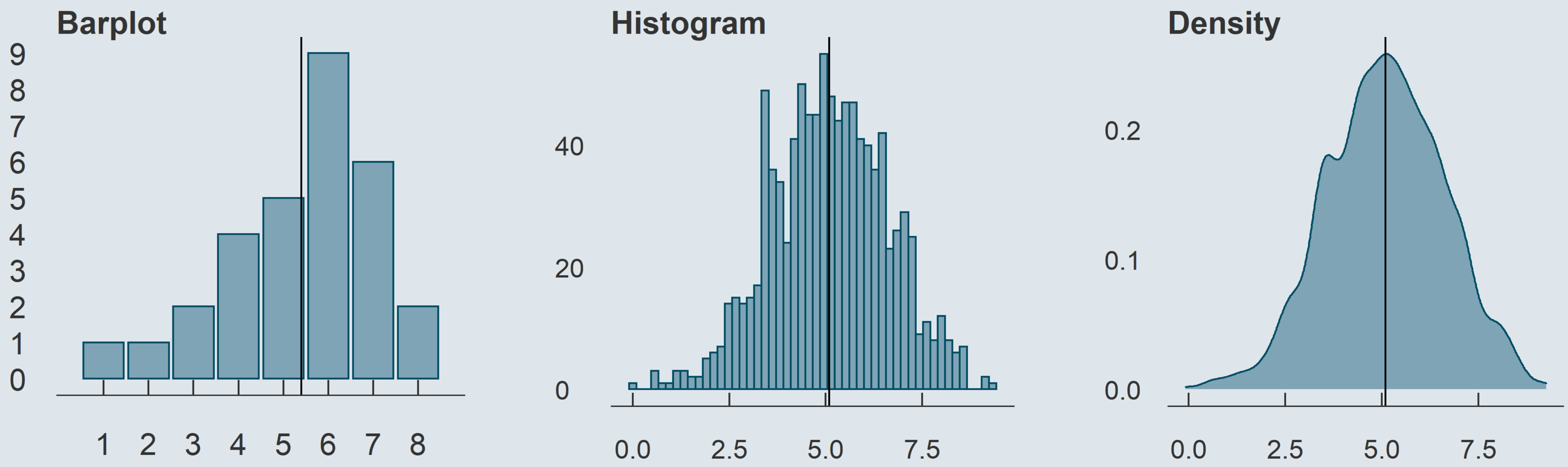


- We can describe a distribution with:

In Part I we saw

Distributions

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

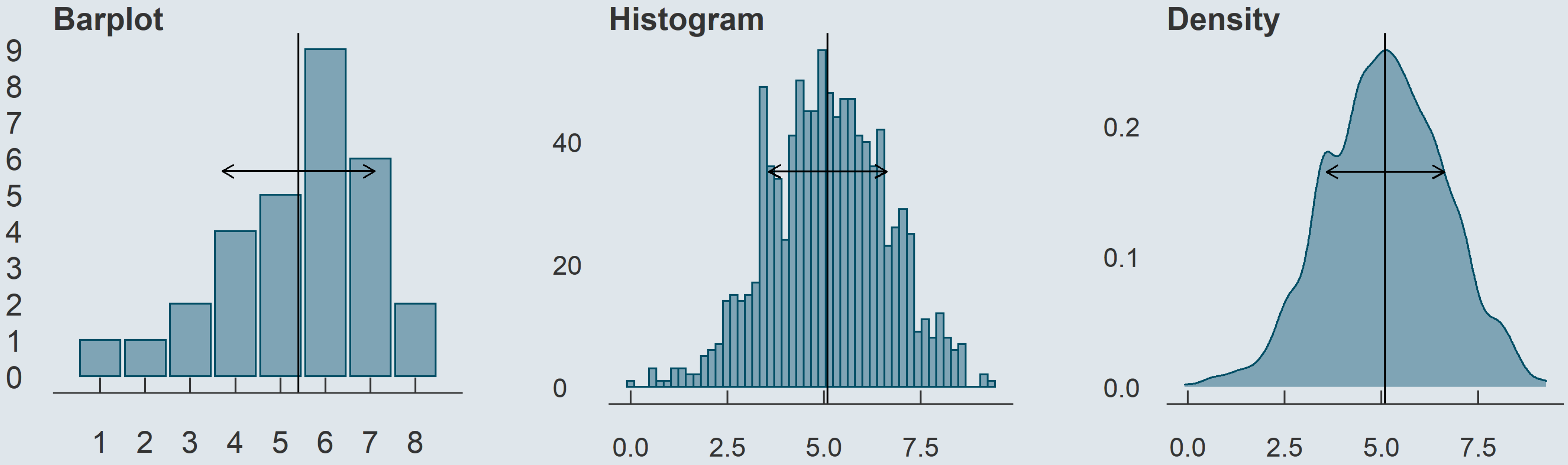


- We can describe a distribution with:
  - Its **central tendency**

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

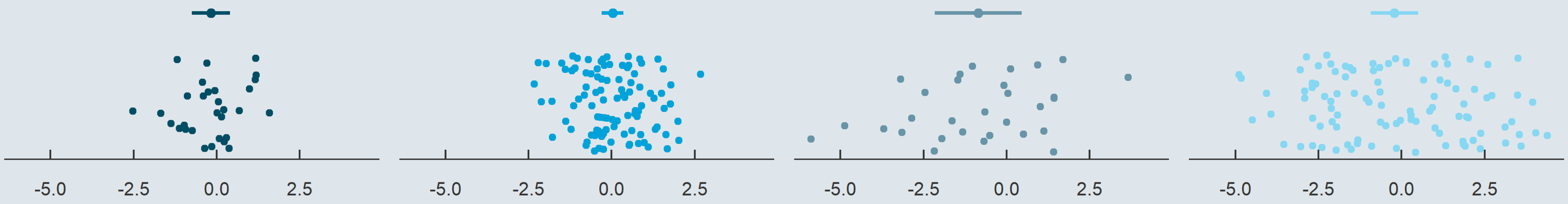
Inference

- In Statistics, we view variables as a given realization of a **data generating process**
    - Hence, the **mean** is what we call an **empirical moment**, which is an **estimation**...
    - ... of the **expected value**, the **theoretical moment** of the DGP we're interested in

- To know how confident we can be in this estimation, we need to compute a **confidence interval**

$$[\bar{x} - t_{n-1,\,97.5\%} \times \frac{\text{SD}(x)}{\sqrt{n}};\; \bar{x} + t_{n-1,\,97.5\%} \times \frac{\text{SD}(x)}{\sqrt{n}}]$$

    - It gets **larger** as the **variance** of the distribution of x increases
    - And gets **smaller** as the **sample size** n increases



In Part I we saw

Read data

```
starbucks <- read.csv("C:/User/Documents/folder/starbucks.csv", sep = ";", encoding = "UTF-8")
```

***→ Make sure to use / and not ***

Chaining operations

```
starbucks %>%
  arrange(-Calories) %>%
  select(Beverage_category, Beverage_prep, Calories) %>%
  filter(row_number() <= 3)
```

```
##              Beverage_category Beverage_prep Calories
## 1   Signature Espresso Drinks        2% Milk      510
## 2   Signature Espresso Drinks        Soymilk      460
## 3 Frappuccino® Blended Coffee     Whole Milk      460
```

→ Make sure to view your data at each step

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

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 customization tools

Item to customize	Main functions
Axes	scale_[x/y]_[continuous/discrete]
Baseline theme	theme_[void/minimal/.../dark]()
Annotations	geom_[[h/v]line/text](), annotate()
Theme	theme(axis.[line/ticks].[x/y] = ...,

Main types of geometry

Geometry	Function
Bar plot	geom_bar()
Histogram	geom_histogram()
Area	geom_area()
Line	geom_line()
Density	geom_density()
Boxplot	geom_boxplot()
Violin	geom_violin()
Scatter plot	geom_point()

In Part I we saw

Main types of aesthetics

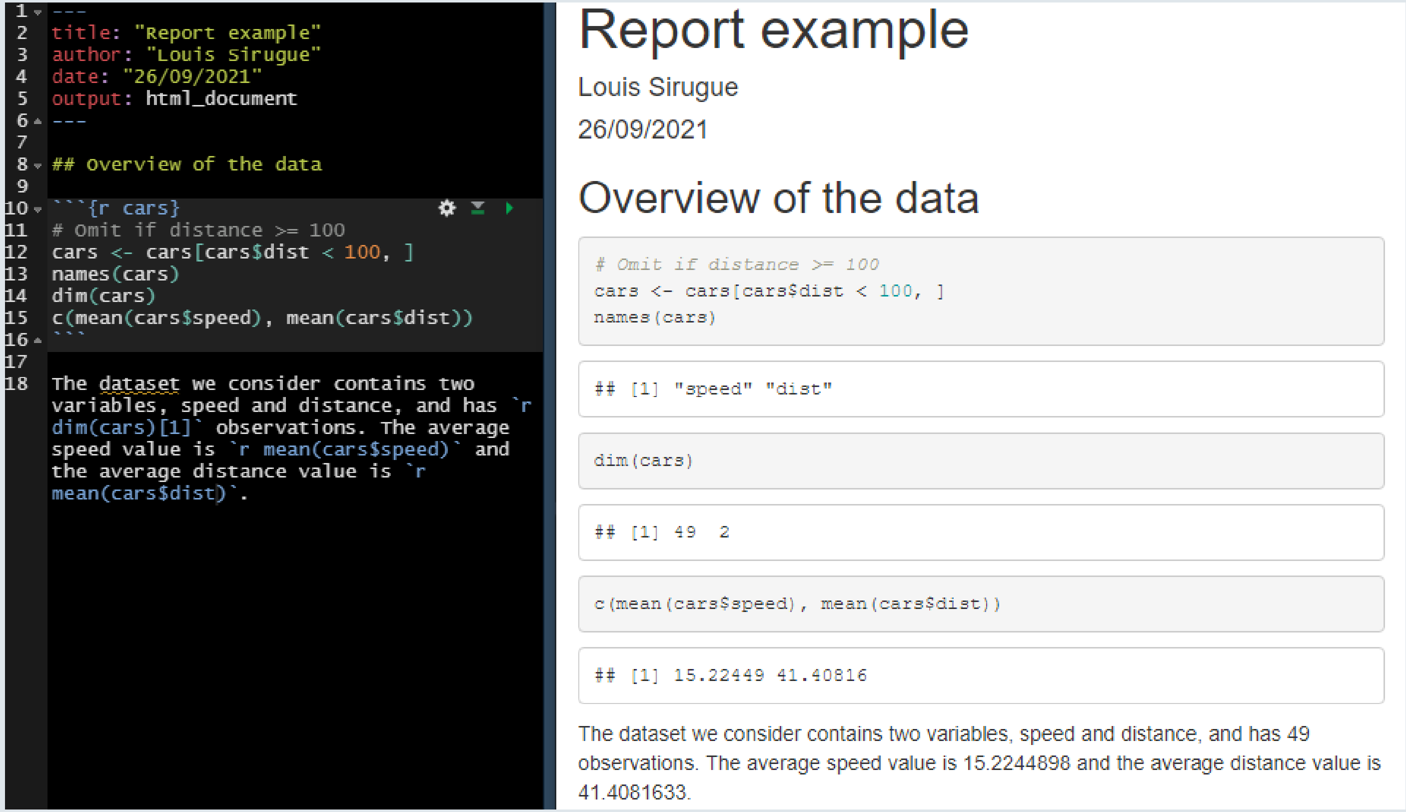
Argument	Meaning
alpha	opacity from 0 to 1
color	color of the geometry
fill	fill color of the geometry
size	size of the geometry
shape	shape for geometries like points
linetype	solid, dashed, dotted, etc.

