APPLIED STATISTICAL ANALYSIS I

Contingency tables, correlation & bivariate regression

Trajche Panov, PhD
panovt@tcd.ie

Department of Political Science
Trinity College Dublin

September 24, 2023

Today's Agenda

(1) Lecture recap
(2) Software check, git and GitHub
(3) Tutorial exercises
(4) Software check, LaTeX and TeXstudio

Joint and conditional probability distributions

What is a contingency table? What is a joint probability distribution? What is a conditional probability distribution? Why is this important?

Contingency tables

What is a contingency table?

- A contingency table: "displays the number of subjects observed at all combinations of possible outcomes for the two variables" (Agresti and Finlay 2009, 221).
- "The row totals and the column totals are called the **marginal distributions**" (Agresti and Finlay 2009, 222).
- Is there an association between gender and party affiliation? Does party affiliation depend on gender?

TABLE 8.1: Party Identification (ID) and Gender, for GSS Data

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

Joint distribution

What is a joint distribution?

- Joint distribution: describes the probability that two variables (e.g., X and Y) simultaneously take some values.
- What is joint probability of Gender=Females and Party Identification=Democrat? $\frac{573}{2771} = 0.2067846 = 21\%$

TABLE 8.1: Party Identification (ID) and Gender, for GSS Data

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

Conditional distribution

What is a conditional distribution?

- Conditional distribution: describes the probability of one variable Y taking different values, conditional on another variable X having a specific value.
- What is the probability of Party Identification=Democrat conditional on Gender=Females? $\frac{573}{1511} = 0.3792191 = 38\%$

TABLE 8.2: Party Identification and Gender: Percentages Computed within Rows of Table 8.1

	Party Identification				
Gender	Democrat	Independent	Republican	Total	*n*
Females	38%	34%	28%	100%	1511
Males	31%	38%	32%	101%	1260

Independence and dependence

Why is this important?

- "Two categorical variables are **statistically independent** if the population conditional distributions on one of them are identical at each category of the other" (Agresti and Finlay 2009, 223).
- "The variables are **statistically dependent** if the conditional distributions are not identical" (Agresti and Finlay 2009, 223).

TABLE 8.3: Population Cross-Classification Exhibiting Statistical Independence. The conditional distribution is the same in each row, (44%, 14%, 42%).

	Party Identification			
Ethnic Group	Democrat	Independent	Republican	Total
White	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Black	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Hispanic	110 (44%)	35 (14%)	105 (42%)	250 (100%)

Chi-square test of independence

What is the Chi-square test of independence? What is the Chi-square distribution?

Chi-square test of independence

What is the Chi-square test of independence?

- Null and alternative hypothesis: Two variables are independent, $f_o = f_e$ (H_0), two variables are dependent, $f_o \neq f_e$ (H_a).
- Test statistics: "compares the observed frequencies in the contingency table with values that satisfy the null hypothesis of independence" (Agresti and Finlay 2009, 225), $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$
- "Let f_o denote an observed frequency in a cell of the table. Let f_e denote an expected frequency. This is the count expected in a cell if the variables were independent. It equals the product of the row and column totals for that cell, divided by the total sample size" (Agresti and Finlay 2009, 225).

Chi-square test of independence

TABLE 8.4: Party Identification by Gender, with Expected Frequencies in Parentheses

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Female	573 (522.9)	516 (540.4)	422 (447.7)	1511
Male	386 (436.1)	475 (450.6)	399 (373.3)	1260
Total	959	991	821	2771

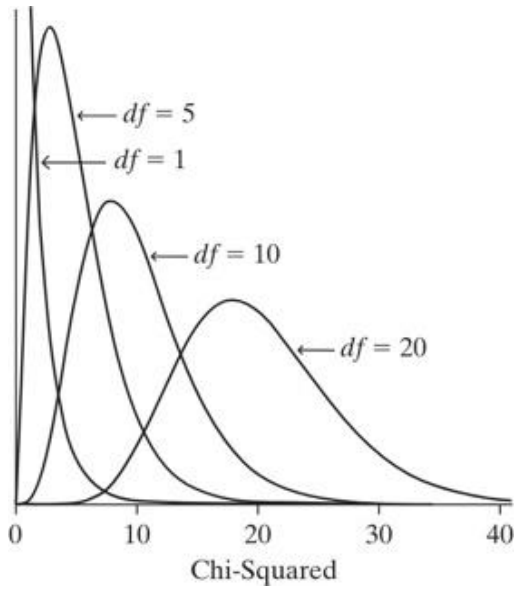
Female and democrat: $f_o = 573$ and $f_e = \frac{1511 \cdot 959}{2771} = 422.9$

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(573-522.9)^2}{522.9} + \frac{(516-540.4)^2}{540.4} + \ldots + \frac{(399-373.3)^2}{373.3} = 16.2$$

How to interpret this value? How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true?
→ Probability distribution

Chi-square distribution

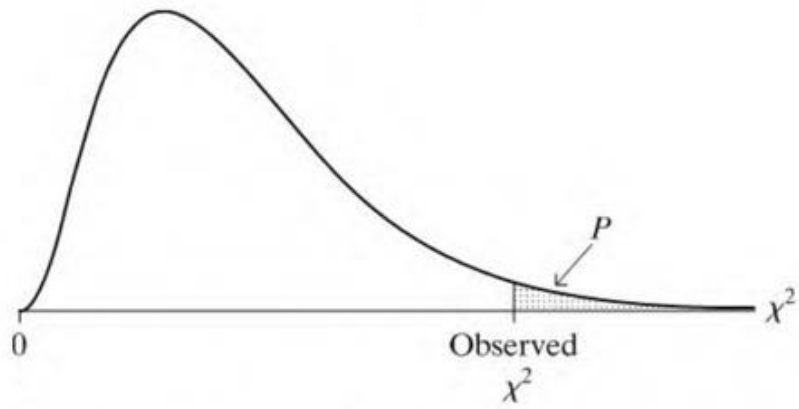
What is the chi-square distribution?



With degrees of freedom (df) for number of rows (r) and number of columns (c), $df = (r - 1)(c - 1)$ (Agresti and Finlay 2009, 226).

Chi-square distribution

How likely are we to observe data in sample (this test statistics), under the assumption that H_0 is true?



What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0% (p=0.0003). → Gender and party affiliation are not independent

Scatter plot

What is a scatter plot?

Scatter plot

What is a scatter plot?

- A plot, showing two continuous variables (e.g., X and Y) alongside each other → for each observation, the value on X is plotted against the value on Y.

Scatter plot

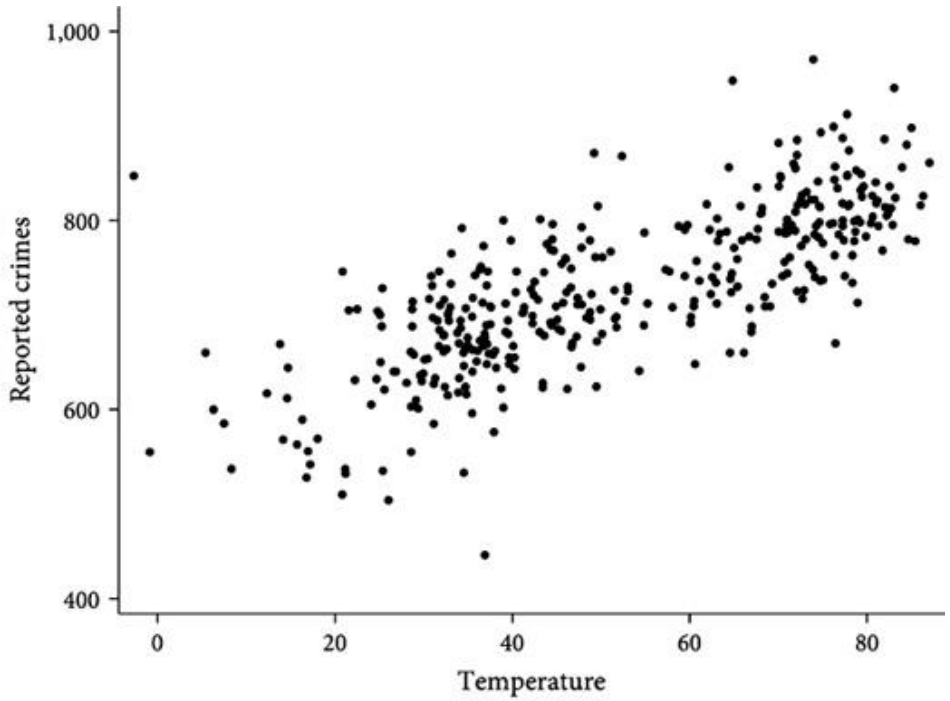


Figure 2.1. Crime and temperature (in degrees Fahrenheit) in Chicago across days in 2018.

