Introductory Econometrics

Lecture 18

Louis SIRUGUE

CPES 2 - Spring 2023

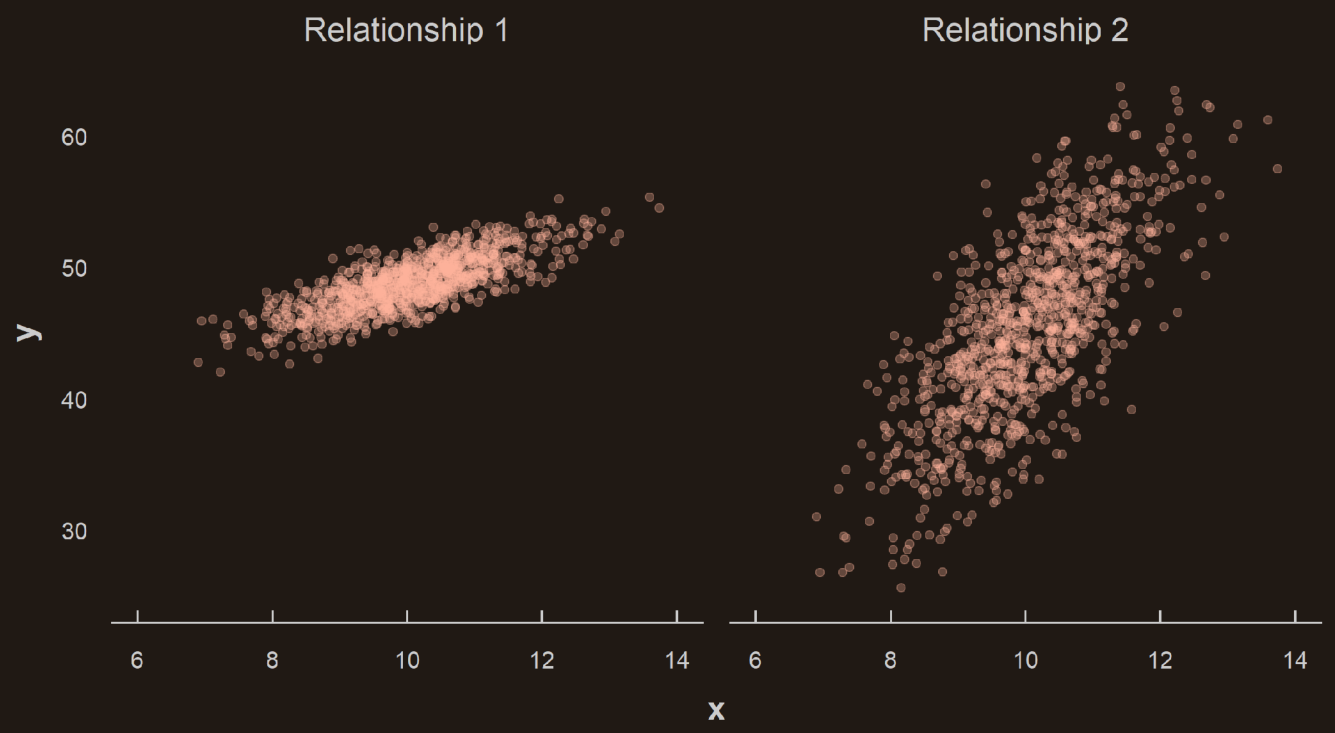
Today: Refresher on Introductory Econometrics

Today: Refresher on Introductory Econometrics

1. Regressions with continuous variables

1.1. Estimation

- Consider these two relationships:



→ One is less noisy but flatter

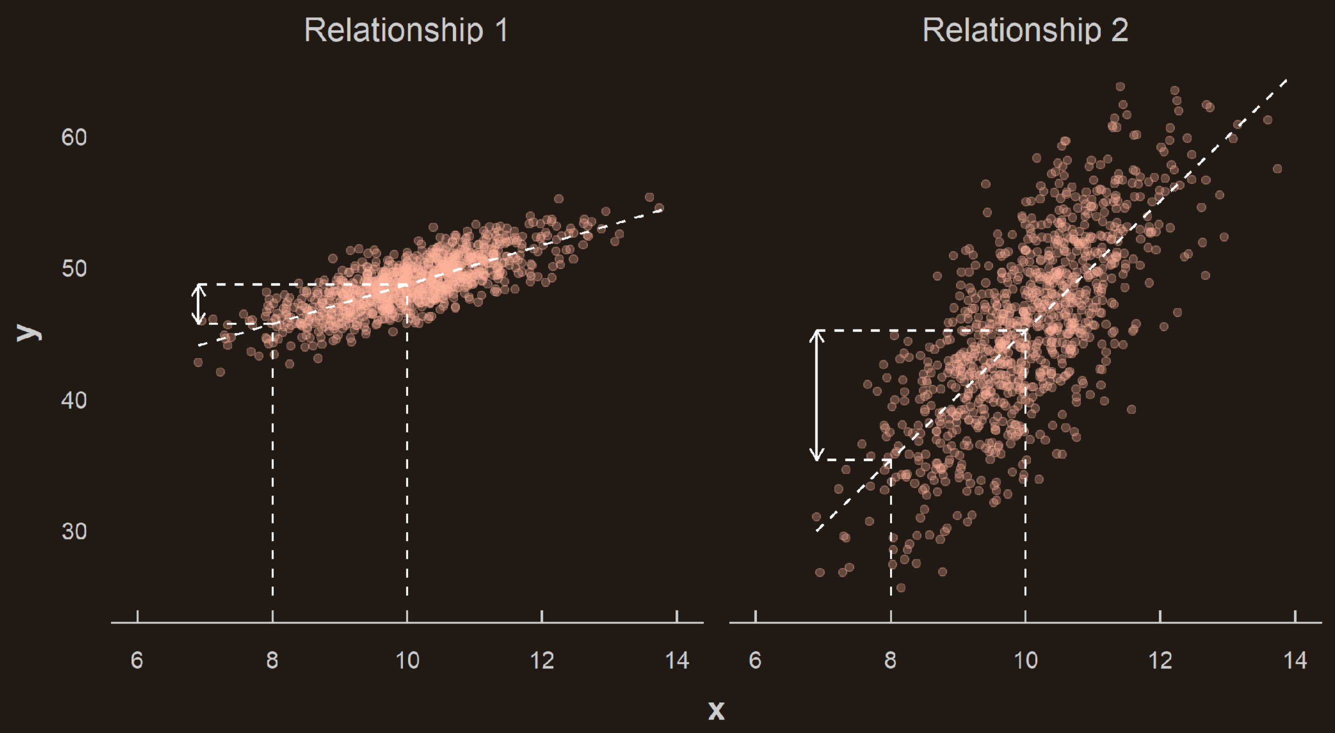
→ One is noisier but steeper

Both have a correlation of .75

1. Regressions with continuous variables

1.1. Estimation

- Consider these two relationships:

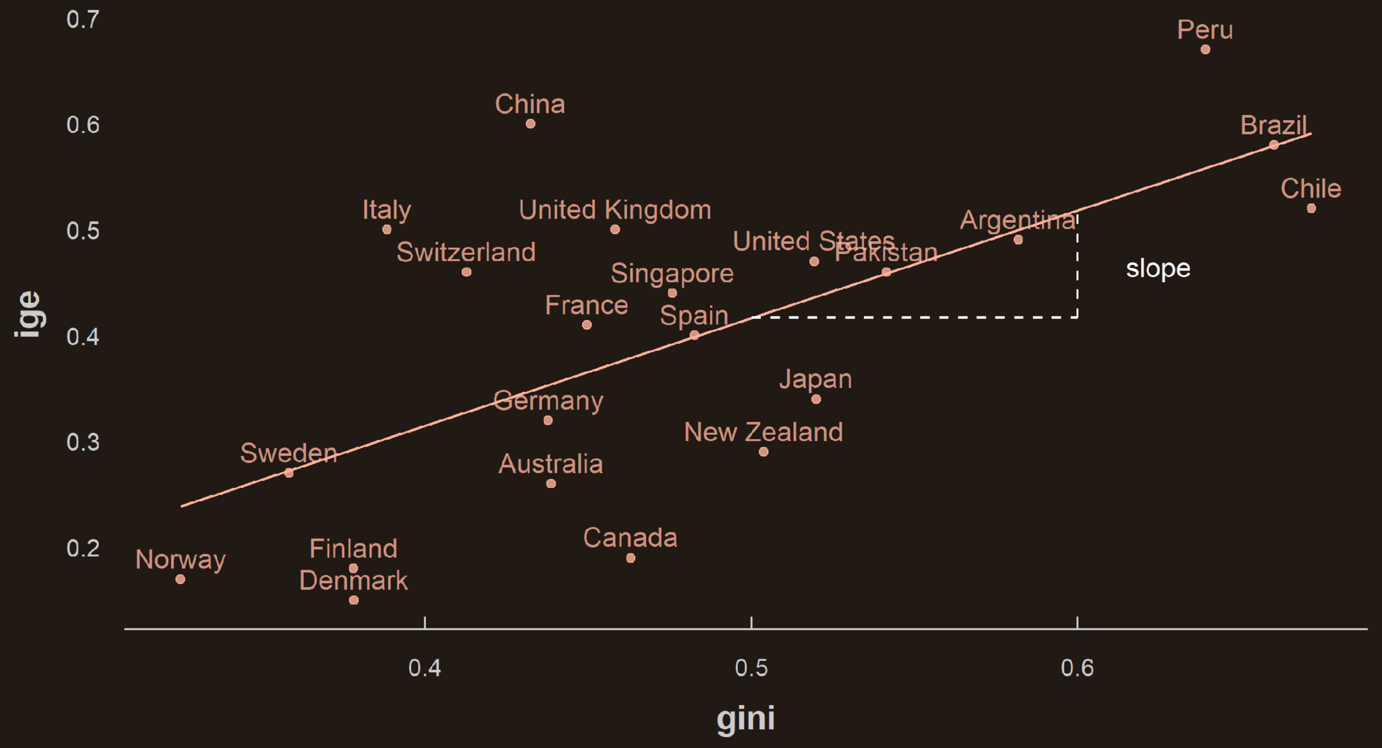


But a given increase in x is not associated with a same increase in y!

1. Regressions with continuous variables

1.1. Estimation

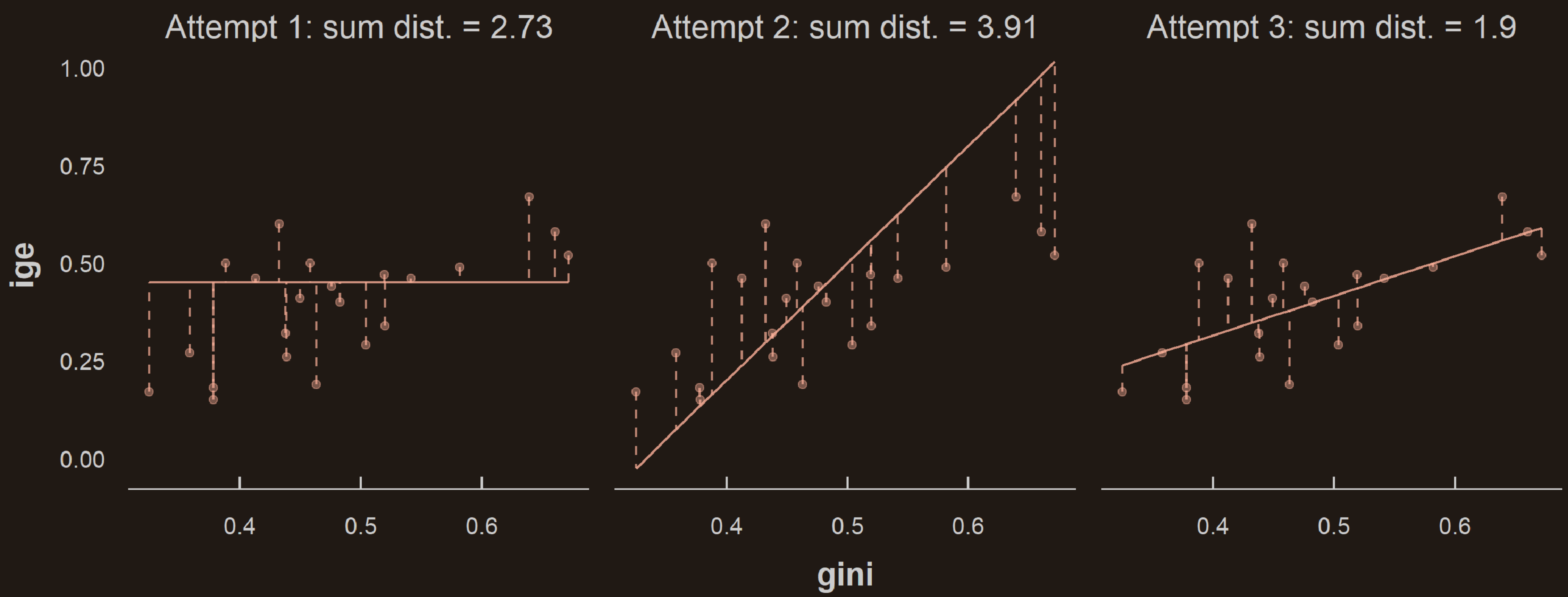
- The idea of a regression is to find the **line** that **fits** the data the **best**
  - Such that its slope can indicate **how we expect y to change if we increase x by 1 unit**



1. Regressions with continuous variables

1.1. Estimation

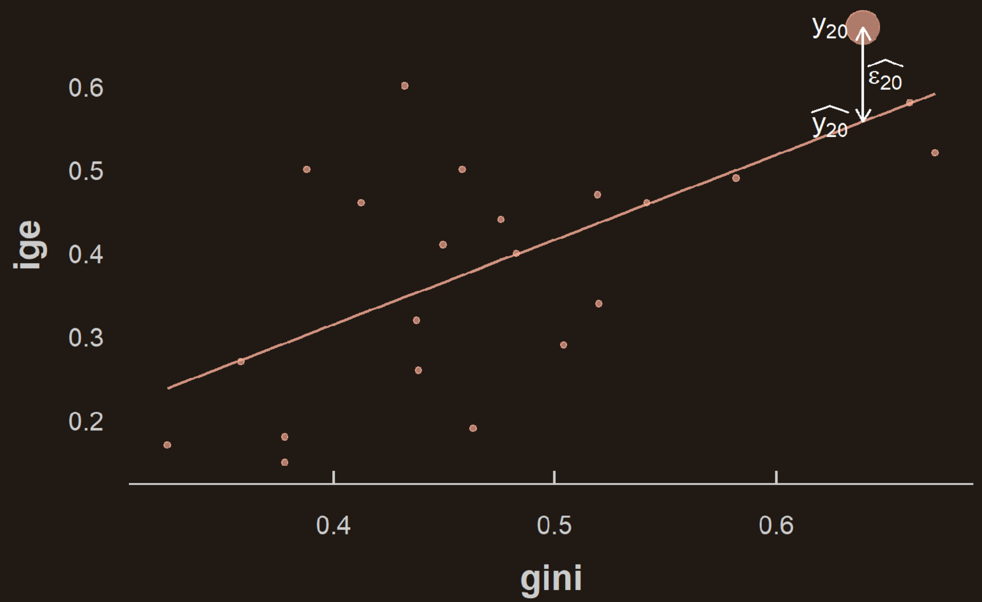
- To do so we should **minimize the distance** between each **point** and the **line**