- If specified **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



YAML header

Code chunks

Text

In Part I we saw

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!*

Today: *We start Econometrics!*

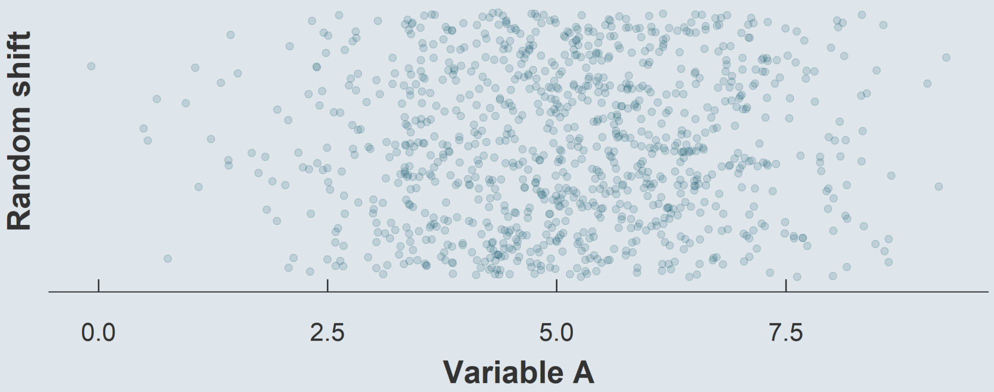
1. Joint distributions

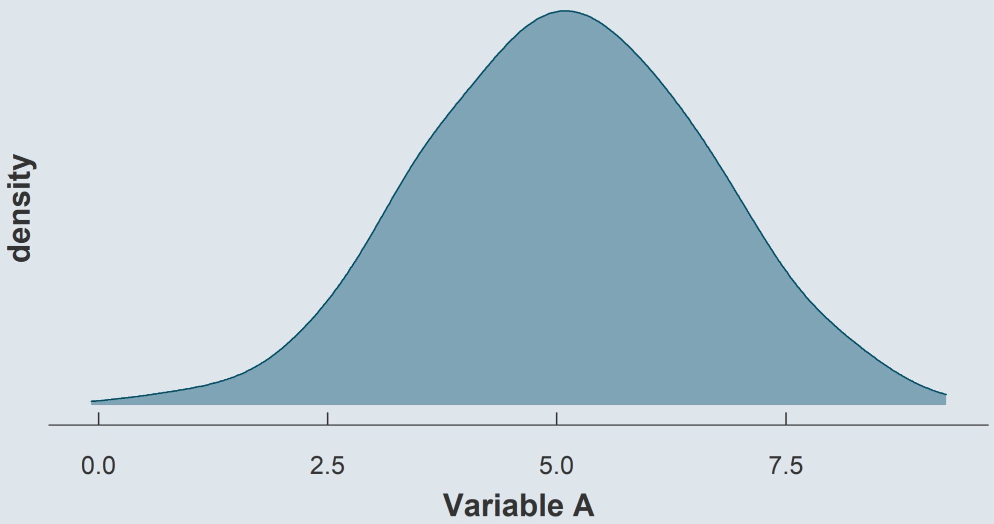
- 1.1. Definition
- 1.2. Covariance
- 1.3. Correlation

1. Joint distributions

1.1. Definition

- The **joint distribution** shows the **values** and associated **frequencies** for **two variables** simultaneously
  - Remember how the **density** could represent the distribution of a **single variable**

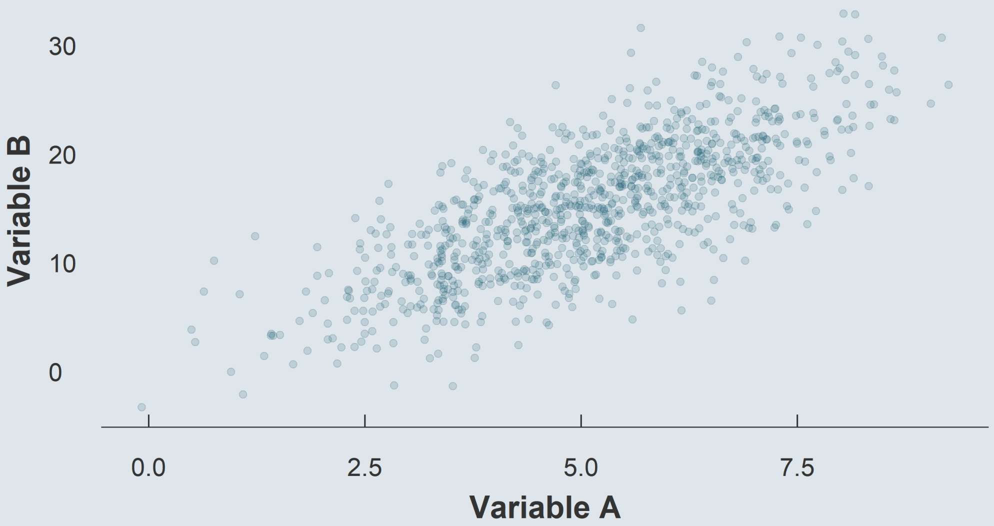


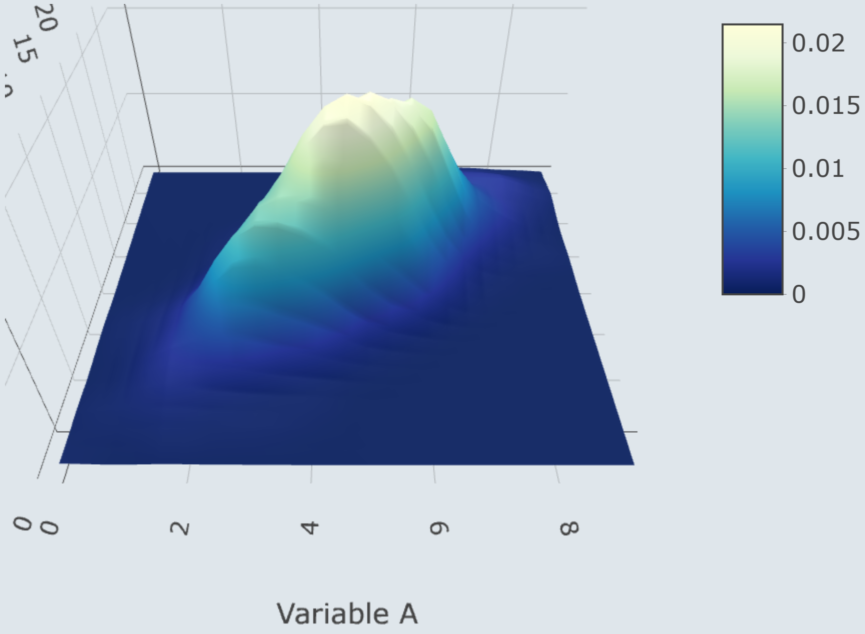


1. Joint distributions

1.1. Definition

- The **joint distribution** shows the **values** and associated **frequencies** for **two variables** simultaneously
  - Remember how the **density** could represent the distribution of a **single variable**
  - The **joint density** can represent the joint distribution of **two variables**



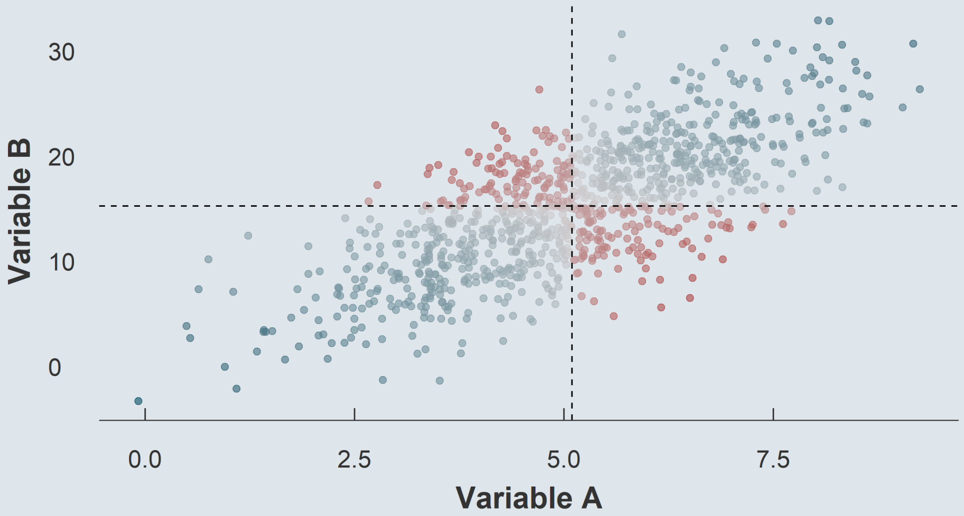


1. Joint distributions

1.2. Covariance

- When describing a **single distribution**, we're interested in its **spread** and **central tendency**
- When describing a **joint distribution**, we're interested in the **relationship** between the two variables
    - This can be characterized by the ***covariance***

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

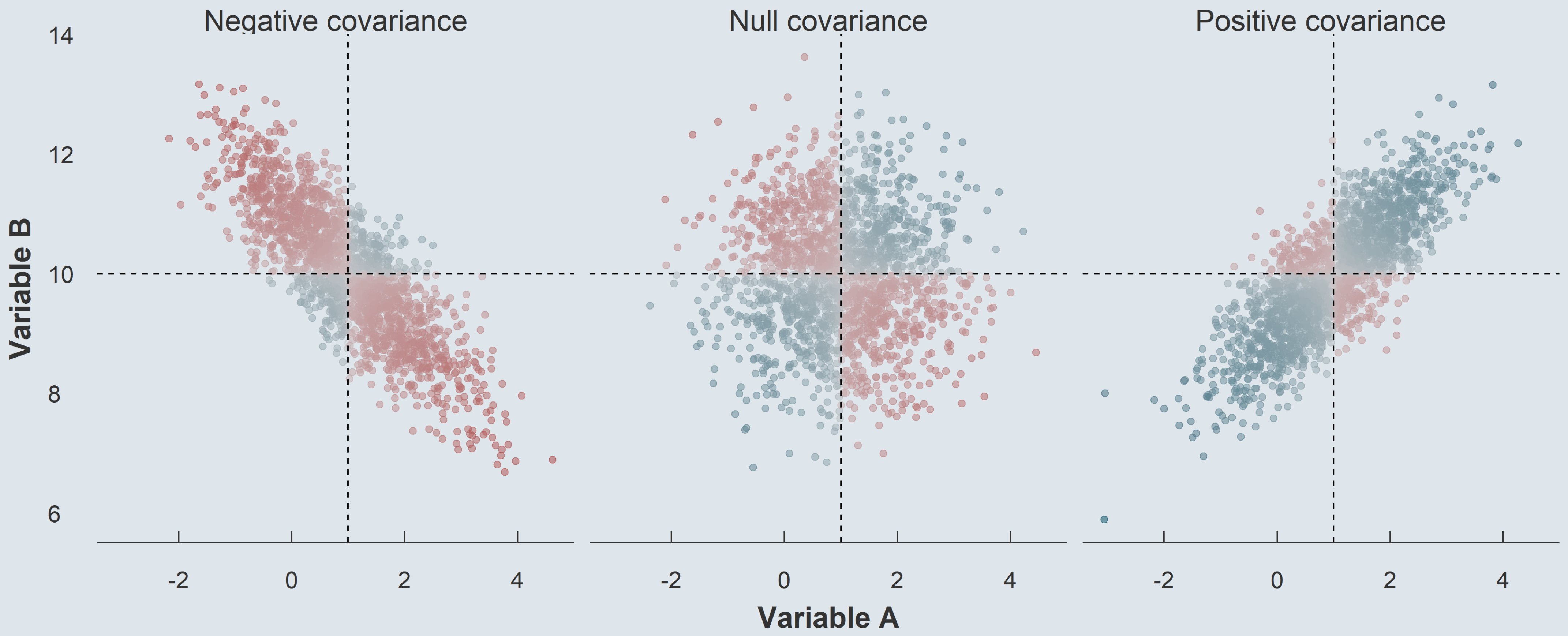


*If **y** tends to be **large** relative to its mean when **x** is **large** relative to its mean, their **covariance** is **positive***