- "each point corresponds to an observation in our data–here, that means each point is a day in Chicago in 2018" (Bueno de Mesquita and Fowler 2021, 15–16).
- "the location of each point shows the average temperature and the amount of crime on a given day" (Bueno de Mesquita and Fowler 2021, 15–16).

Correlation

What is correlation? How can we measure correlation?

Correlation

What is correlation?

- "The *correlation* between two features of the world is the extent to which they tend to occur together" (Bueno de Mesquita and Fowler 2021, 13).
- "If two features of the world tend to occur together, they are *positively correlated*" (Bueno de Mesquita and Fowler 2021, 13).
- "If the occurrence of another feature of the world is unrelated to the occurrence of another feature of the world, they are *uncorrelated*" (Bueno de Mesquita and Fowler 2021, 13).
- "And if when one feature of the world occurs the other tends not to occur, they are *negatively correlated*" (Bueno de Mesquita and Fowler 2021, 13).

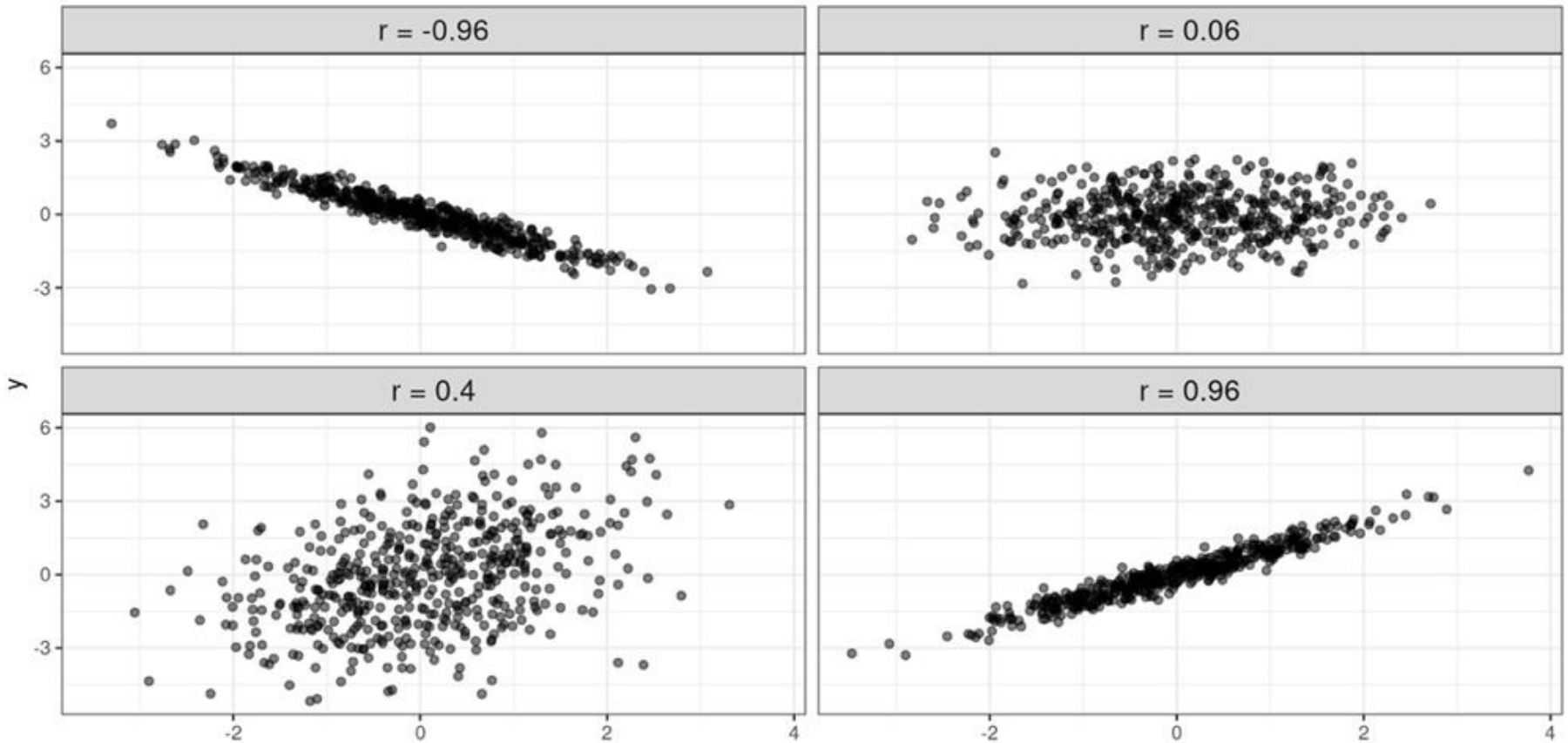
Correlation

How can we measure correlation?

- Correlation: (correlation coefficient, Pearson correlation coefficient, Pearson's r, r) standardized average of the product of deviations of two variables from the mean (=standardized covariance) $r = \frac{1}{n-1} \sum \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$
- ranges between -1 and 1, with 0=no association, the larger the absolute value, the stronger the association

Correlation

What is correlation?



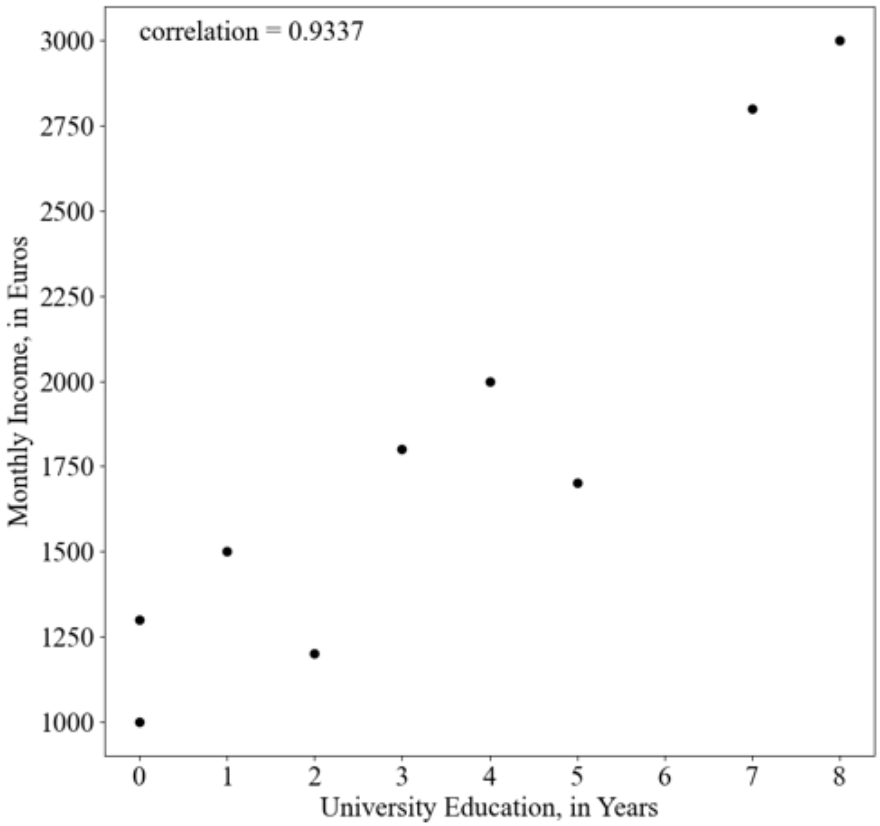
Shortcomings of correlation analysis

- no indication on the "substantive importance or size of the relationship between X and Y" (Bueno de Mesquita and Fowler 2021, 29).
- Slope: "tells us, descriptively, how much Y changes, on average, as X increases by one unit" (Bueno de Mesquita and Fowler 2021, 29).

Linear regression model

What is a linear regression model? What interpretations can we make?