1. Regressions with continuous variables

1.1. Estimation

Take for instance the 20th observation: Peru



And consider the following notations:

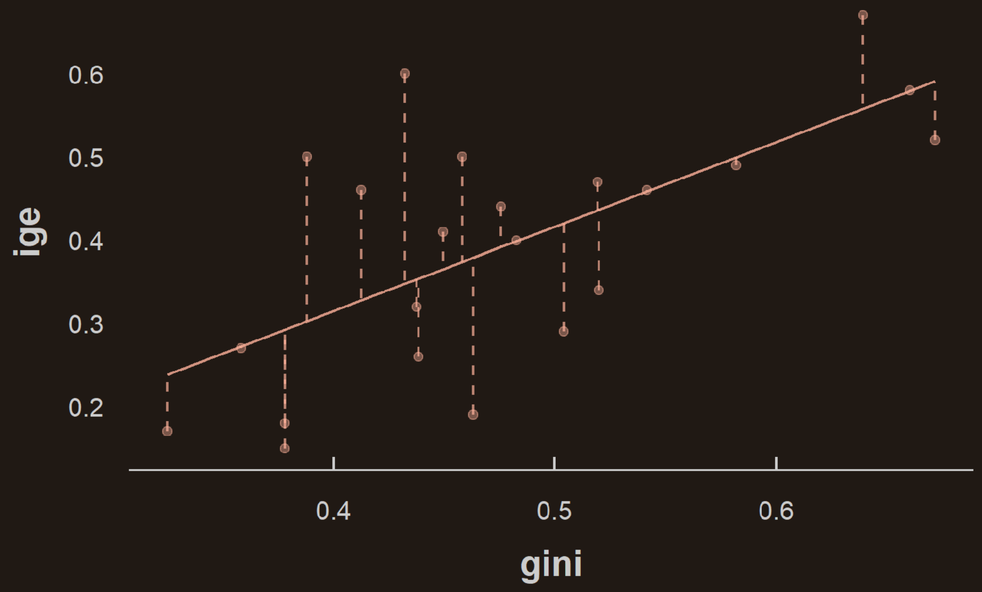
- We denote y_i the ige of the i^{th} country
- We denote x_i the gini of the i^{th} country
- We denote $\widehat{y_i}$ the value of the y coordinate of our line when $x = x_i$

→ 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. Regressions with continuous variables

1.1. 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
  - Both are estimations of the actual α and β of the unknown DGP

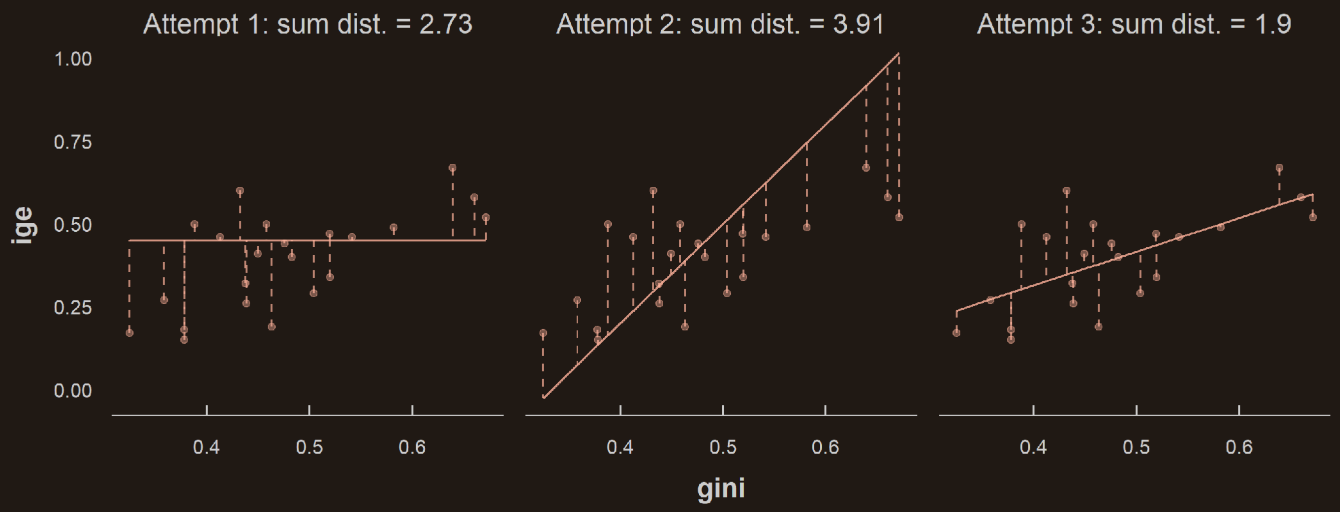
1. Regressions with continuous variables

1.1. 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. Regressions with continuous variables

1.1. 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)**

- If we replace $\widehat{\varepsilon_i}$ with $y_i - \hat{\alpha} - \hat{\beta}x_i$
  - We can solve the minimization problem (see Lecture 7) to obtain:

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

Vocabulary

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

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

  - 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. Regressions with continuous variables

1.2. Inference

- Inference refers to the fact of being able to **conclude** something from our estimation
  - The $\hat{\beta}$ from our sample is actually an **estimation** of the unobserved β of the underlying population
  - We would like to know how reliable $\hat{\beta}$ is, **how confident we are** in its estimation
  - The first step of inference is to compute the **standard error** of $\hat{\beta}$

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

- 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

1. Regressions with continuous variables

1.2. Inference

- 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

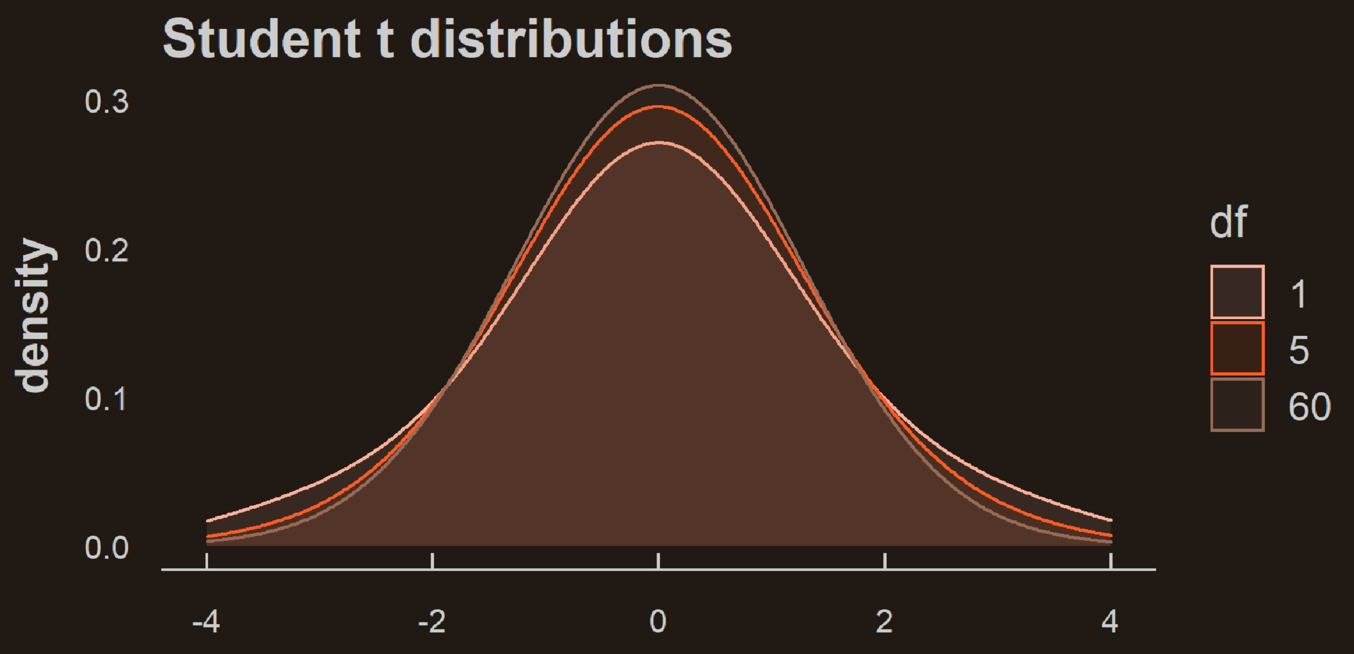
- To obtain a confidence interval we can use the fact that under specific conditions (that you're gonna see next year) it is possible to derive how this object is distributed:

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

1. Regressions with continuous variables

1.2. Inference

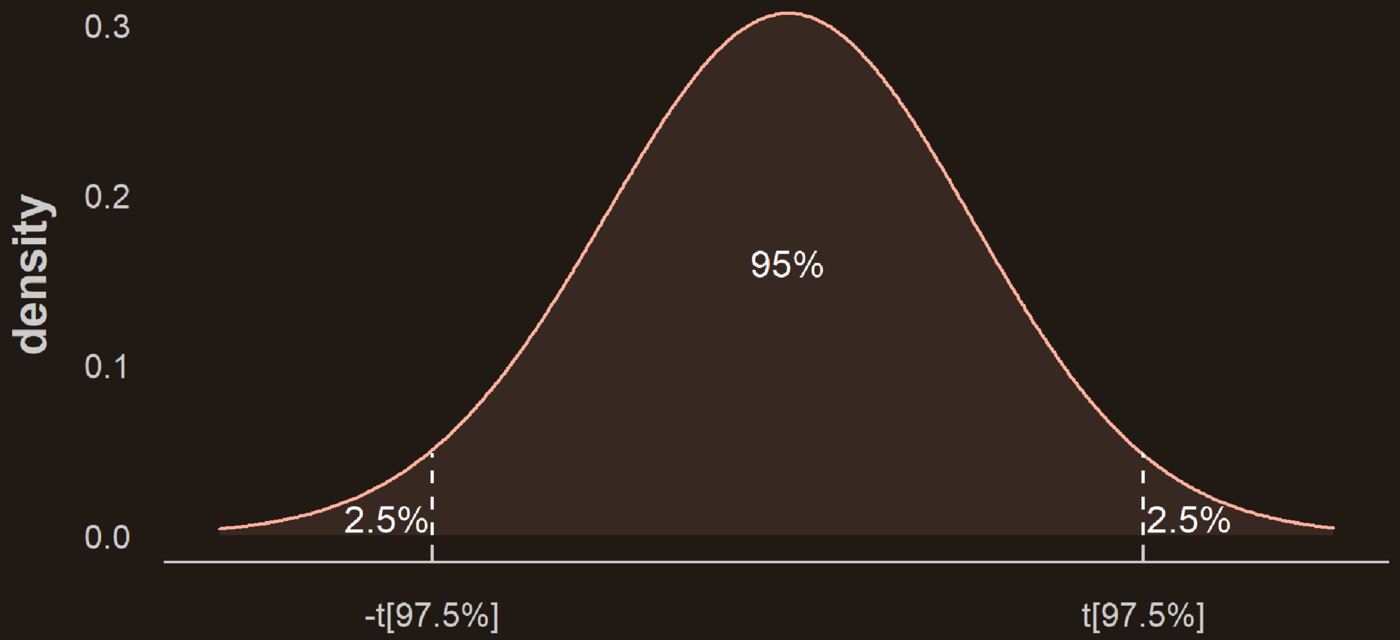
- Theory shows that $\hat{t} \equiv \frac{\hat{\beta}-\beta}{\text{se}(\hat{\beta})}$ follows a Student t distribution whose number of degrees of freedom is equal to n (in our case 22 countries) minus the number of parameters estimated in the model (in our case 2: α and β)



1. Regressions with continuous variables

1.2. Inference

- 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 $t_{97.5\%}$



1.2. Inference

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

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

- Rearranging the terms yields:

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

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

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

- To get $t_{97.5\%}$ with 20 df:

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

1. Regressions with continuous variables

1.2. Inference

- ***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 our estimation of the true parameter is different from a given value (generally 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 a given value, 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