*Conversely, if **one** tends to be **large** when the **other** tends to be **low**, the **covariance** is **negative***

1. Joint distributions

1.2. Covariance



1. Joint distributions

1.3. Correlation

- One disadvantage of the **covariance** is that is it **not standardized**
  - You **cannot** directly **compare** the covariance of two pairs of completely different variables
  - Given distance variables will have a larger covariance in centimeters than in meters

➜ Theoretically the **covariance** can take **values** from $-\infty$ to $+\infty$

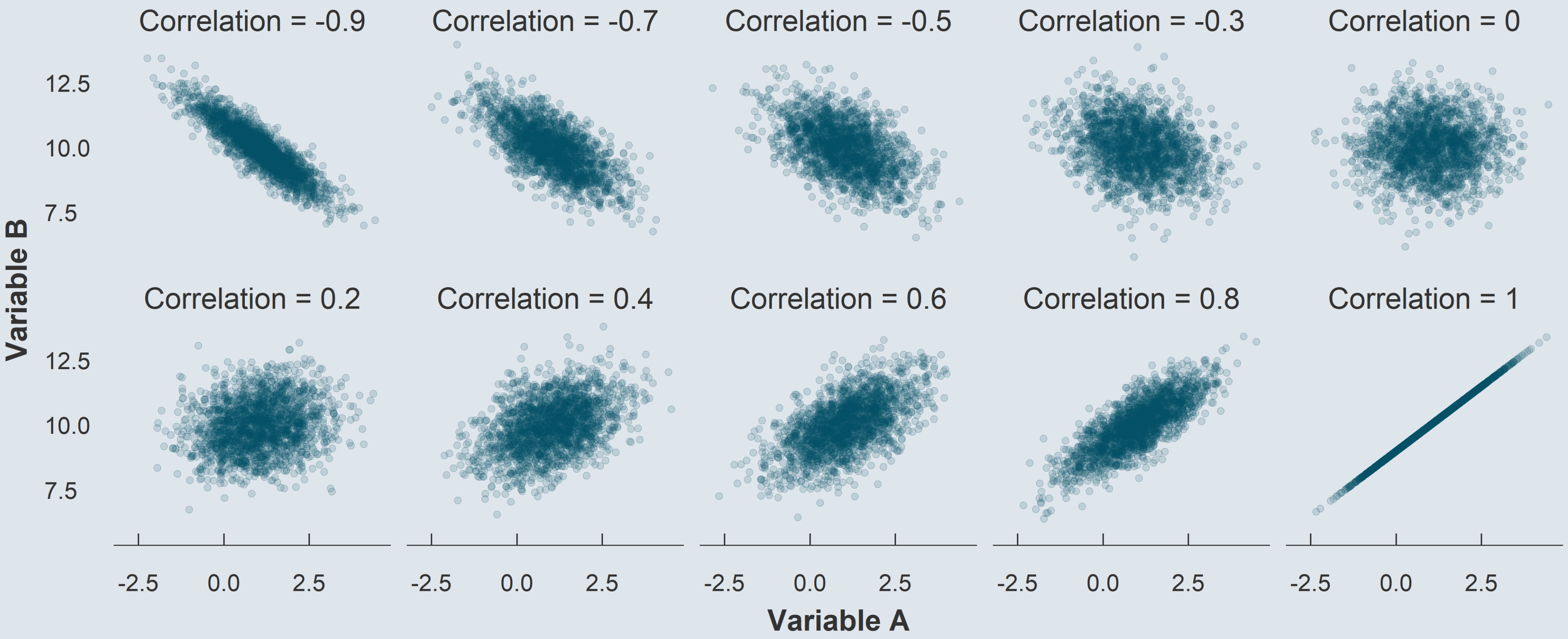
- To **net out** the covariance from the **unit** of the data, we can **divide** it by $\text{SD}(x) \times \text{SD}(y)$
  - We call this **standardized** measure the **correlation**
  - Correlations coefficients are **comparable** because they are independent from the unit of the data

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

➜ The **correlation** coefficient is bounded between **values** from -1 to 1

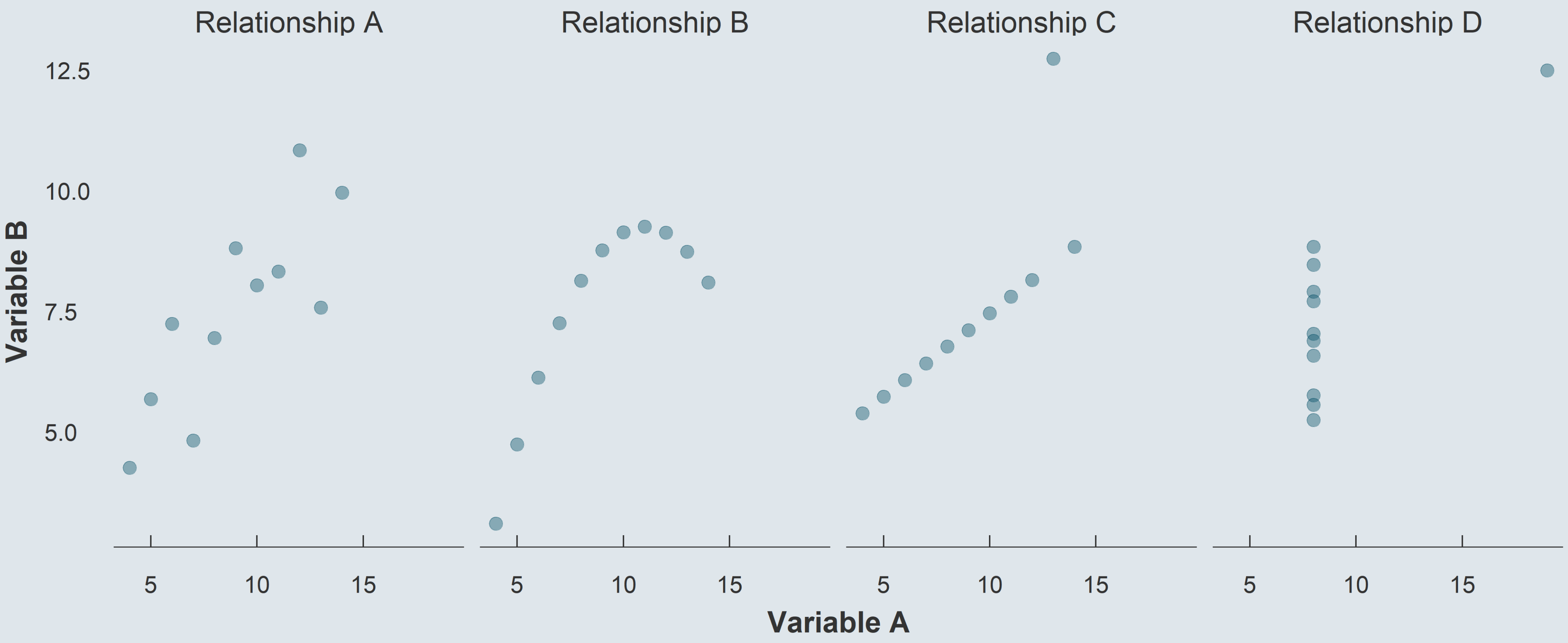
1. Joint distributions

1.3. Correlation



1. Joint distributions

➜ But a same correlation can hide very different relationships



1. Joint distributions

→ Covariance and correlation in R

```r
x <- c(50, 70, 60, 80, 60)
y <- c(10, 30, 20, 30, 40)
```

- The **covariance** can be obtain with the function `cov()`

```r
cov(x, y)
```

```
## [1] 70
```

- The **correlation** can be obtain with the function `cor()`

```r
cor(x, y)
```

```
## [1] 0.5384615
```

Overview

1. Joint distributions ✓

- 1.1. Definition
- 1.2. Covariance
- 1.3. Correlation

2. Univariate regressions

- 2.1. Introduction to regressions
- 2.2. Coefficients estimation

3. Binary variables

- 3.1. Binary dependent variables
- 3.2. Binary independent variables

4. Wrap up!

Overview

2. Univariate regressions

2.1. Introduction to regressions

- Consider the following dataset

```
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

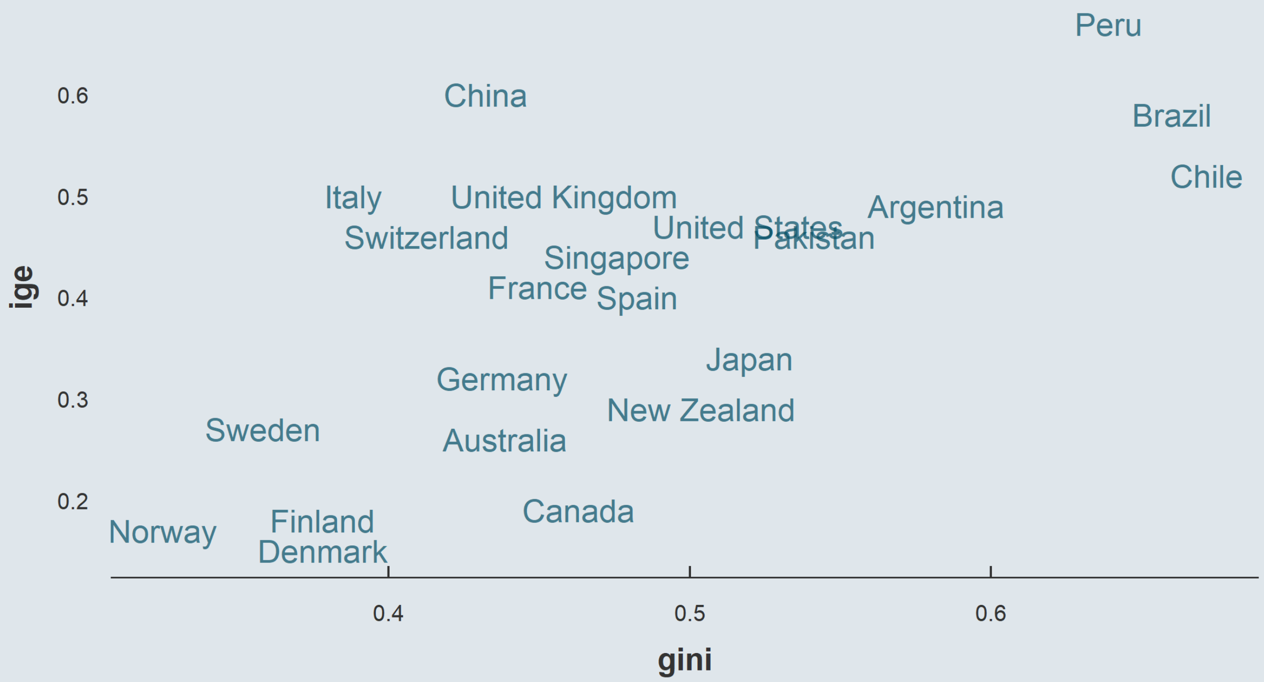
2. **Gini:** Gini index of income inequality between
   0: everybody has the same income
   1: a single individual has all the income

2. Univariate regressions

2.1. 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()
```