So far, correlation analysis



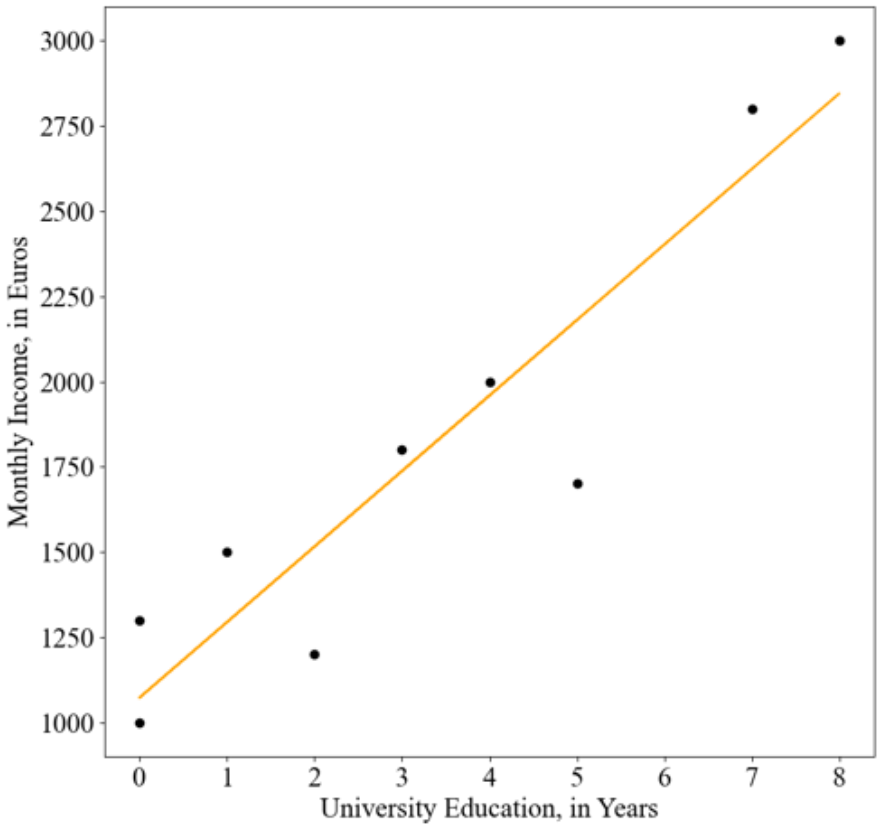
Now, just by looking at the plot, can you identify the straight line which best describes the joint variation between X and Y?

*This is fictional data.

Regression analysis

What is a linear regression model?

- Find linear line of best fit, $Y_i = \alpha + \beta X_i + \epsilon_i$



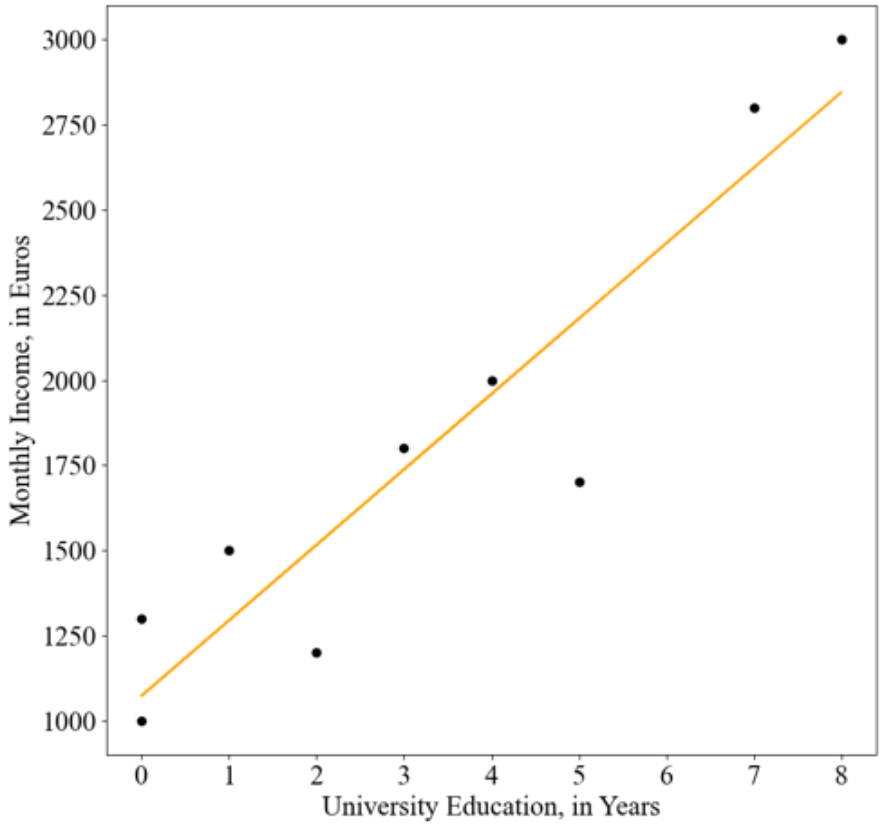
Regression analysis

What is a linear regression model?

- Find linear line of best fit, $Y_i = \alpha + \beta X_i + \epsilon_i$
- α (intercept): expected value of Y when $X = 0$
- β (slope): expected change in Y when X increases by one unit
- $\hat{Y}$(expected value): predicted outcome based on the regression model, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- ϵ (error/residual): difference between actual and predicted outcome, $\epsilon_i = Y_i - \hat{Y}_i$

Regression analysis

What interpretations can we make?

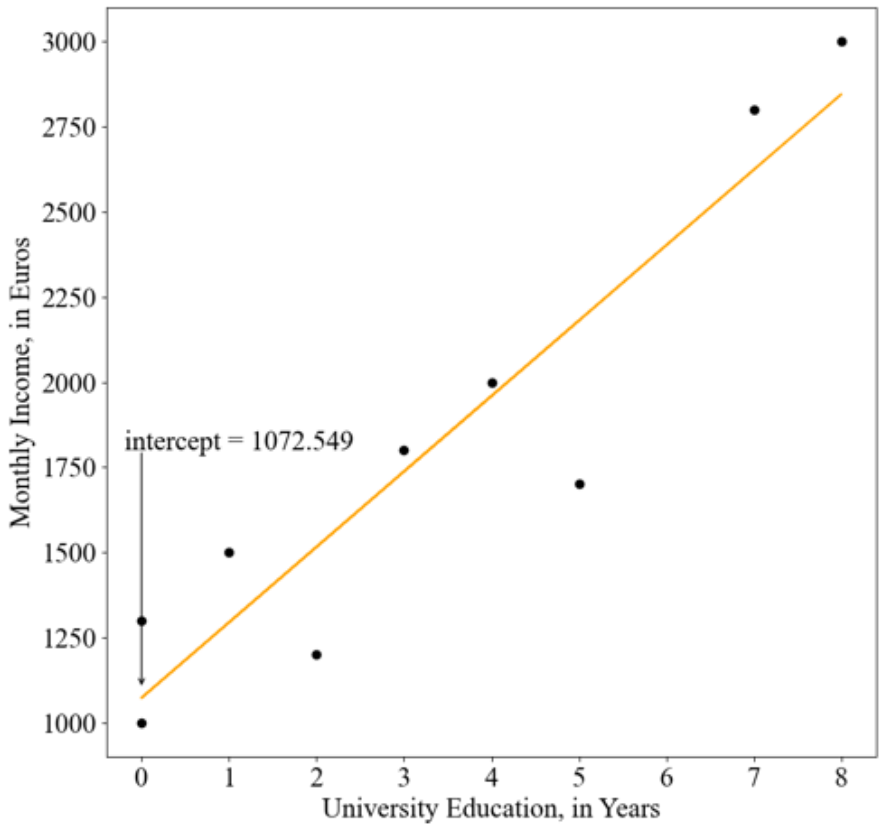


$$income = \alpha + \beta * education$$

$$income = 1072.5490 + 221.5686 * education$$

Regression analysis

What interpretations can we make? (intercept)