1. Regressions with continuous variables

1.2. Inference

- **P-val. 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

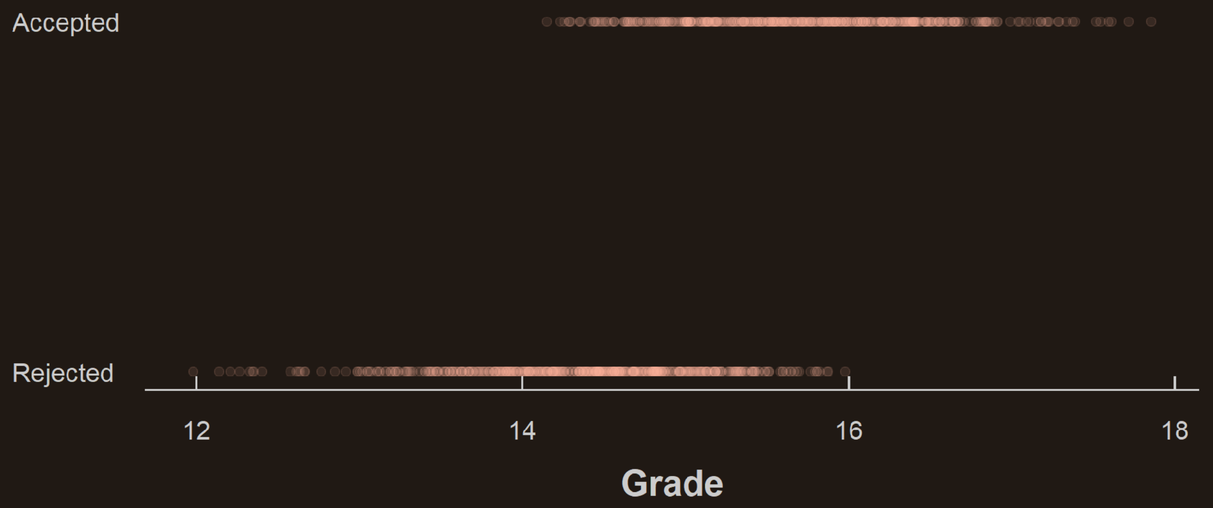
Overview

Overview

2. Regressions with discrete variables

2.1. Binary dependent variable

- 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

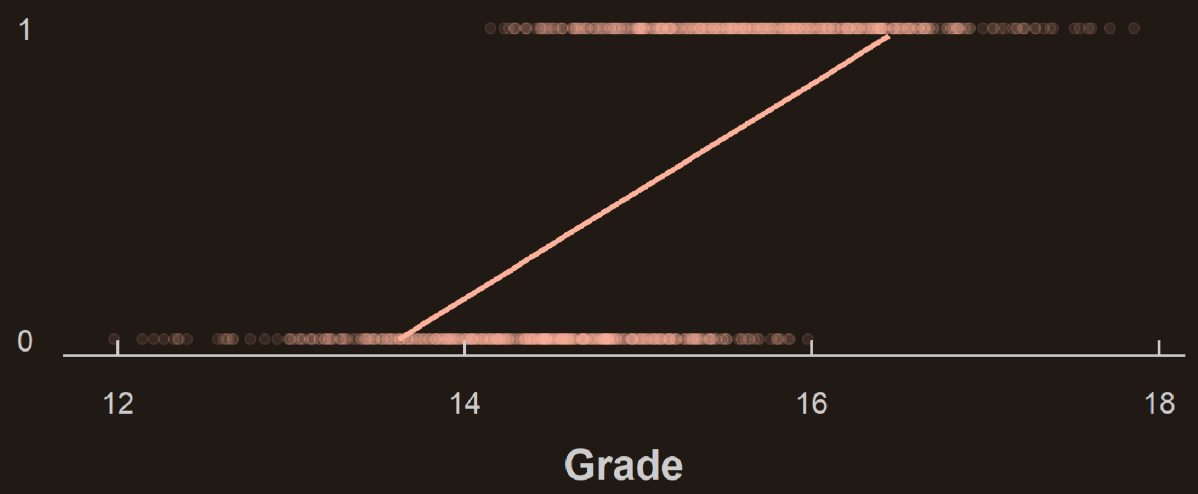


2. Regressions with discrete variables

2.1. Binary dependent variable

- 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

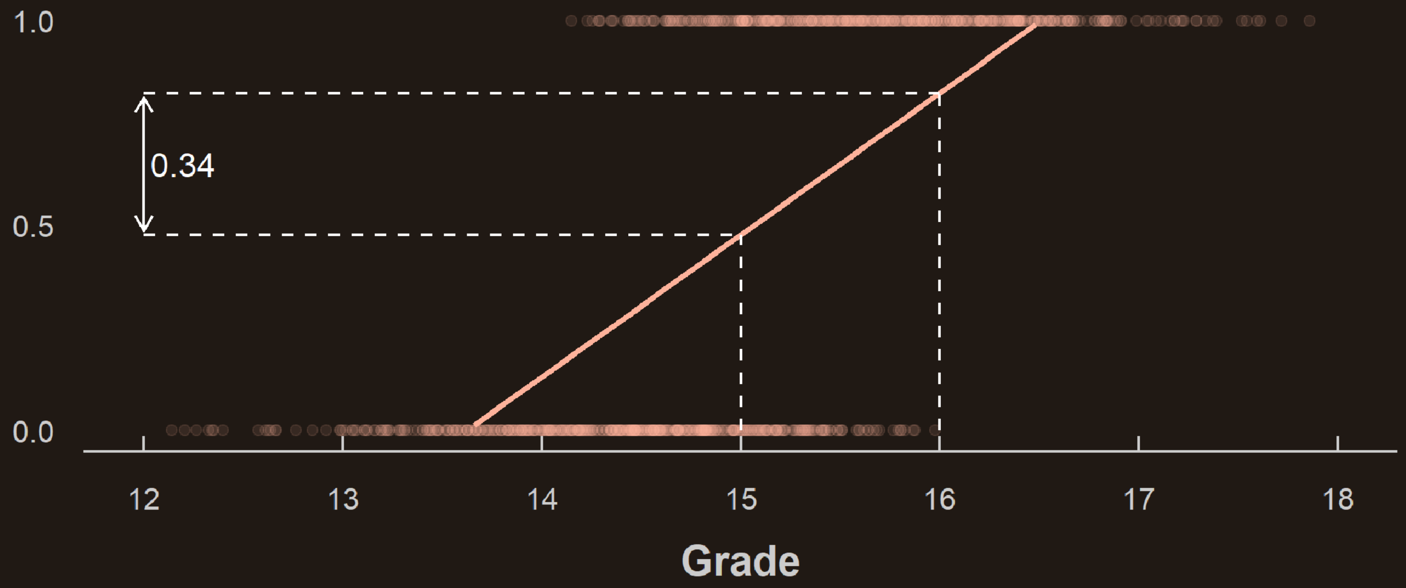
$$1\{y_i = \text{Accepted}\} = \hat{\alpha} + \hat{\beta} \times \text{Grade}_i + \hat{\varepsilon}_i$$



2. Regressions with discrete variables

2.1. Binary dependent variable

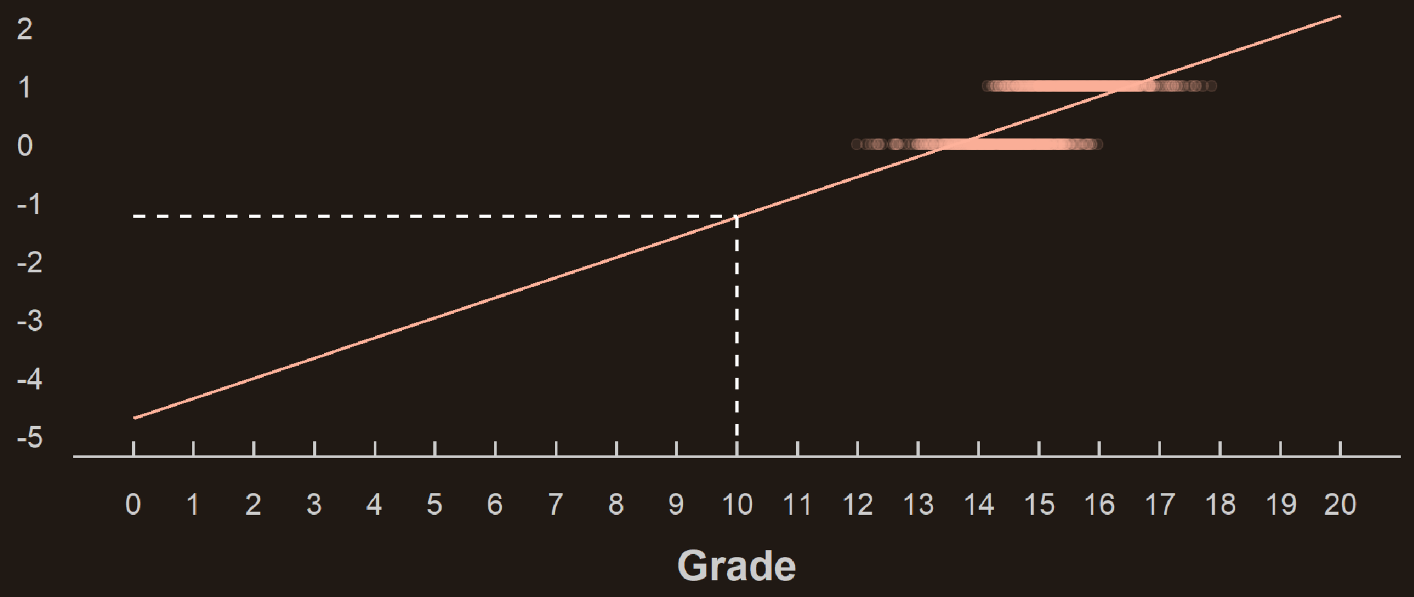
- The fitted values can be viewed as the probability to be accepted for a given grade
  - The slope is thus by how much the probability of being accepted would increase on expectation for a 1 point increase in the grade
  - That's why we call OLS regression models with a binary outcome *Linear Probability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



2. Regressions with discrete variables

2.1. Binary dependent variable

- But 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



2. Regressions with discrete variables

2.2. Binary independent variable

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

- So instead of
  - having a binary dependent variable :