2. Univariate regressions

2.1. 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
```

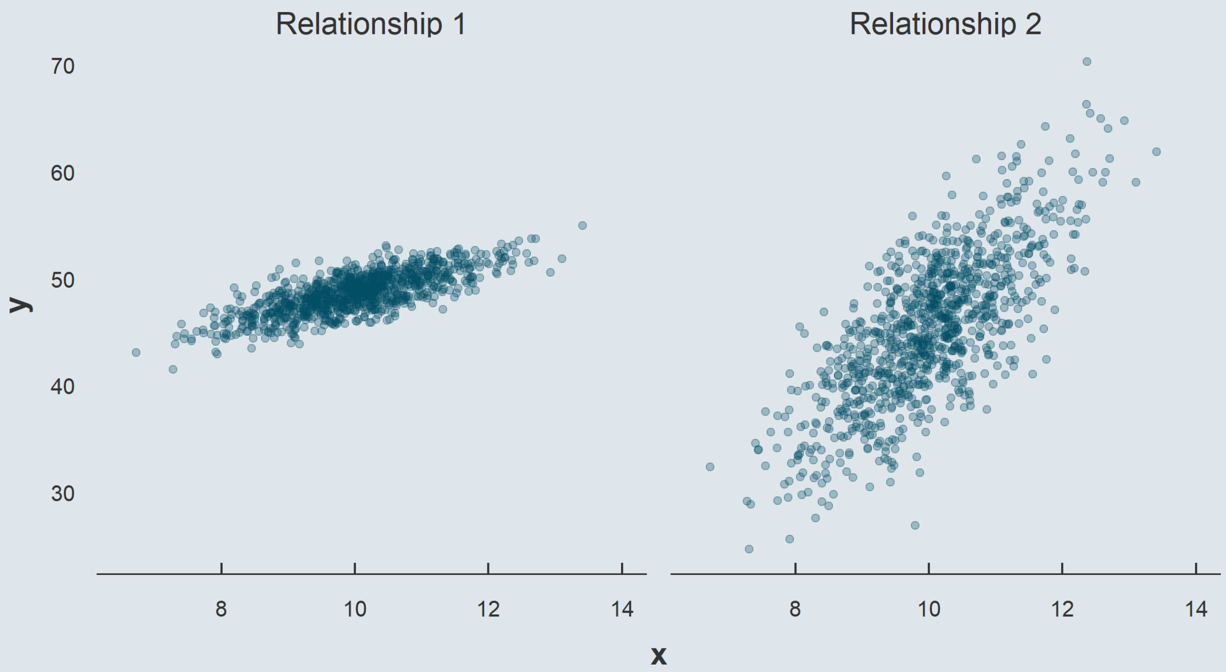
- The **correlation** coefficient is equal to **.65**
    - Remember that the correlation can take values from -1 to 1
    - Here the correlation is indeed **positive** and **fairly strong**

- But how useful is this for real-life applications? We may want more **practical** information:
    - Like by how much y is **expected** to **increas**e for a given change in x
    - This is of particular interest for economists and **policy** makers

2. Univariate regressions

2.1. Introduction to regressions

- Consider these two relationships :



→ One is less noisy but flatter

→ One is noisier but steeper

Both have a correlation of .75

2. Univariate regressions

2.1. Introduction to regressions

- Consider these two relationships :

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

2. Univariate regressions

2.1. Introduction to regressions

- Knowing that income inequality is **negatively correlated** with intergenerational mobility is one thing

- But how much more intergenerational mobility could we expect for a given reduction in inequality?
  - We need to know to characterize the ***"steepness"*** of the relationship!

- It is usually the **type of questions** 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?*

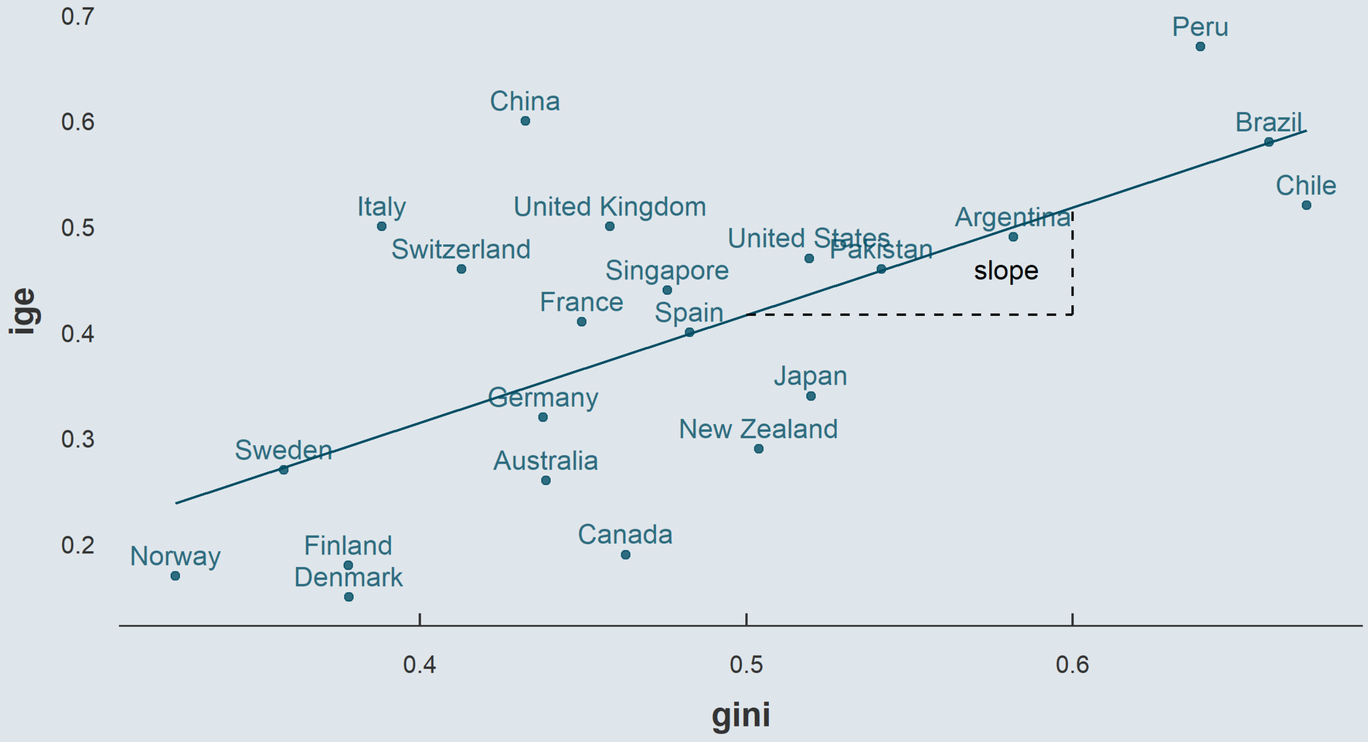
- And once again, this is typically what is of interest for **policymakers**

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

2. Univariate regressions

2.2. Coefficients estimation

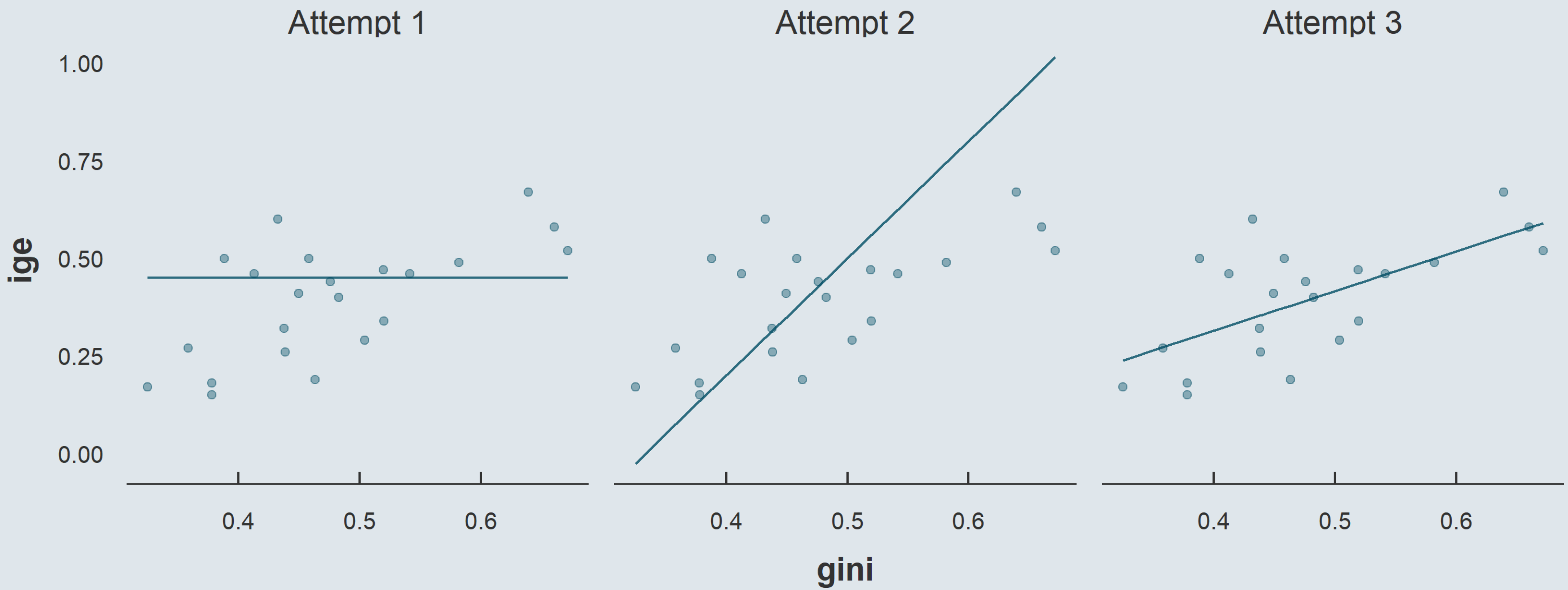
- The idea is to find the **line that fits the data** the best
  - Such that its **slope** can indicate how we **expect y to change** if we **increase x by 1** unit



2. Univariate regressions

2.2. Coefficients estimation

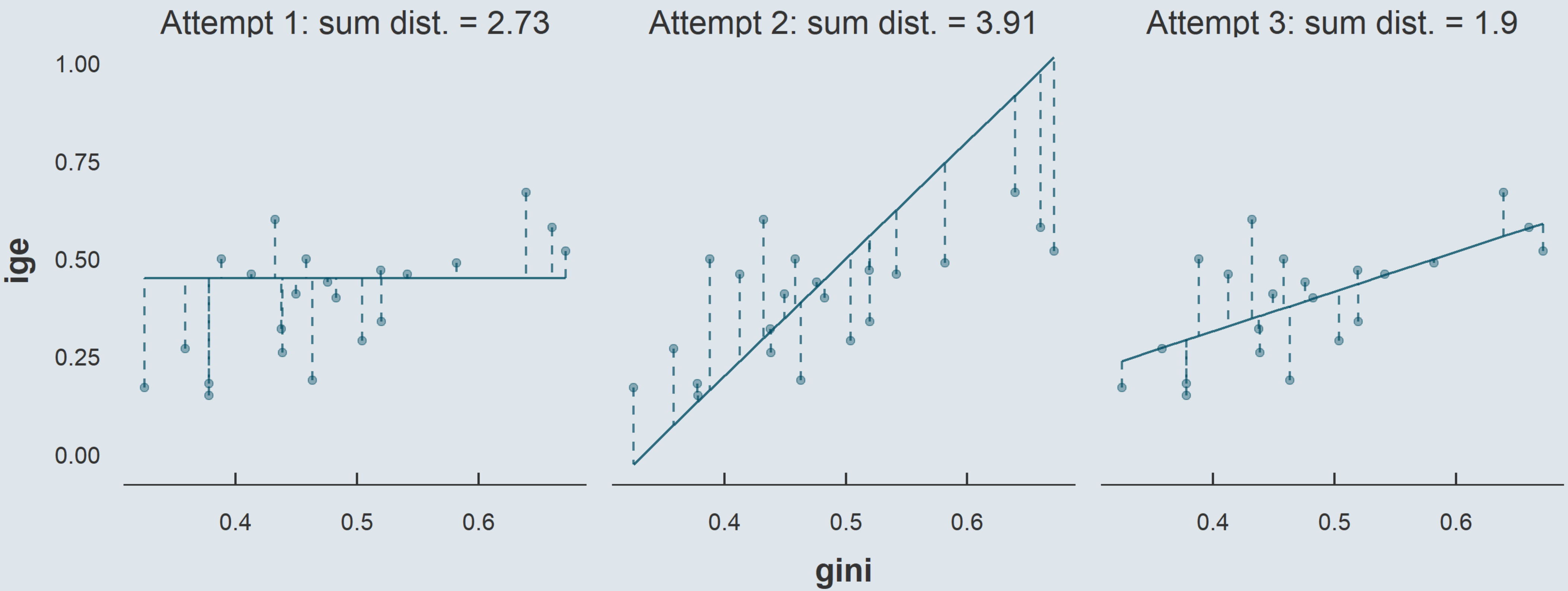
- But how do we **find that line?**



2. Univariate regressions

2.2. Coefficients estimation

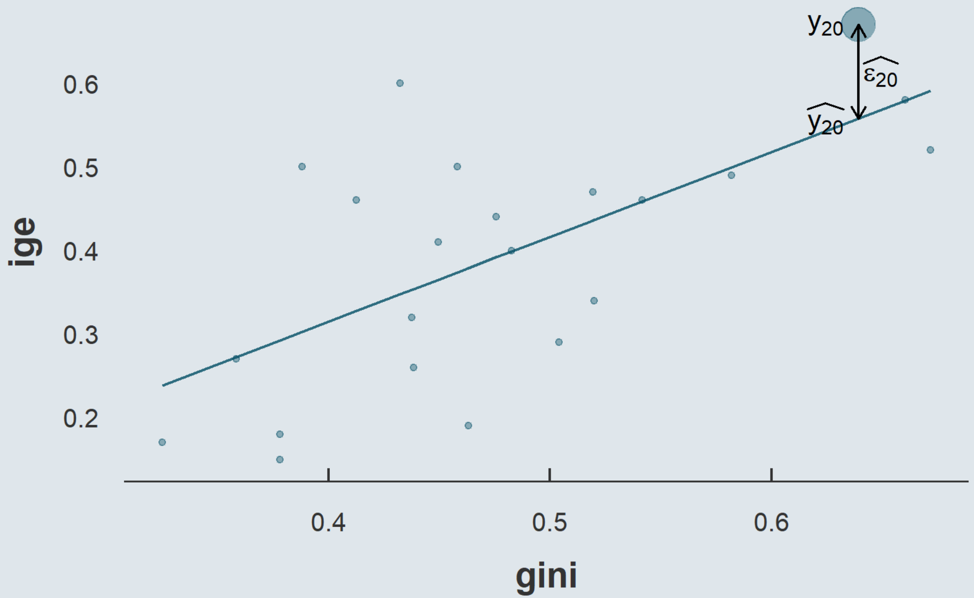
- We try to **minimize the distance** between each point and our line



2. Univariate regressions

2.2. 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 for $x = x_i$

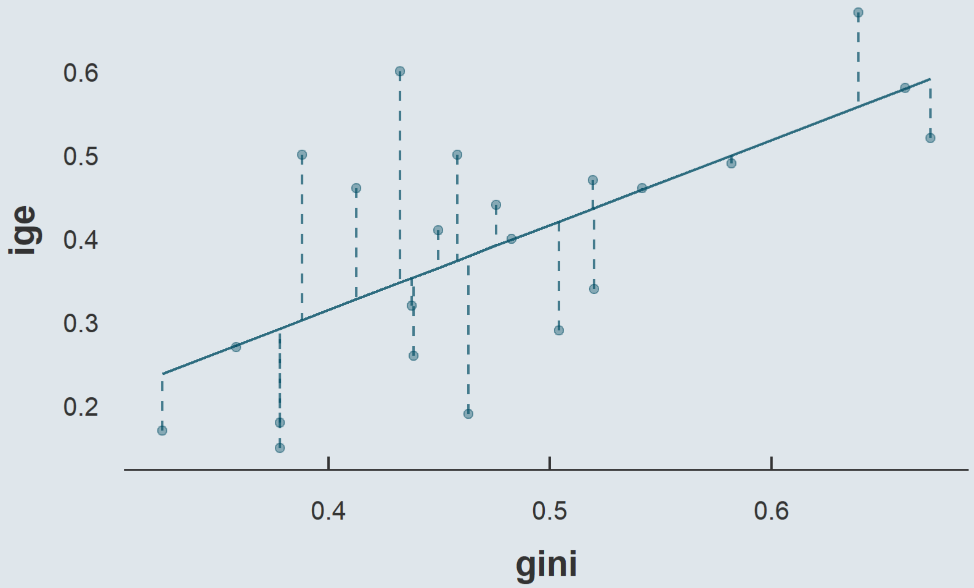
→ The distance between the i^{th} y value and the line is

$$y_i - \widehat{y_i}$$

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

2. Univariate regressions

2.2. Coefficients estimation



- $\widehat{\varepsilon_i}$ being 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 **intercept**
  - $\hat{\beta}$ is the **slope**

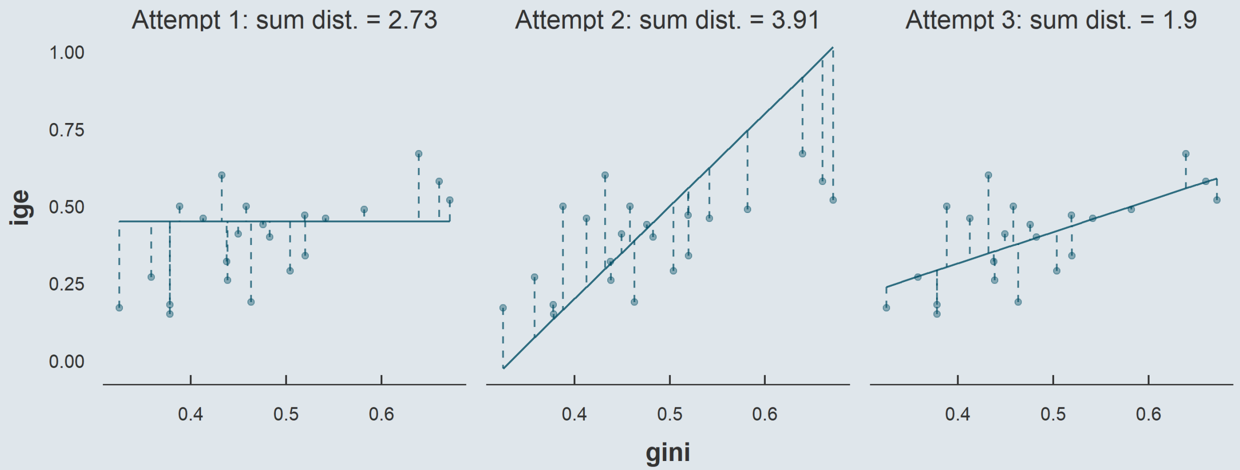
2. Univariate regressions

2.2. 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: both $\hat{\alpha}$ and $\hat{\beta}$ seem right ➜ $\widehat{\varepsilon_i}$ are low

2. Univariate regressions

2.2. 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 $\hat{\beta}$

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

$$\iff$$

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

2. Univariate regressions

2.2. Coefficients estimation

- And our minimization problem writes

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

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

- Rearranging the first equation yields

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

2. Univariate regressions

2.2. 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:

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

Practice

1) Import `ggcurve.csv` and compute the $\hat{\alpha}$ and $\hat{\beta}$ coefficients of that equation:

$$\text{gini}_i = \hat{\alpha} + \hat{\beta} \times \text{IGE}_i + \widehat{\varepsilon_i}$$

2) Create a new variable in the dataset for $\widehat{\text{gini}}$

3) Plot your results (scatter plot + line)

Hints: You can use different y variables for different geometries by specifying the mapping within the geometry function: geom_point(aes(y = y))

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

You've got 10 minutes!

Solution

1) Import `ggcurve.csv` and compute the $\hat{\alpha}$ and $\hat{\beta}$ coefficients of that equation:

```
# Read the data
ggcurve <- read.csv("ggcurve.csv")
# Compute beta
beta <- cov(ggcurve$gini, ggcurve$ige) / var(ggcurve$gini)
# Compute alpha
alpha <- mean(ggcurve$ige) - (beta * mean(ggcurve$gini))
```

```
c(alpha, beta)
```

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

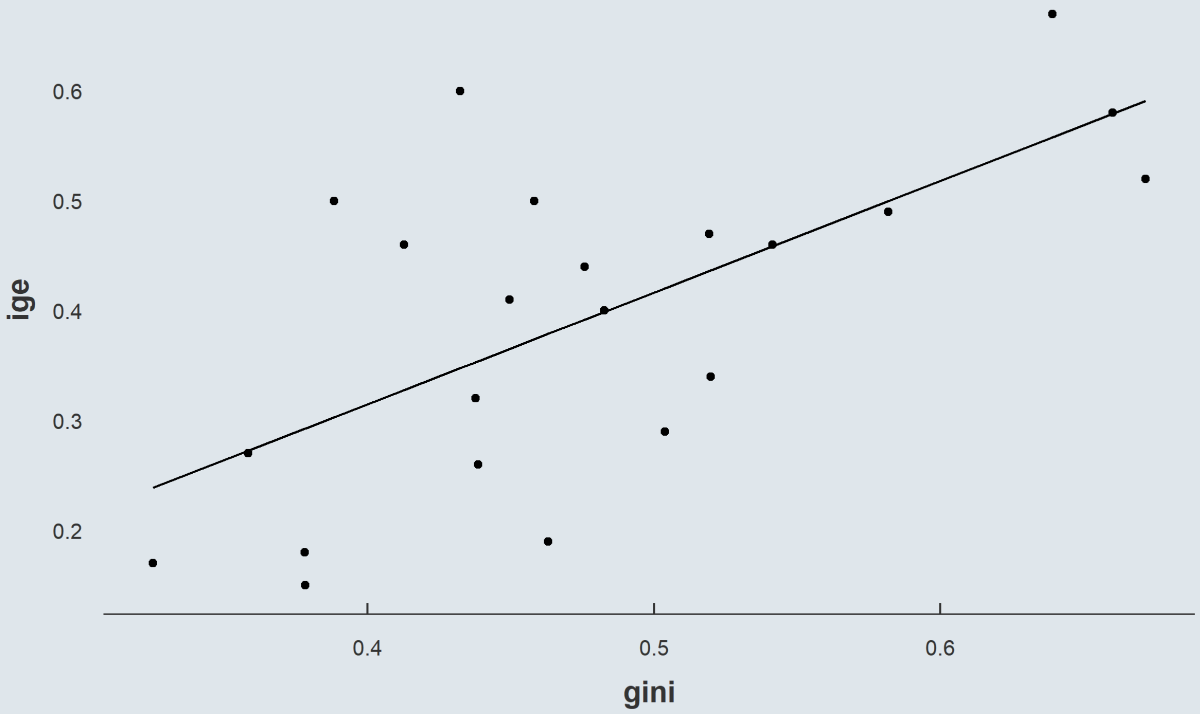
2) Create a new variable in the dataset for $\widehat{\text{gini}}$

```
ggcurve <- ggcurve %>%
  mutate(fit = alpha + beta * gini)
```

Solution

3) Plot your results (scatter plot + line)

```
ggplot(ggcurve, aes(x = gini)) +
  geom_point(aes(y = ige)) + geom_line(aes(y = fit))
```



2. Univariate regressions

2.2. Coefficients estimation

- As usual there are **functions** to do that **in R**

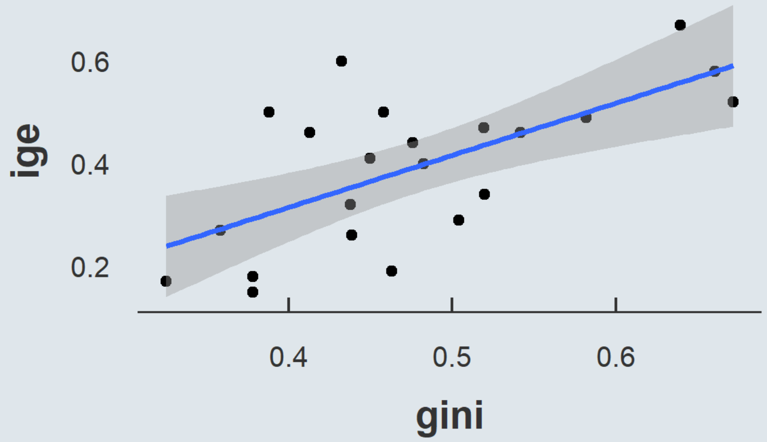
- **lm()** to estimate regression coefficients
- It has two main **arguments:**
  - **Formula:** written as **y ~ x**
  - **Data:** where y and x are

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

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

- **geom_smooth()** to plot the fit