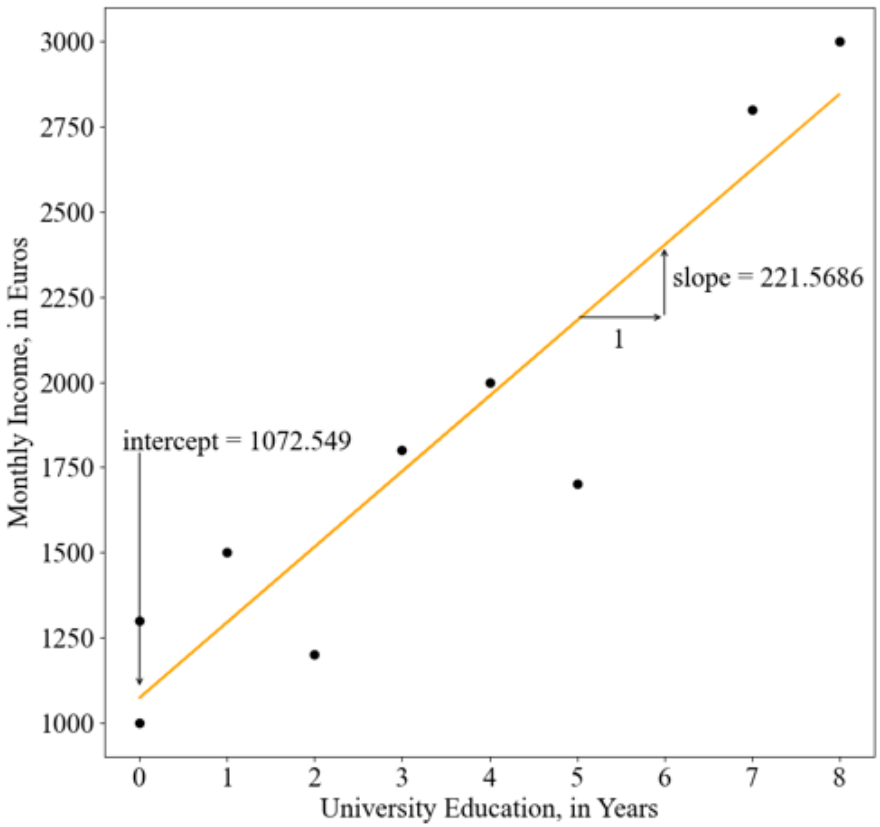


If an individual has a university education of 0 years, what income would we expect for that person?

$$income = 1072.5490 + 221.5686 * 0 = 1072.5490$$

Regression analysis

What interpretations can we make? (slope)

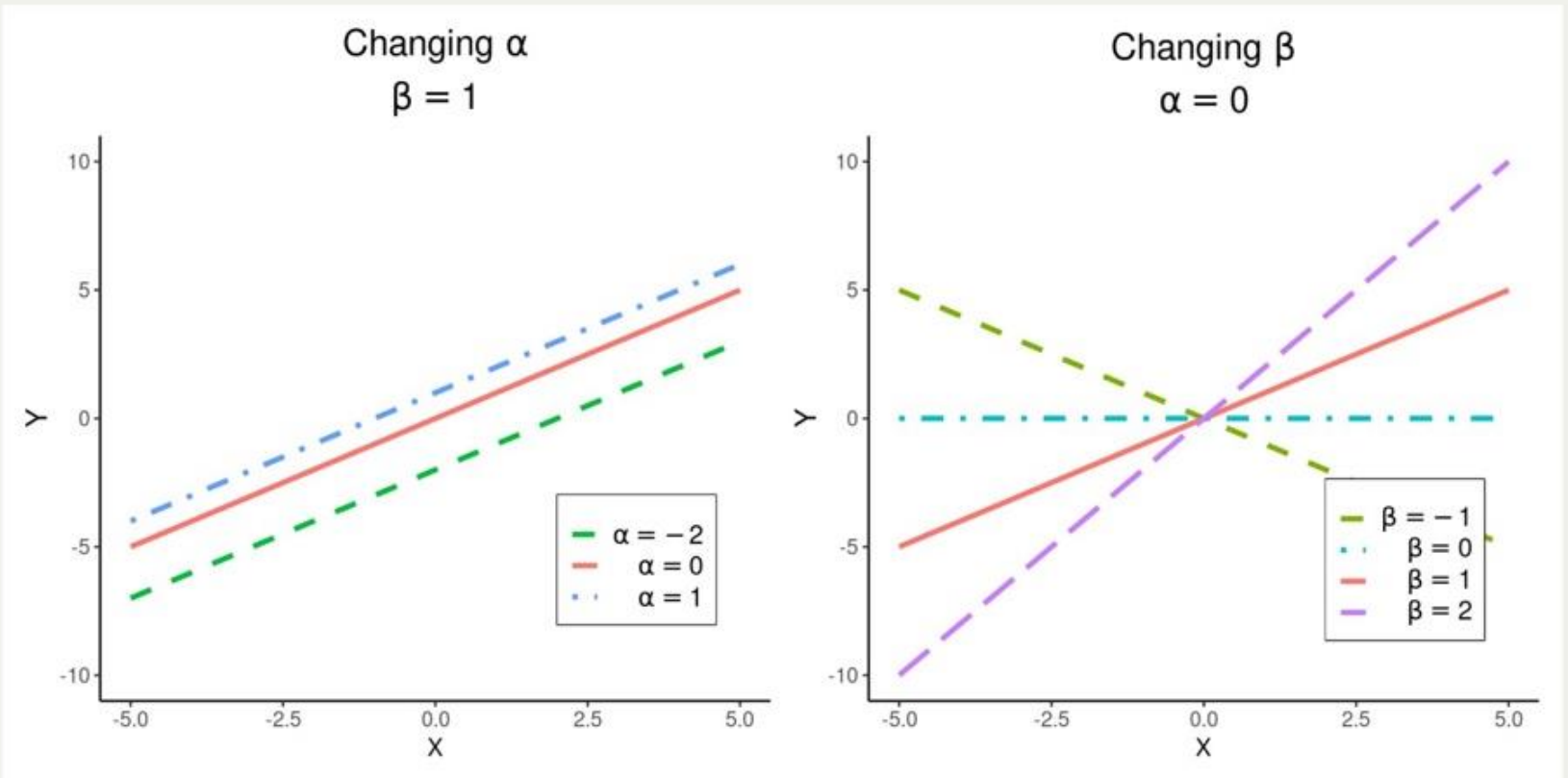


If the university education increases by one year, how much more Euros would we expect an individual to earn? $income = 1072.5490 + 221.5686 * 1 = 1294.1176$

→ With every additional year of university education, the expected income increases by 221.5686 Euros.

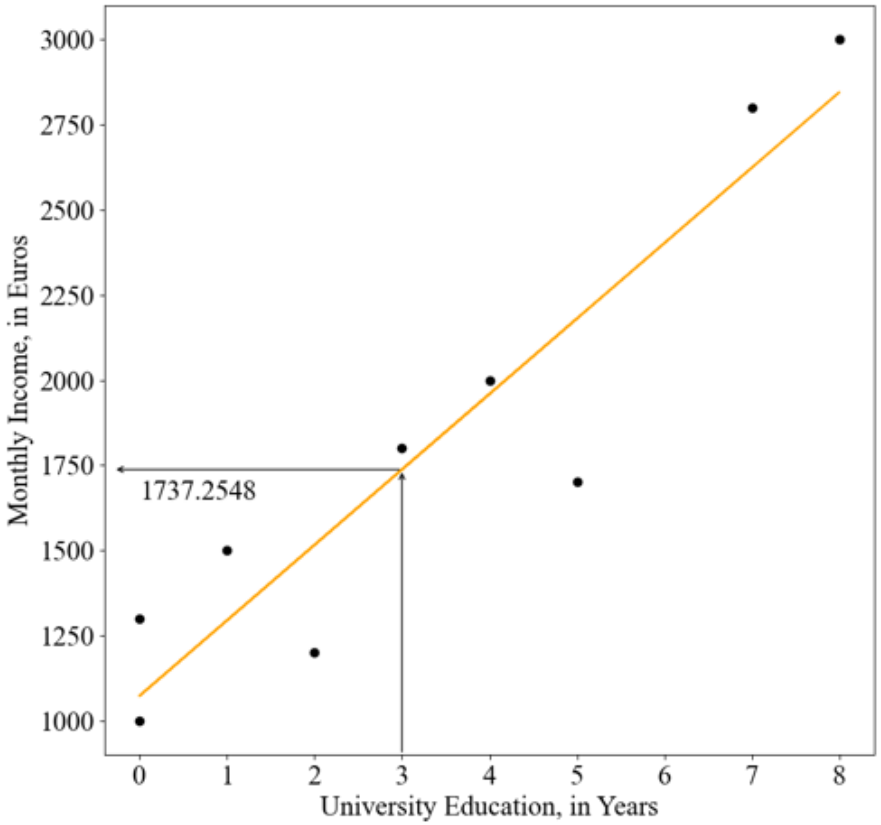
Regression analysis

Varieties of linear relationships



Regression analysis

What interpretations can we make? (expected value)