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

  - we have a binary independent variable

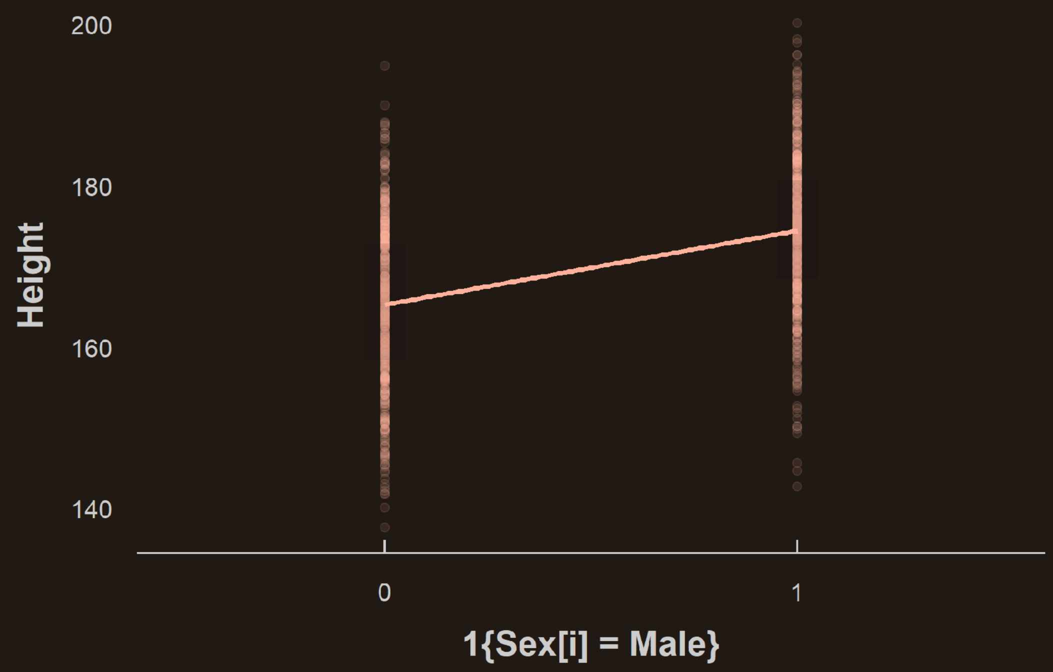
$$\text{Height}_i = \hat{\alpha} + \hat{\beta} \times 1\{x_i = \text{Male}\} + \hat{\varepsilon}_i$$

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

2. Regressions with discrete variables

2.2. Binary independent variable

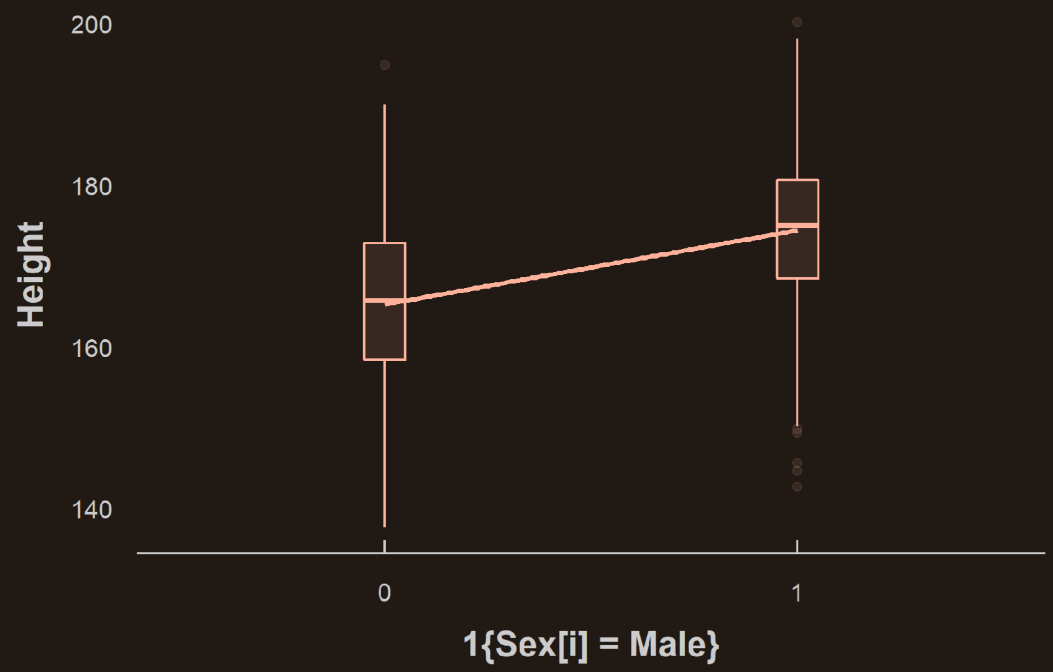
- If the sex variable was continuous it would be the expected increase in height for a *'1 unit increase'* in sex
  - Here the *'1 unit increase'* is switching from 0 to 1, i.e. from female to male
  - Here is the traditionnal scatter plot representation



2. Regressions with discrete variables

2.2. Binary independent variable

- Replacing 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



2. Regressions with discrete variables

2.2. Binary independent variable

- Let's have a look at the regression results and at the summary statistics of both distributions:

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                        Height
## ----------------------------------------
## SexMale                9.2***
##                        (0.6)
##
## Constant              165.3***
##                        (0.4)
##
## ----------------------------------------
## Observations           1,000
## R2                      0.2
## ========================================
## Note:      *p<0.1; **p<0.05; ***p<0.01
```

Height summary statistics by sex

Sex	Min	Q1	Med	Mean	Q3	Max
Female	137.7	158.4	165.8	165.3	172.9	194.9
Male	142.7	168.5	175.0	174.5	180.7	200.2

→ The $\hat{\alpha}$ coefficient is equal to the expected value of y when $x = 0$, i.e., to the average height for females

→ The $\hat{\beta}$ coefficient is equal to expected increase in y when going from $x = 0$ to $x = 1$, i.e., to the difference between male and female average height

2. Regressions with discrete variables

2.2. Binary independent variable

- Let's think of it in terms of a regression model:

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

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

$$\text{Height}_i = 165.0 + 9.8 \times 1\{x_i = \text{Male}\} + \hat{\varepsilon}_i$$

- The fitted values write:

$$\widehat{\text{Height}_i} = 165.0 + 9.8 \times 1\{x_i = \text{Male}\}$$

- When the dummy equals 0 (females):

$$\begin{aligned}\widehat{\text{Height}_i} &= 165.0 + 9.8 \times 0 \\ &= 165.0 = \overline{\text{Height}}_{[x_i=\text{Female}]}\end{aligned}$$

- When the dummy equals 1 (males):

$$\begin{aligned}\widehat{\text{Height}_i} &= 165.0 + 9.8 \times 1 \\ &= 174.8 = \overline{\text{Height}}_{[x_i=\text{Male}]}\end{aligned}$$

2. Regressions with discrete variables

2.3. Categorical independent variable

- So far we've been working with binary categorical variables:
  - Accepted vs. Rejected, Male vs. Female
  - But what about discrete variables with more than two categories?

- Take for instance the race variable:

```
asec_2020 <- read.csv("asec_2020.csv")
kable(asec_2020 %>% group_by(Race) %>% summarise(N = n()) %>% t(),
      caption = "Distribution of the Race categorical variable")
```

Distribution of the Race categorical variable

Race	Asian	Black	Other	White
N	4528	6835	2422	50551

→ How can we use this variable as an independent variable in our regression framework?

2. Regressions with discrete variables

2.3. Categorical independent variable

- Just as we converted our 2-category variable into 1 dummy variable, we can convert an n-category variable into $n-1$ dummy variables:

Sex	Male		Race	Black	Other	White
Female	0		Asian	0	0	0
Female	0		Asian	0	0	0
Female	0		Black	1	0	0
Female	0		Black	1	0	0
Male	1		Other	0	1	0
Male	1		Other	0	1	0
Male	1		White	0	0	1
Male	1		White	0	0	1

→ ***But why do we omit one category every time?***

- Females are observations for which Male equals 0
- Asians are observations for which Black, Other, and White each equals 0

→ Females and Asians are ***reference categories***

- The coefficient associated with the Male dummy was interpreted ***relative*** to females
- The coefficients associated with the Black, Other, and White dummies will be interpreted ***relative*** to Asians

2. Regressions with discrete variables

2.3. Categorical independent variable

- Thus, regressing earnings on the race categorical variable amounts to estimate the equation:

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}_1 1\{\text{Race}_i = \text{Black}\} + \hat{\beta}_2 1\{\text{Race}_i = \text{Other}\} + \hat{\beta}_3 1\{\text{Race}_i = \text{White}\} + \hat{\varepsilon}_i$$

- And if we compare the regression results to the average earnings by group:

```
summary(lm(Earnings ~ Race, asec_2020))$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  77990.78   1149.552  67.84449 0.000000e+00
## RaceBlack   -27413.29   1482.197 -18.49503 3.571079e-76
## RaceOther   -28512.08   1947.305 -14.64181 1.819073e-48
## RaceWhite   -15110.29   1199.933 -12.59262 2.559272e-36
```

Mean earnings by race

Race	Mean earnings
Asian	77990.78
Black	50577.49
Other	49478.70
White	62880.49

  - α is still the average earnings for the reference category
  - coefficient are still *relative* to the reference category

2. Regressions with discrete variables

2.3. Categorical independent variable

- As you can see from the previous regression results, by default R sorts categories by alphabetical order:

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   77990.78   1149.552  67.84449 0.000000e+00
## RaceBlack    -27413.29   1482.197 -18.49503 3.571079e-76
## RaceOther    -28512.08   1947.305 -14.64181 1.819073e-48
## RaceWhite    -15110.29   1199.933 -12.59262 2.559272e-36
```

- But oftentimes we would prefer the reference category to be the majority group
  - In R we can use the `relevel()` function to change the reference category of a factor

```
summary(lm(Earnings ~ relevel(as.factor(Race), "White"), asec_2020))$coefficients[, c(1, 2, 4)]
```

```
##                                         Estimate Std. Error     Pr(>|t|)
## (Intercept)                             62880.49   344.0464 0.000000e+00
## relevel(as.factor(Race), "White")Asian  15110.29  1199.9326 2.559272e-36
## relevel(as.factor(Race), "White")Black -12302.99   996.8981 5.947231e-35
## relevel(as.factor(Race), "White")Other -13401.79  1609.0045 8.294160e-17
```

Overview

1. Regressions with continuous variables ✓
- 1.1. Estimation
- 1.2. Inference

2. Regressions with discrete variables ✓
- 2.1. Binary dependent variable
- 2.2. Binary independent variable
- 2.3. Categorical independent variable

3. Controls and interactions

4. Interpretation

Overview

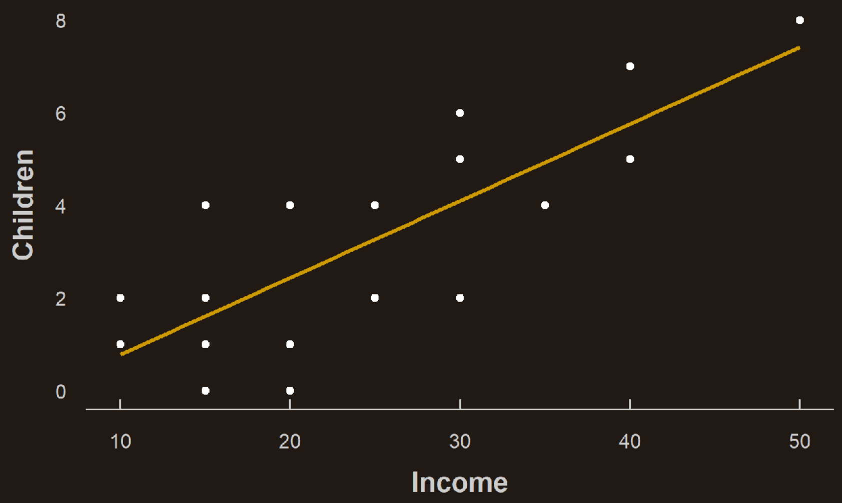
3. Controls and interactions

- We can add a third variable z in the regression for two reasons:
  - **Controlling for z** allows to **net out** the relationship between x and y from how they both relate to **z**
  - **Interacting x with z** allows to **estimate how the relationship** between x and y **varies with z**

- Consider the following fictitious dataset at the household level
  - Household annual income
  - Number of children in the household
  - Parents' education level

```r
data <- read.csv("household_data.csv")
head(data, 7) # fictitious data
```

```
##   Income Children    Education
## 1     20        1 < Highschool
## 2     10        1 < Highschool
## 3     10        2 < Highschool
## 4     15        0 < Highschool
## 5     15        1 < Highschool
## 6     20        0 < Highschool
## 7     15        2   Highschool
```