```
ggplot(ggcurve, aes(x = gini, y = ige)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x)
```



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 we tune them 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**

Overview

1. Joint distributions ✓

- 1.1. Definition
- 1.2. Covariance
- 1.3. Correlation

2. Univariate regressions ✓

- 2.1. Introduction to regressions
- 2.2. Coefficients estimation

3. Binary variables

- 3.1. Binary dependent variables
- 3.2. Binary independent variables

4. Wrap up!

Overview

1. Joint distributions ✓

- 1.1. Definition
- 1.2. Covariance
- 1.3. Correlation

2. Univariate regressions ✓

- 2.1. Introduction to regressions
- 2.2. Coefficients estimation

3. Binary variables

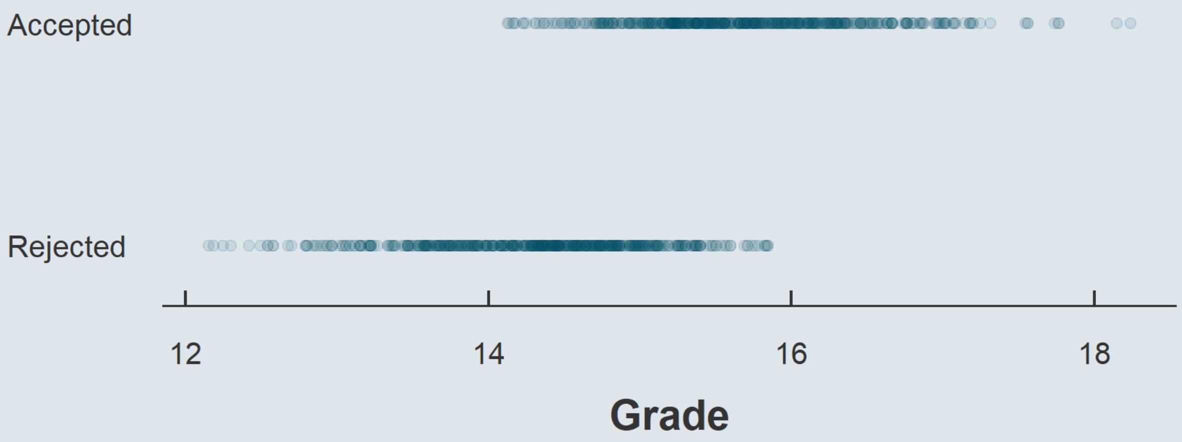
- 3.1. Binary dependent variables
- 3.2. Binary independent variables

3. Binary variables

3.1. Binary dependent variables

- **So far** we've considered only **continuous variables** in our regression models
  - But what if our **dependent** variable is **discrete?**

- Consider that we have data on candidates to a job:
  - Their *Baccalauréat* grade (/20)
  - Whether they got accepted

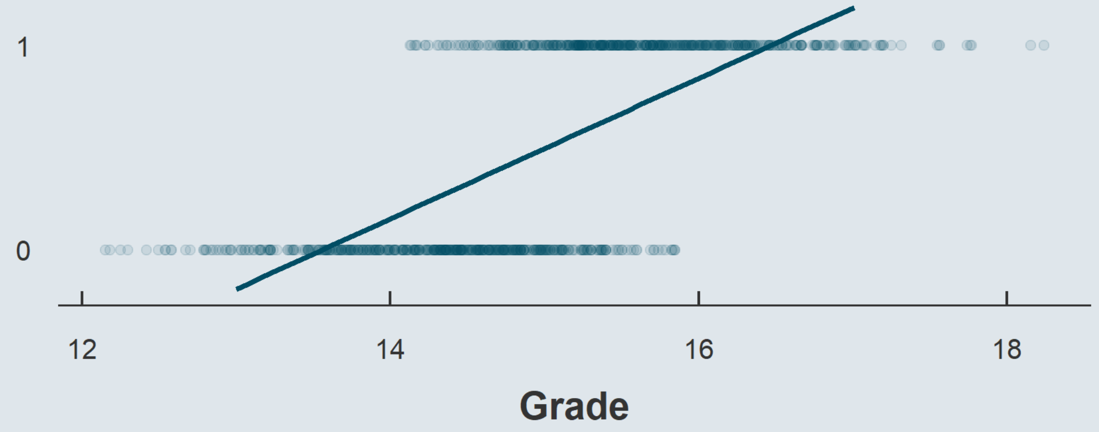


3. Binary variables

3.1. Binary dependent variables

- Even if the **outcome variable** is binary we can regress it on the grade variable
  - We can convert it into a **dummy** variable, a variable taking either the value **0 or 1**
  - Here consider a dummy variable taking the value 1 if the person was accepted

$$1\{y_i = \text{Accepted}\} = \hat{\alpha} + \hat{\beta} \times \text{Grade}_i + \hat{\varepsilon}_i$$

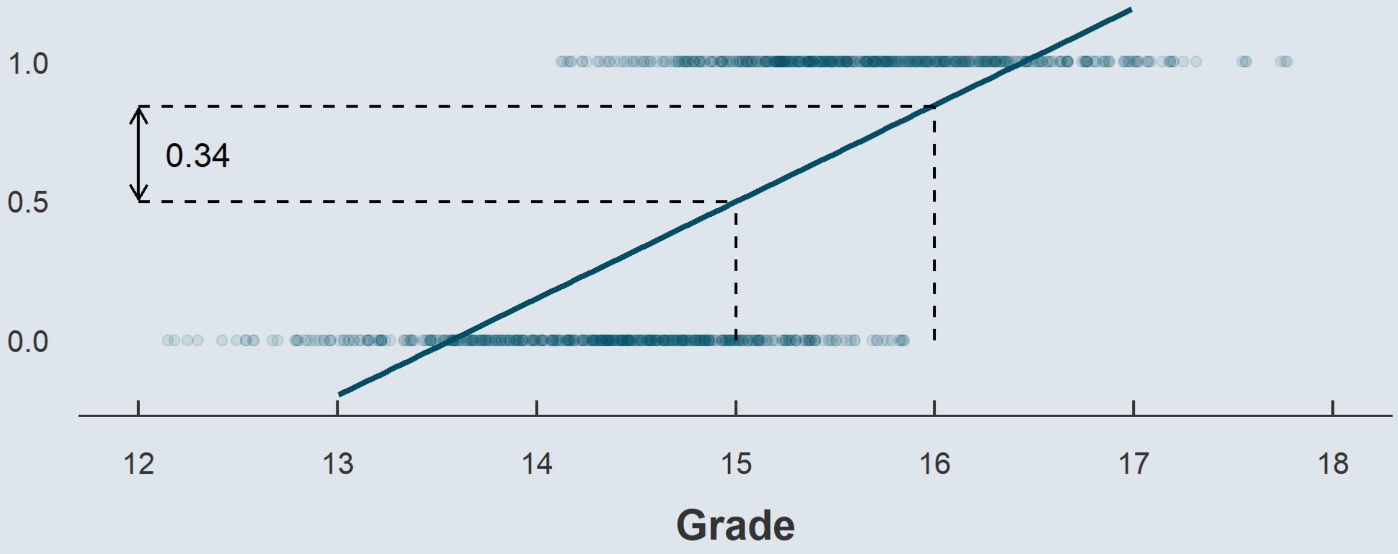


➜ ***How would you interpret the beta coefficient from this regression?***

3. Binary variables

3.1. Binary dependent variables

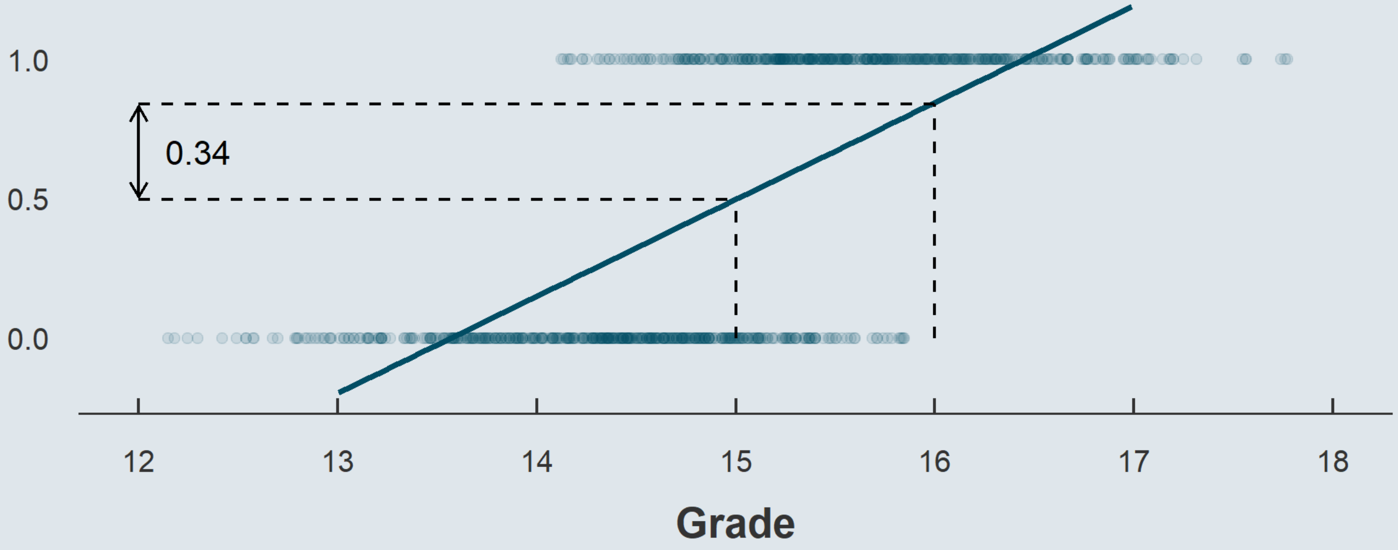
- The **fitted values** can be viewed as the **probability** to be accepted for a given grade
    - $\hat{\beta}$ is thus by how much this probability would vary on expectation for a 1 point increase in the grade
    - That's why we call OLS regression models with a binary outcome **Linear *Probability* Models**



3. Binary variables

3.1. Binary dependent variables

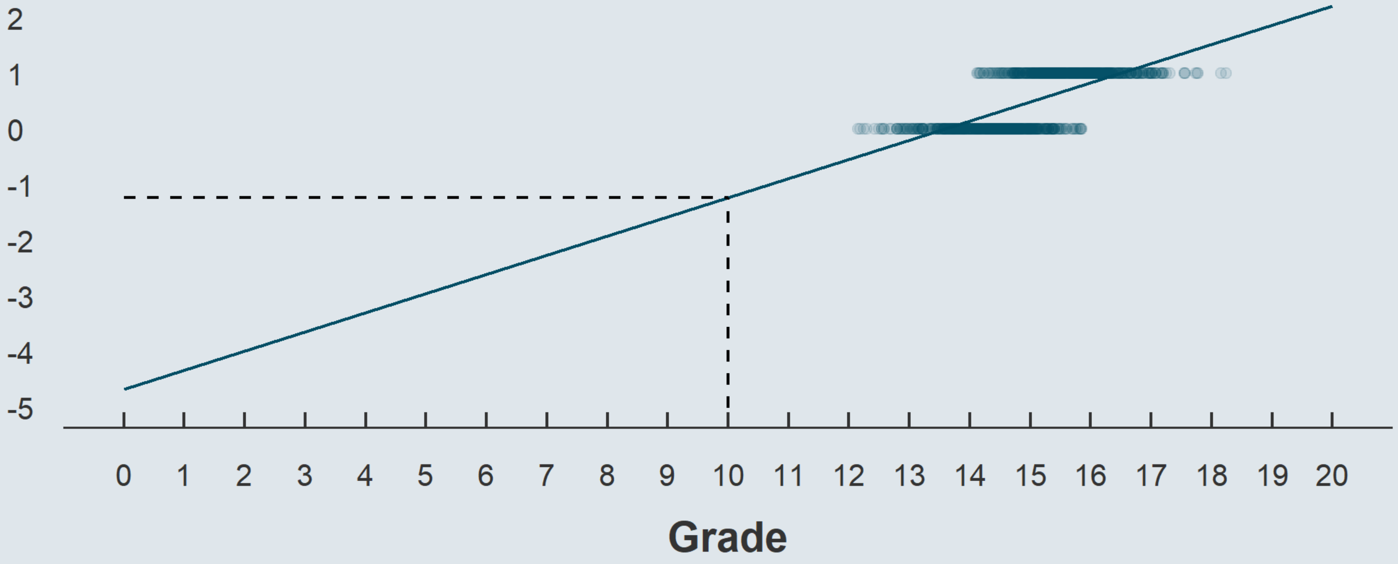
- But what kind of **problems** could we encounter with **such models?**
  - What would be the $\hat{\alpha}$ coefficient here?
  - And what's the probability to be accepted for a grade of 18?



3. Binary variables

3.1. Binary dependent variables

- With an **LPM** you can end up with ***"probabilities"*** that are **lower than 0** and **greater than 1**
  - **Interpretation** is only **valid** for values of x sufficiently **close to the mean**
  - Keep that in mind and be **careful** when interpreting the results of an LPM



3. Binary variables

3.2. Binary independent variables

- Now consider that we have individual **data** containing
  - The **sex**
  - The **height** (centimeters)

- So the situation is different
  - We used to have a **binary dependent variable:**

$$1\{y_i = \text{Accepted}\} = \hat{\alpha} + \hat{\beta} \times \text{Grade}_i + \hat{\varepsilon}_i$$