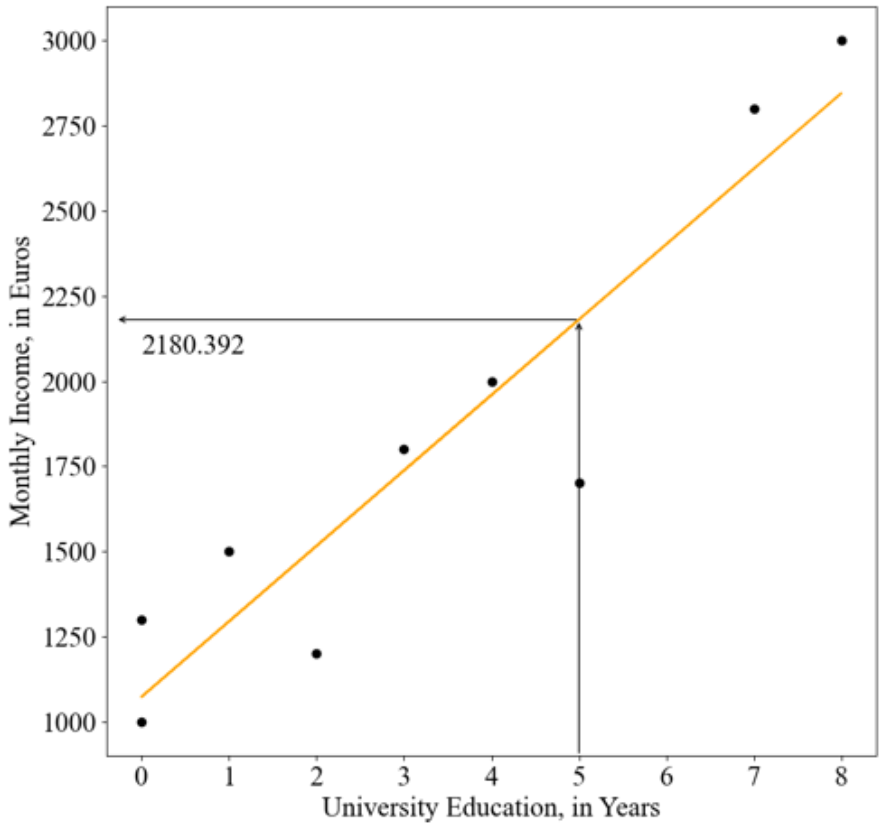


If an individual has 3 university education years, what income would we expect for that person?

$$income = 1072.5490 + 221.5686 * 3 = 1737.2548$$

Regression analysis

What interpretations can we make? (residual)



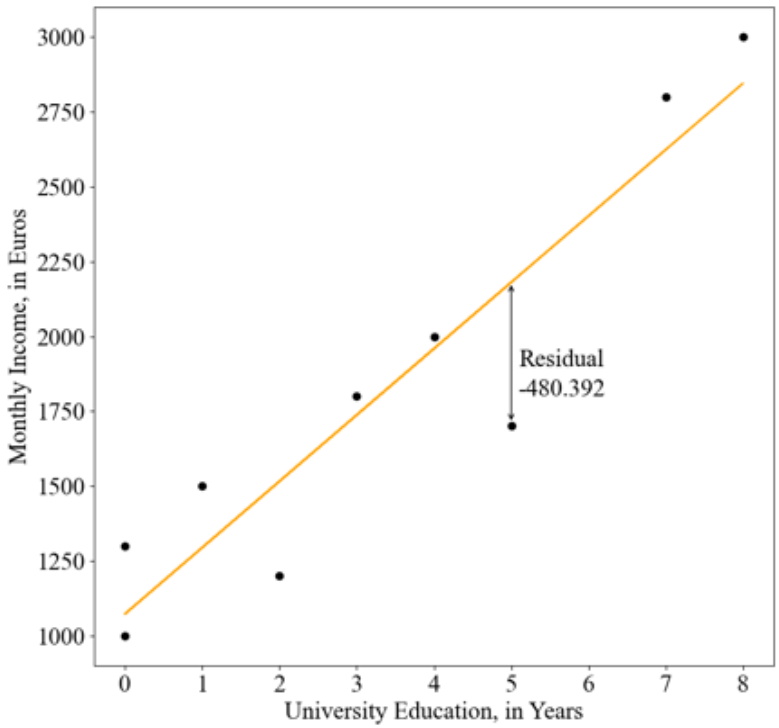
$$income = 1072.5490 + 221.5686 * 5 = 2180.392$$

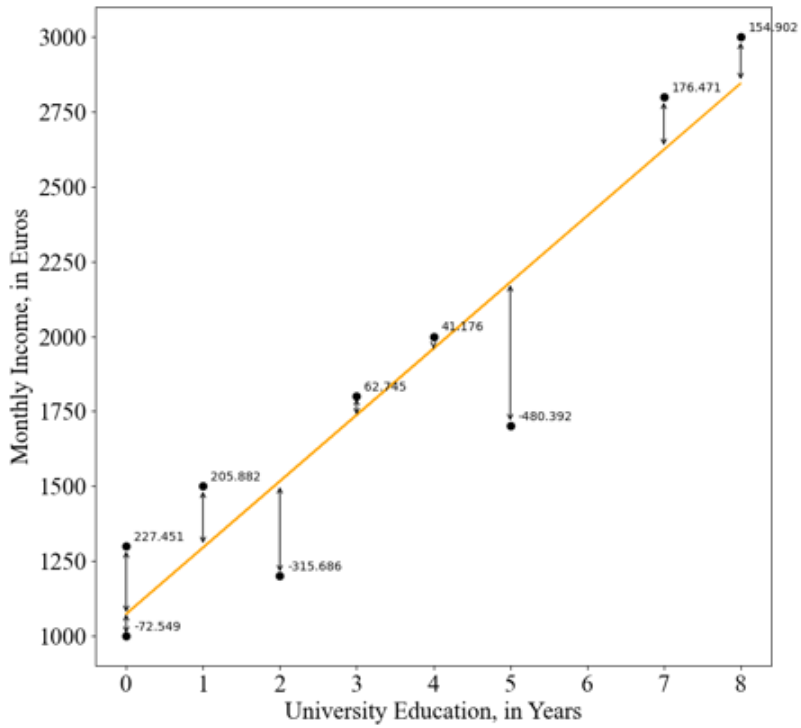
$$\text{Residual} = \text{Actual} - \text{Predicted}$$

$$\text{Residual} = 1700 - 2180.392 = -480.392$$

Regression analysis

What interpretations can we make? (residuals)





Ordinary least squares (OLS)

How are intercept and slope estimated?

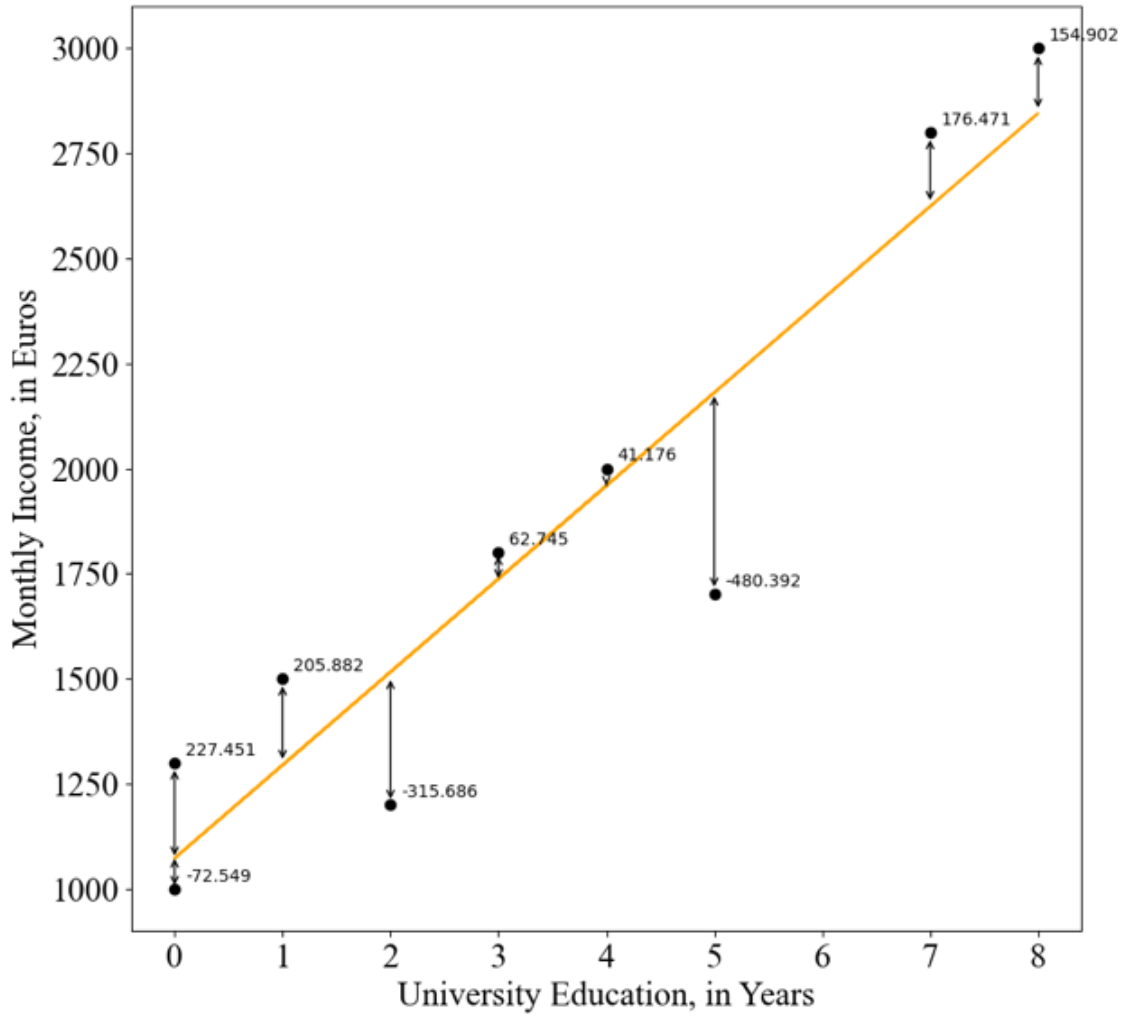
Ordinary least squares (OLS)