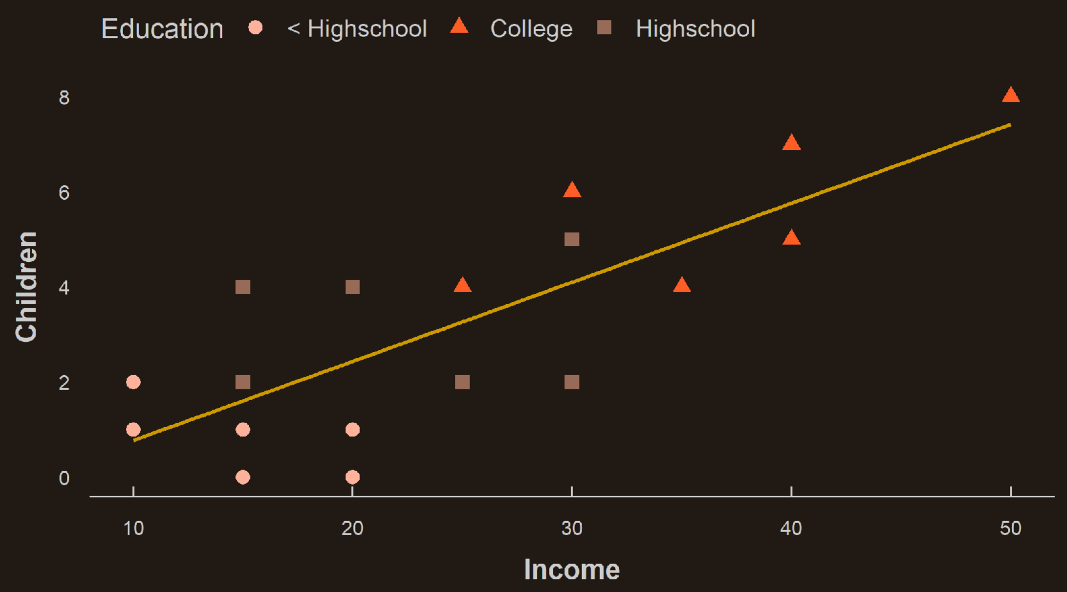


3. Controls and interactions

- There's a clear positive relationship

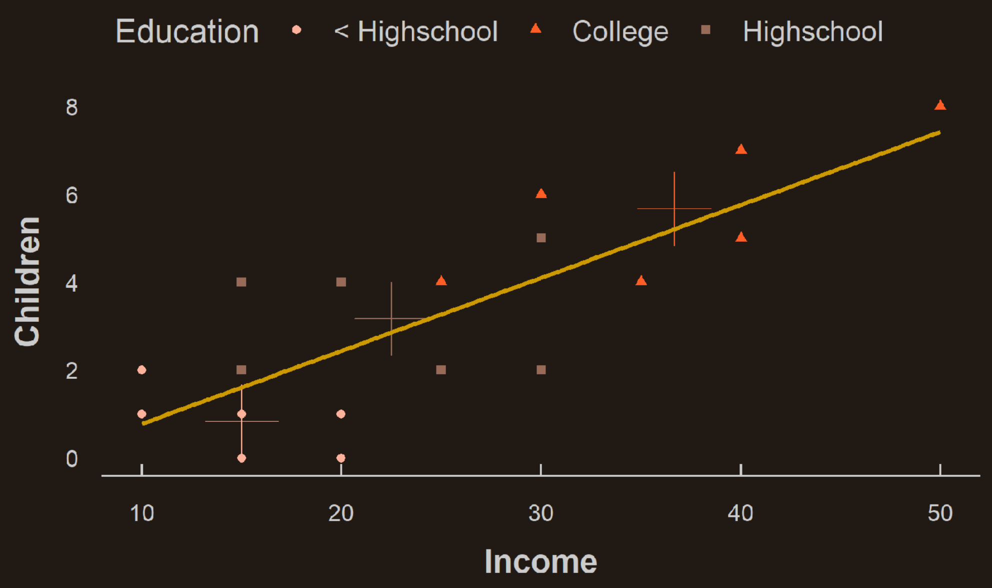
```
##               Estimate Pr(>|t|)
## (Intercept)    -0.885    0.319
## Income          0.166    0.000
```

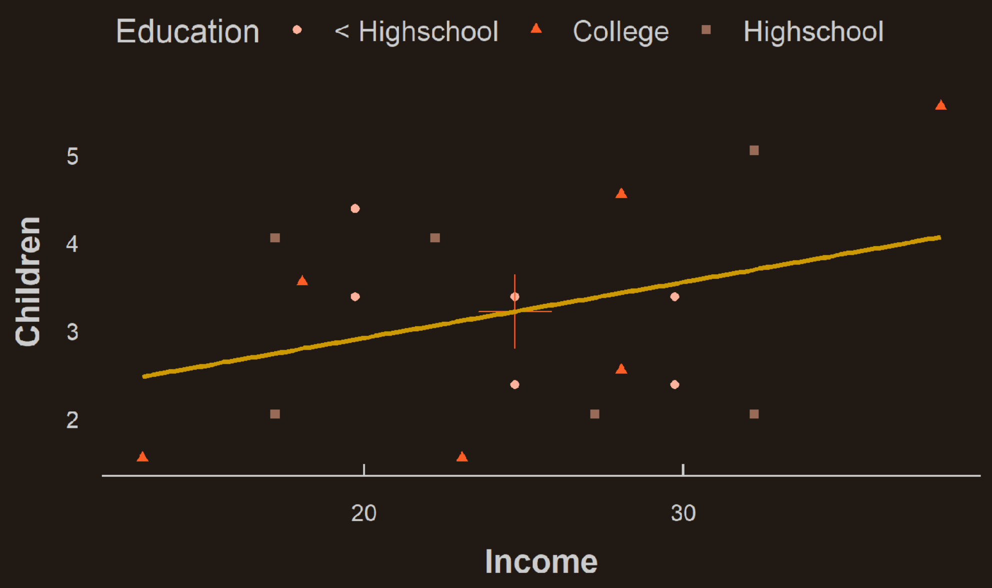
  - But what if this relationship was driven by a third variable?
  - Maybe it's just that more educated parents tend to earn more and to have more children



3. Controls and interactions

- **Controlling** for education does the same to the slope **as recentering** the graph with respect to education
  - In that way, when moving along the x axis, **z** does not increase but **remains constant**





- The crosses are located at the average x and y values for each education group
  - Controlling for education shifts x and y by group such that crosses superimpose

```
##                       Estimate Pr(>|t|)
## (Intercept)             -0.120    0.892
## Income                   0.064    0.196
## EducationCollege         3.456    0.015
## EducationHighschool      1.856    0.037
```

3. Controls and interactions

- Here when we **do not control** for education:

$$Children_i = \alpha + \beta Income_i + \varepsilon_i$$

  - We estimate the overall relationship (here, significantly positive)

- But when we **control** for education:

$$Children_i = \alpha + \beta Income_i + \gamma_1 1\{Education_i = \text{Highschool}\} + \gamma_2 1\{Education_i = \text{College}\} + \varepsilon_i$$

  - We estimate the relationship net of the effect of education (here, not significant)

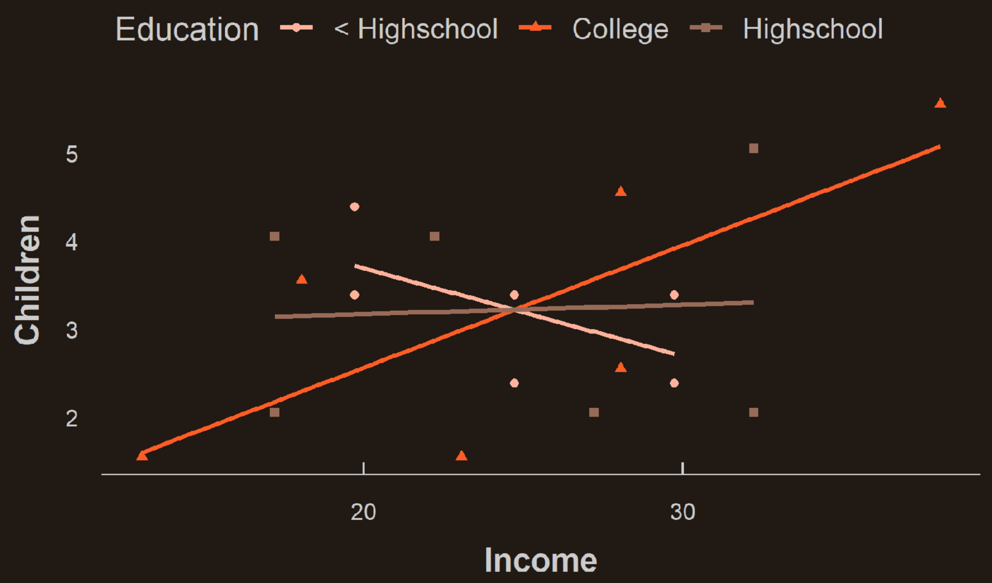
- **Interacting** the two variables is going one step further:

$$\begin{aligned} Children_i = \alpha + \beta Income_i + \gamma_1 1\{Education_i = \text{Highschool}\} + \gamma_2 1\{Education_i = \text{College}\} + \\ \delta_1 Income_i \times 1\{Education_i = \text{Highschool}\} + \delta_2 Income_i \times 1\{Education_i = \text{College}\} + \varepsilon_i \end{aligned}$$

  - It is not simply taking into account the fact that education may plays a role
  - It estimates by how much the relationship between x and y varies according to z

3. Controls and interactions

- **Interacting** income with education provides **one slope per education group**:




```
##                            Estimate Pr(>|t|)
## (Intercept)                   2.333    0.225
## Income                       -0.100    0.411
## EducationCollege             -1.768    0.553
## EducationHighschool           0.596    0.819
## Income:EducationCollege       0.239    0.095
## Income:EducationHighschool    0.111    0.445
```

- The principle is the same when the third variable is continuous:
  - Controlling nets out the slope from how the third variable enters the relationship
  - Interacting gives by how much the slope changes on expectation when the third variable increases by 1
  - And we can control for/interact with multiple third variables

Overview

4. Interpretation

Train at interpreting coefficients from randomly drawn relationships