  - We now have a **binary independent variable:**

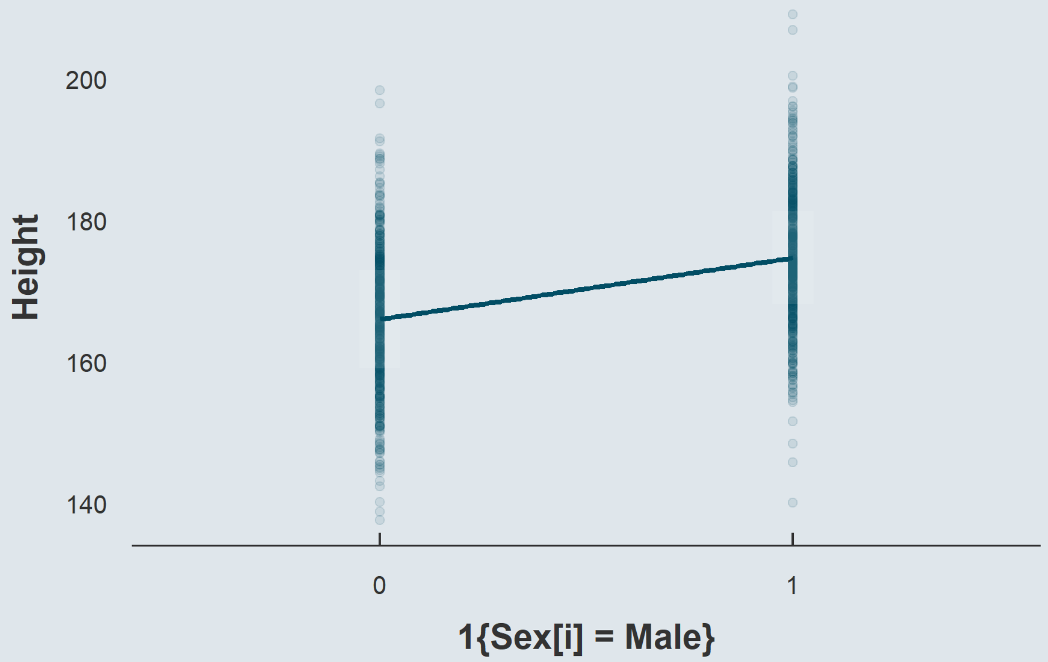
$$\text{Height}_i = \hat{\alpha} + \hat{\beta} \times 1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

➜ ***How would you interpret the coefficient $\hat{\beta}$ from this regression?***

3. Binary variables

3.2. Binary independent variables

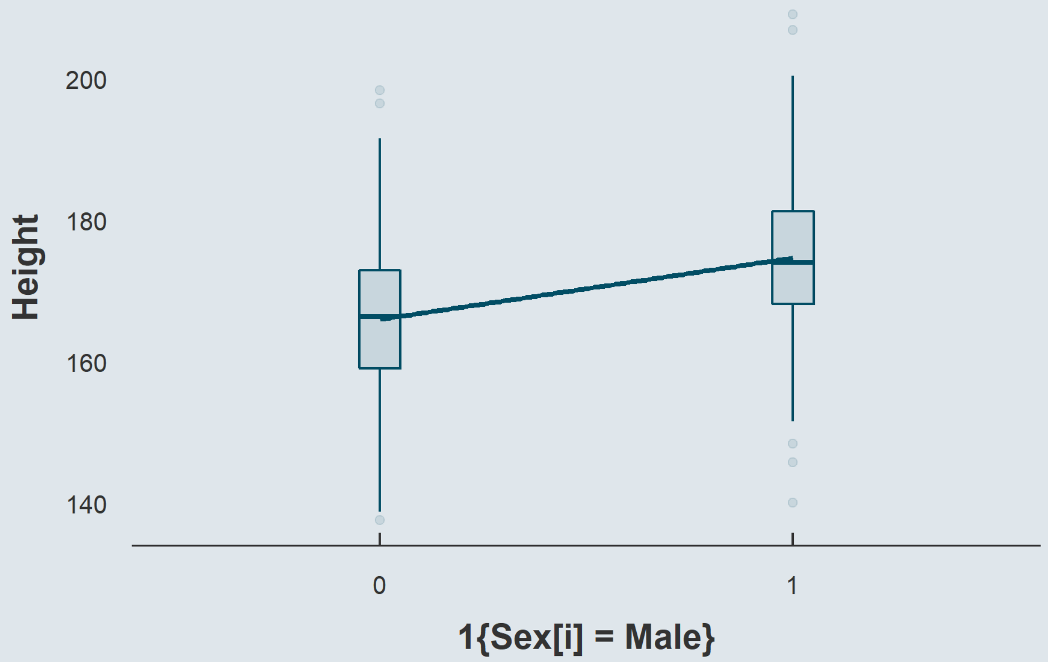
- If the sex variable was **continuous** it would be the expected increase in height for a ***"1 unit increase"*** in sex
  - Here the ***"1 unit increase"*** is switching from 0 to 1, i.e. **from female to male**
  - With that in mind, how would you interpret the coefficient $\hat{\beta}$?



3. Binary variables

3.2. Binary independent variables

- If I replace the point geometry by the corresponding **boxplots**
  - What this ***"1 unit increase"*** corresponds to should be **clearer**
  - The coefficient $\hat{\beta}$ is actually the **difference** between the **average height** for males and females



3. Binary variables

3.2. Binary independent variables

$$\overline{\text{Height}}_{[\text{Sex}_i=\text{Female}]} = 165$$

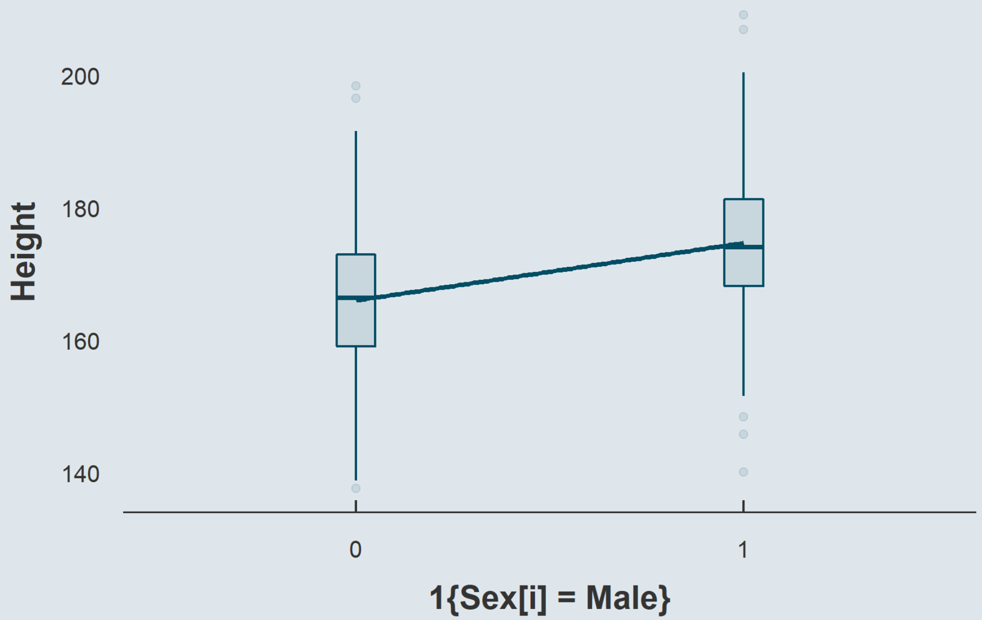
$$\overline{\text{Height}}_{[\text{Sex}_i=\text{Male}]} = 176$$

$$\text{Height}_i = \hat{\alpha} + \hat{\beta} \times 1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

$$\hat{\alpha} = 165 \qquad \hat{\beta} = 11$$

$$\text{Height}_i = \hat{\alpha} + \hat{\beta} \times 1\{\text{Sex}_i = \text{Female}\} + \hat{\varepsilon}_i$$

$$\hat{\alpha} = 176 \qquad \hat{\beta} = -11$$



3. Binary variables

3.2. Binary independent variables

- In terms of **fitted values:**

$$\text{Height}_i = \hat{\alpha} + \hat{\beta} \times 1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

- We now have $\hat{\alpha}$ and $\hat{\beta}$:

$$\text{Height}_i = 165 + 11 \times 1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

- The fitted values write:

$$\widehat{\text{Height}}_i = 165 + 11 \times 1\{\text{Sex}_i = \text{Male}\}$$

- When the dummy equals 0 *(females)*:

$$\begin{aligned} \widehat{\text{Height}}_i &= 165 + 11 \times 0 \\ &= 165 = \overline{\text{Height}_{[\text{Sex}_i = \text{Female}]}} \end{aligned}$$

- When the dummy equals 1 *(males)*:

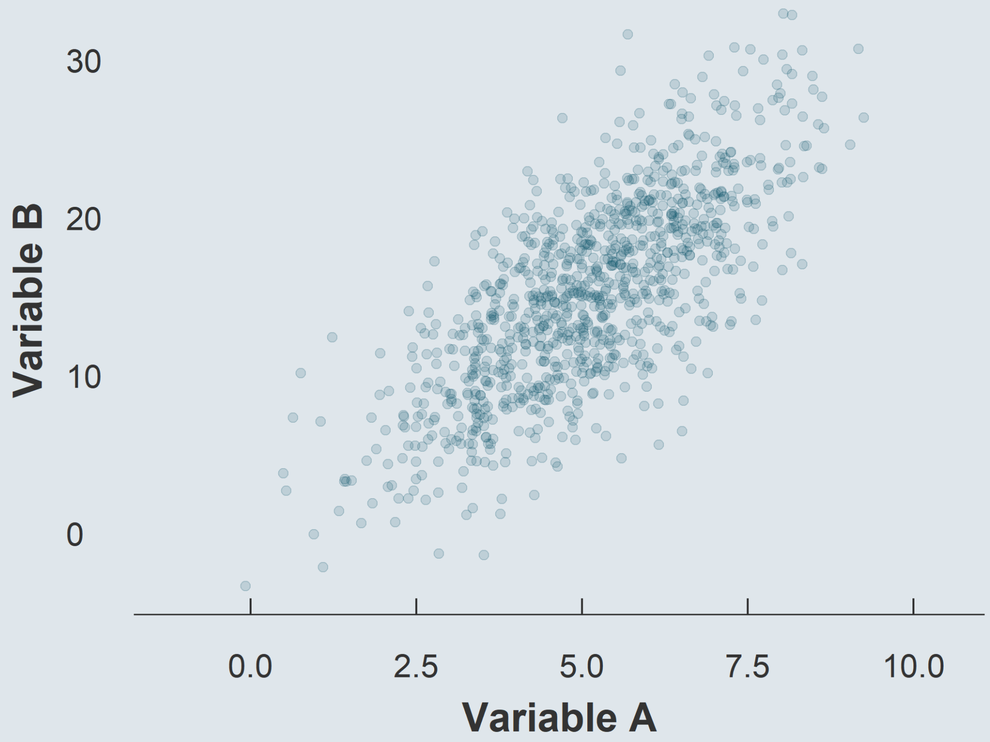
$$\begin{aligned} \widehat{\text{Height}}_i &= 165 + 11 \times 1 \\ &= 176 = \overline{\text{Height}_{[\text{Sex}_i = \text{Male}]}} \end{aligned}$$

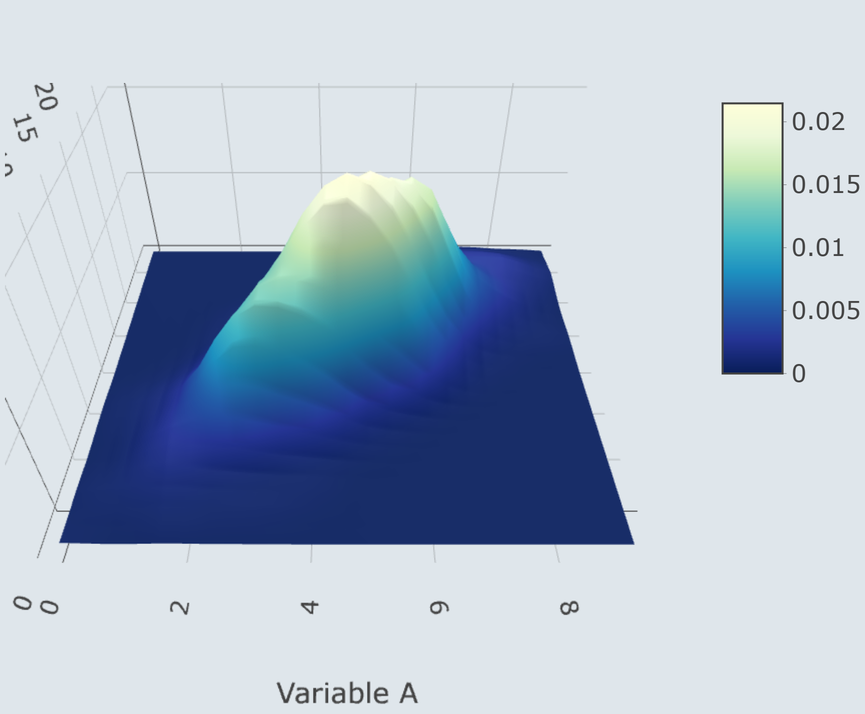
Overview

4. Wrap up!

1. Joint distribution

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





4. Wrap up!

1. 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
  - It can take values from $-\infty$ to ∞ and depends on the unit of the data

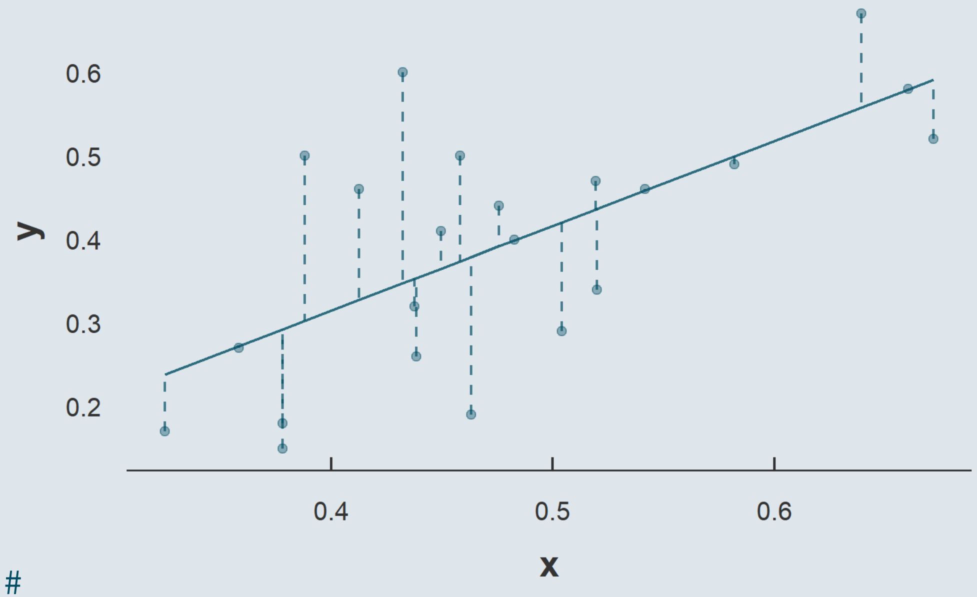
$$\text{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- The **correlation** is the covariance of two variables divided by the product of their standard deviation
  - It can take values from -1 to 1 and is independent from the unit of the data

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

4. Wrap up!

2. Regression