How are intercept and slope estimated?

- How do we find the line which best fits the data?
- Apply the OLS (Ordinary Least Squares) method, which minimizes the sum of squared errors (SSE).
- Sum of squared errors = the sum of squared differences between actual and predicted values of Y.
- $SSE = \sum_{i=1}^{n}(\hat{\epsilon_i})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\alpha} - \hat{\beta}X_i))^2$ → minimize this!

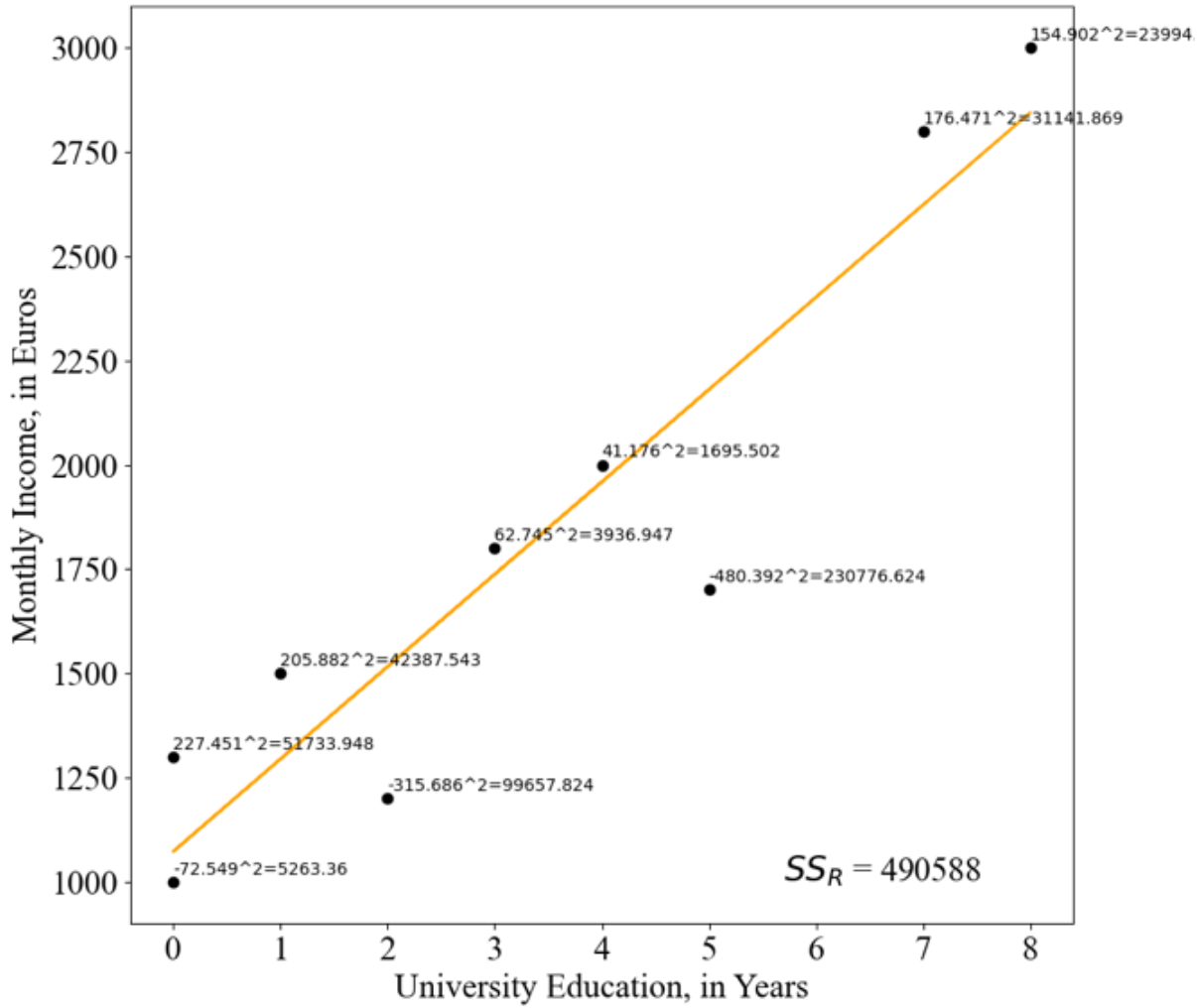
Ordinary least squares (OLS)

How are intercept and slope estimated?



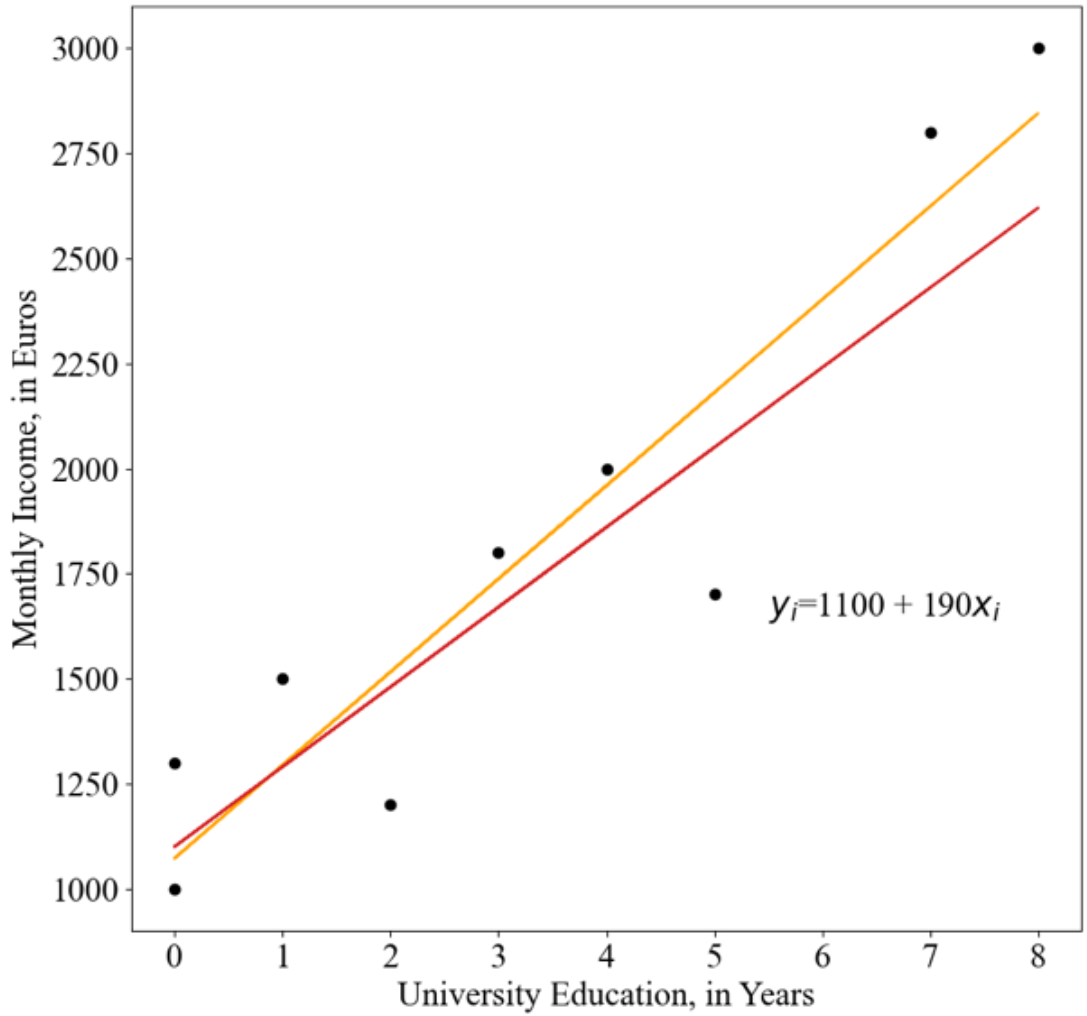
Ordinary least squares (OLS)