```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Coefficients:
## (Intercept)            x
##    -0.09129      1.01546
```

- 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

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

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

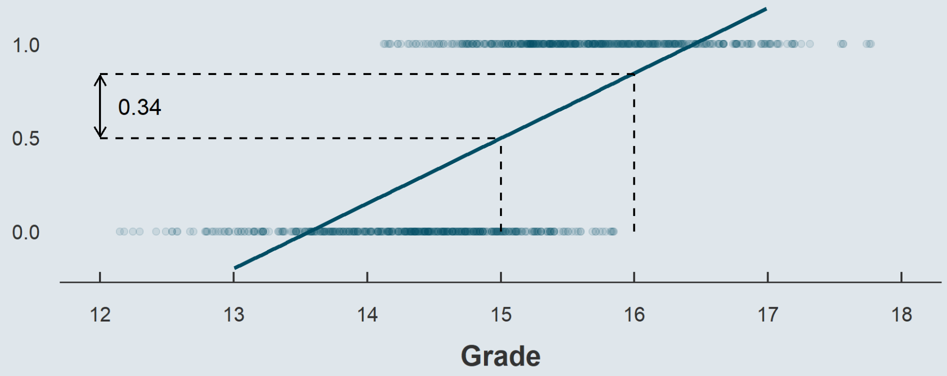
- $\hat{\alpha}$ and $\hat{\beta}$ minimize $\hat{\varepsilon}_i$

4. Wrap up!

3. Binary variables

Binary **dependent** variables

- The **fitted values** can be viewed as **probabilities**
  - $\hat{\beta}$ is the expected increase in the probability that $y = 1$ for a one unit increase in x



  - We call that a **Linear Probability model**

Binary **independent** variables

- The x variable should be viewed as a **dummy 0/1**
  - $\hat{\beta}$ is the difference between the average y for the group $x = 1$ and the group $x = 0$