How are intercept and slope estimated?



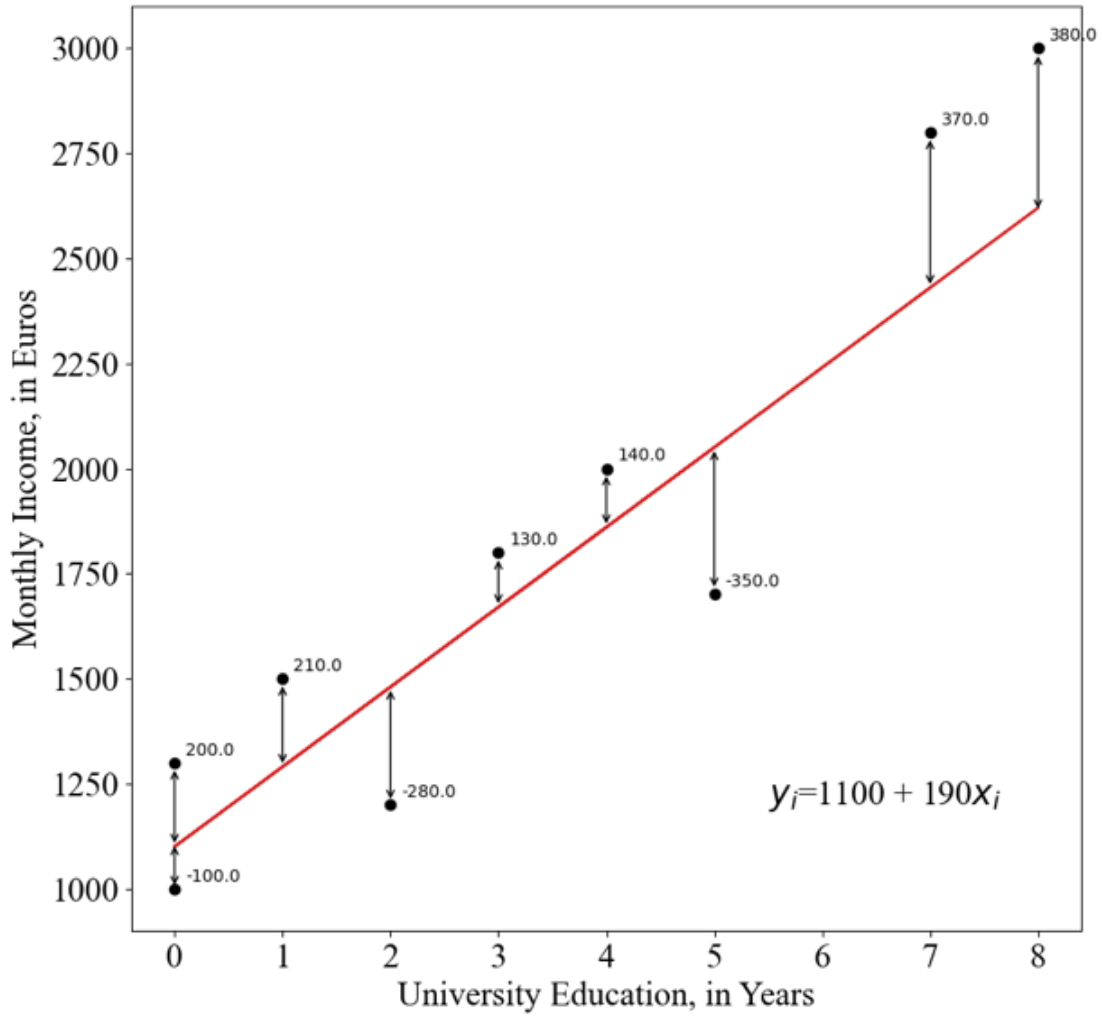
Ordinary least squares (OLS)

How are intercept and slope estimated?



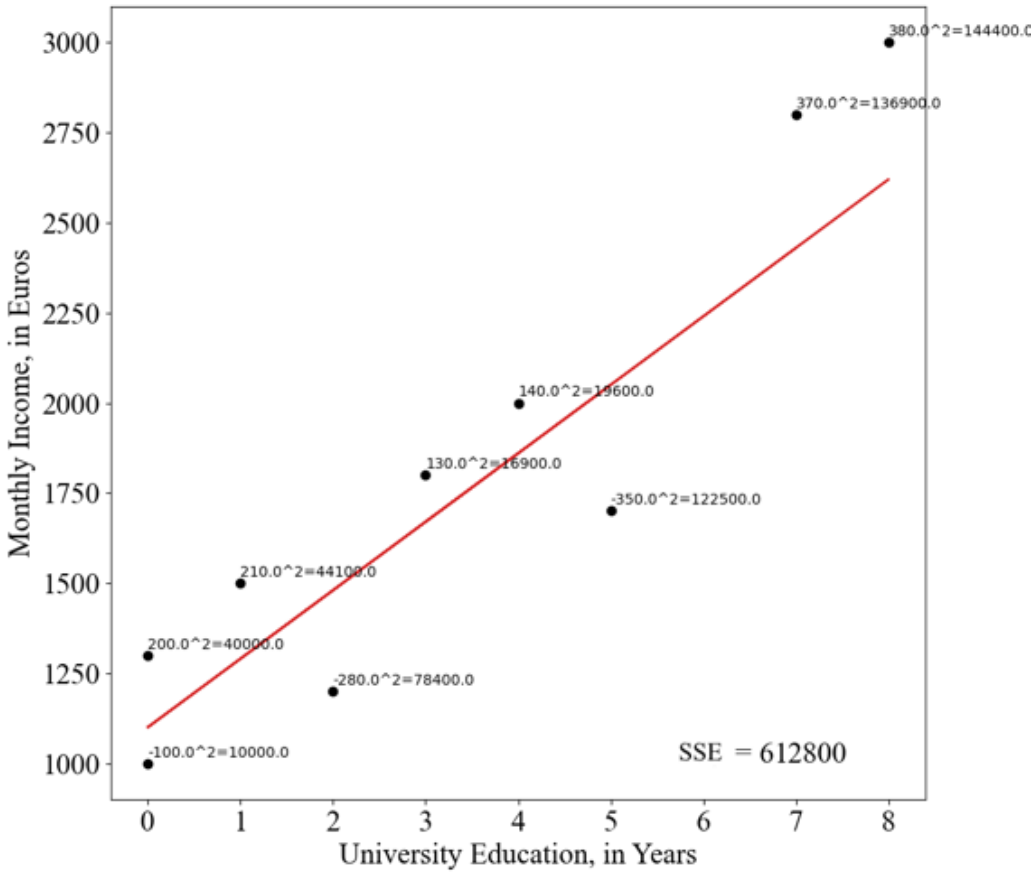
Ordinary least squares (OLS)

How are intercept and slope estimated?



Ordinary least squares (OLS)

How are intercept and slope estimated?



$612,800 > 490,588 \longrightarrow SSE(RED) > SSE(ORANGE)$

$\longrightarrow$ Orange regression line has better fit.

OLS assumptions

What are the assumptions of linear regression?

Assumptions of linear regression

Assumptions about the error (ϵ_i):

$$\epsilon_i \sim N(0, \sigma^2)$$

* ϵ_i is normally distributed
* $E(\epsilon_i) = 0$, no bias
* ϵ_i has constant variance σ^2 (Homoscedasticity)
* No autocorrelation
* X values are measured without error

(Kellstedt and Whitten 2018, 190–194)

Assumptions of linear regression

Assumptions about the model specification:

* No causal variables left out and no noncausal variables included
* Parametric linearity

(Kellstedt and Whitten 2018, 190–194)

Assumptions of linear regression

Minimal mathematical requirements:

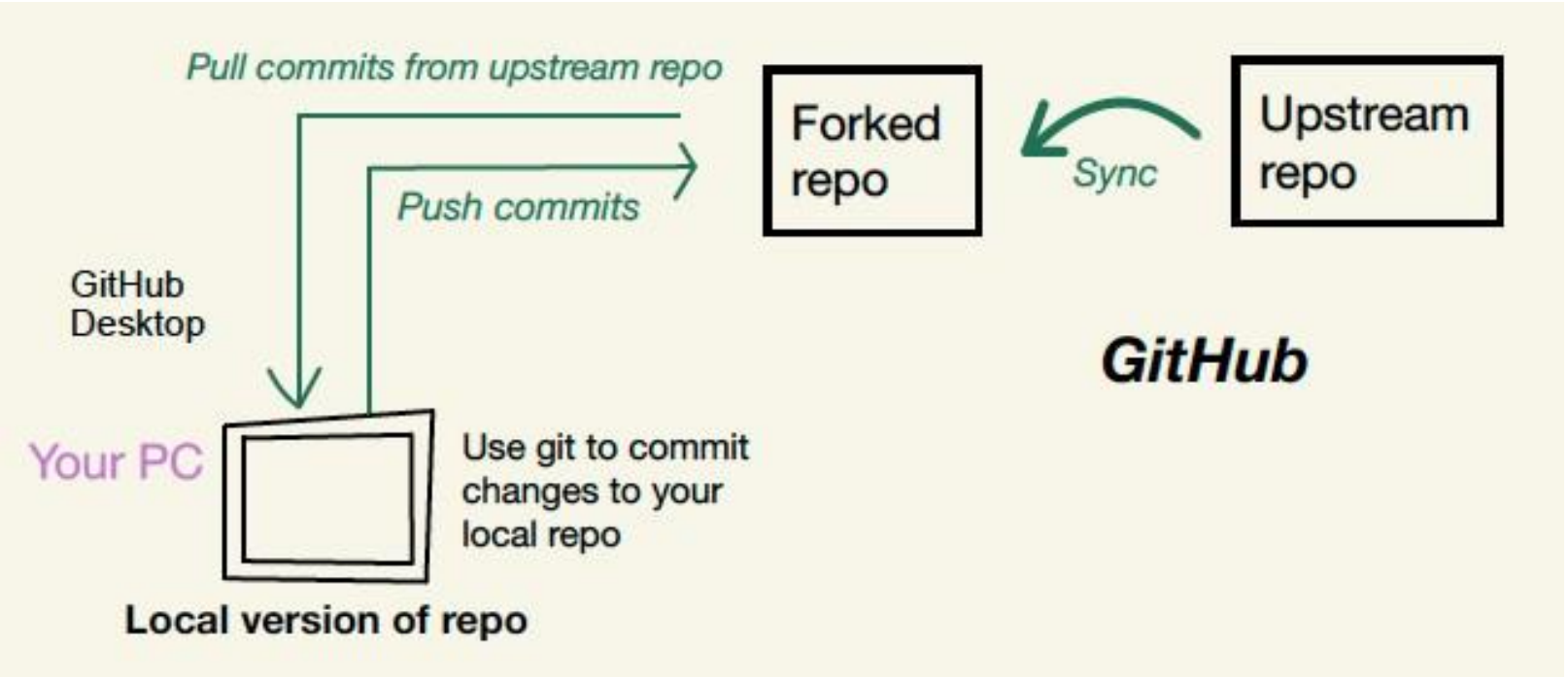
* X must vary
* Number of observations must be larger than the number of predictors
* In multiple regression: No perfect multicollinearity

(Kellstedt and Whitten 2018, 190–194)

Software check

How to update your local repository? How to git pull?

Software check

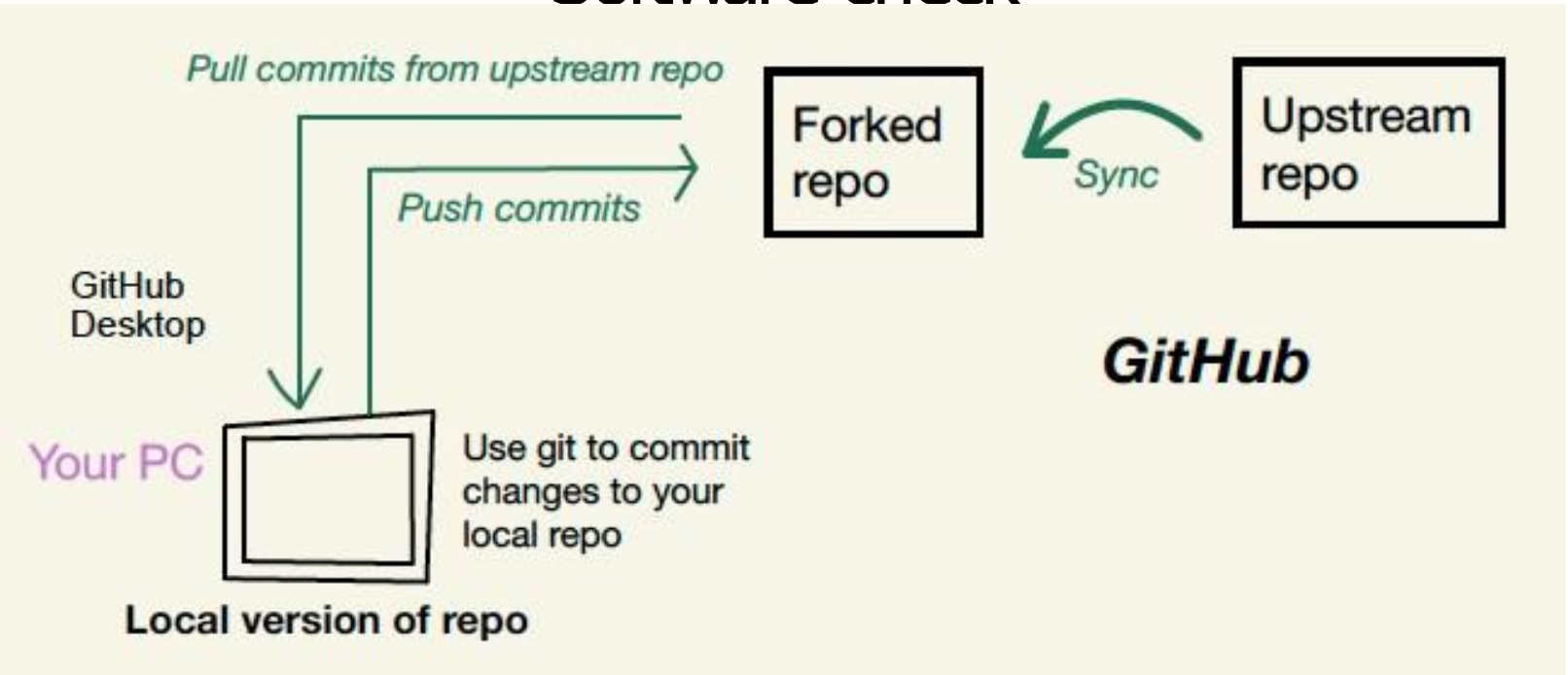


1. Synchronize fork
2. Fetch origin

Software check

How to update your repository on GitHub? How to git push?

Software check



1. Make sure your local repository is updated:
    * Synchronize fork
    * Fetch origin
2. Commit changes in file (local)
3. Push commits to fork

References I

- Agresti, Alan, and Barbara Finlay. 2009. *Statistical methods for the social sciences.* Essex: Pearson Prentice Hall.
- Bueno de Mesquita, Ethan, and Anthony Fowler. 2021. *Thinking clearly with data: A guide to quantitative reasoning and analysis.* Princeton: Princeton University Press.
- Kellstedt, Paul M., and Guy D. Whitten. 2018. *The fundamentals of political science research.* Cambridge: Cambridge University Press.