Applied Statistical Analysis I

Bivariate regression, inference & prediction

Hannah Frank
frankh@tcd.ie

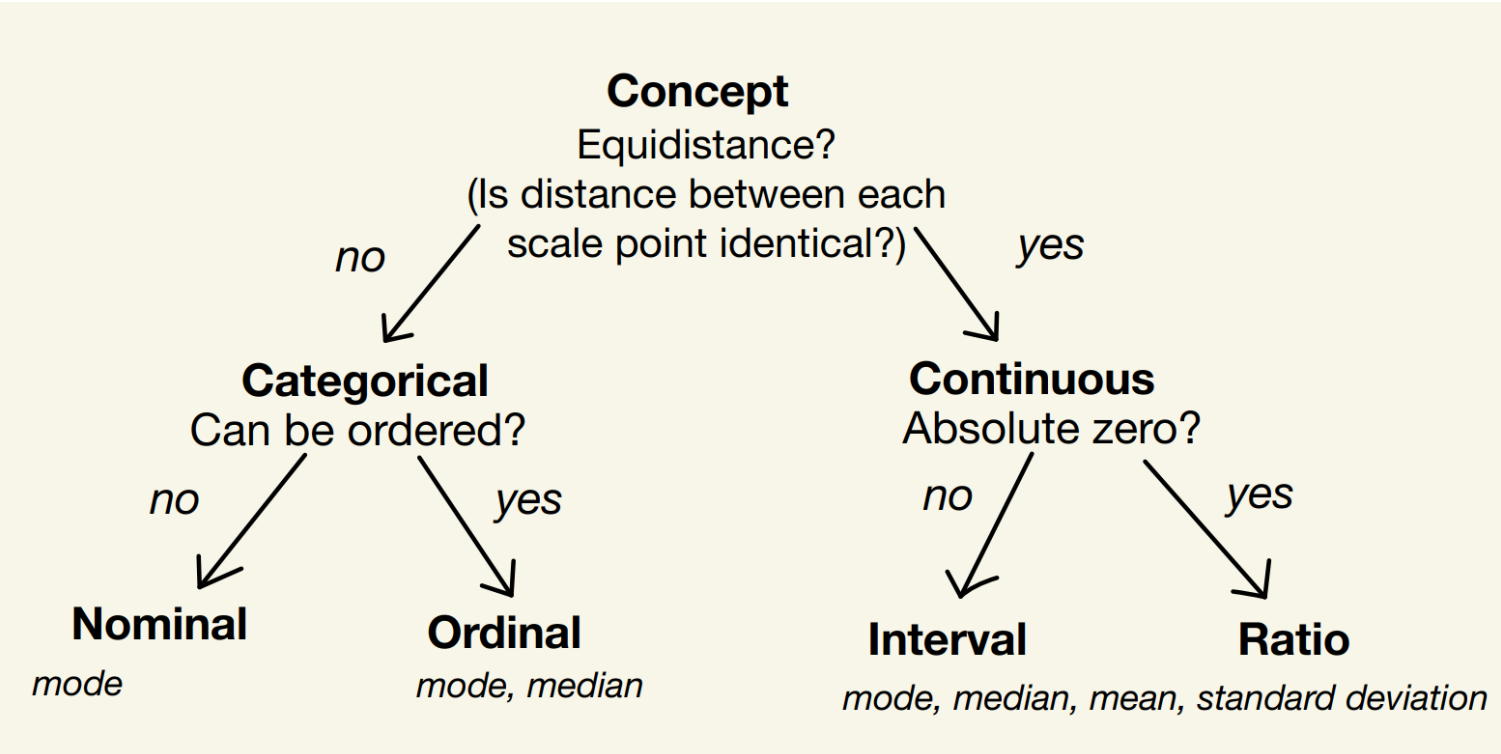
Department of Political Science
Trinity College Dublin

October 4, 2023

Today's Agenda

(1) Lecture recap
(2) Git pull
(3) Tutorial exercises

Importance of measurement scales



(Kellstedt and Whitten 2018, Chap. 5)

Discrete: finite set of possible values (Contingency tables, chi-square test)
Continuous: infinite set of possible values (t-test for mean and difference in means, correlation, scatter plot, dependent variable in linear regression)

Correlation

What is correlation? How can we measure correlation?

How can we test the statistical significance of 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?

- Covariance: covariance is the average of the product of deviations of two quantitative variables from the mean, $cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$
- Positive association, if larger-than-average X_i co-occurs with larger-than-average Y_i, and vice versa.
- Negative association, if larger-than-average X_i co-occurs with smaller-than-average Y_i, and vice versa.
- only interpret sign, not magnitude of association, given that covariance is scale-dependent

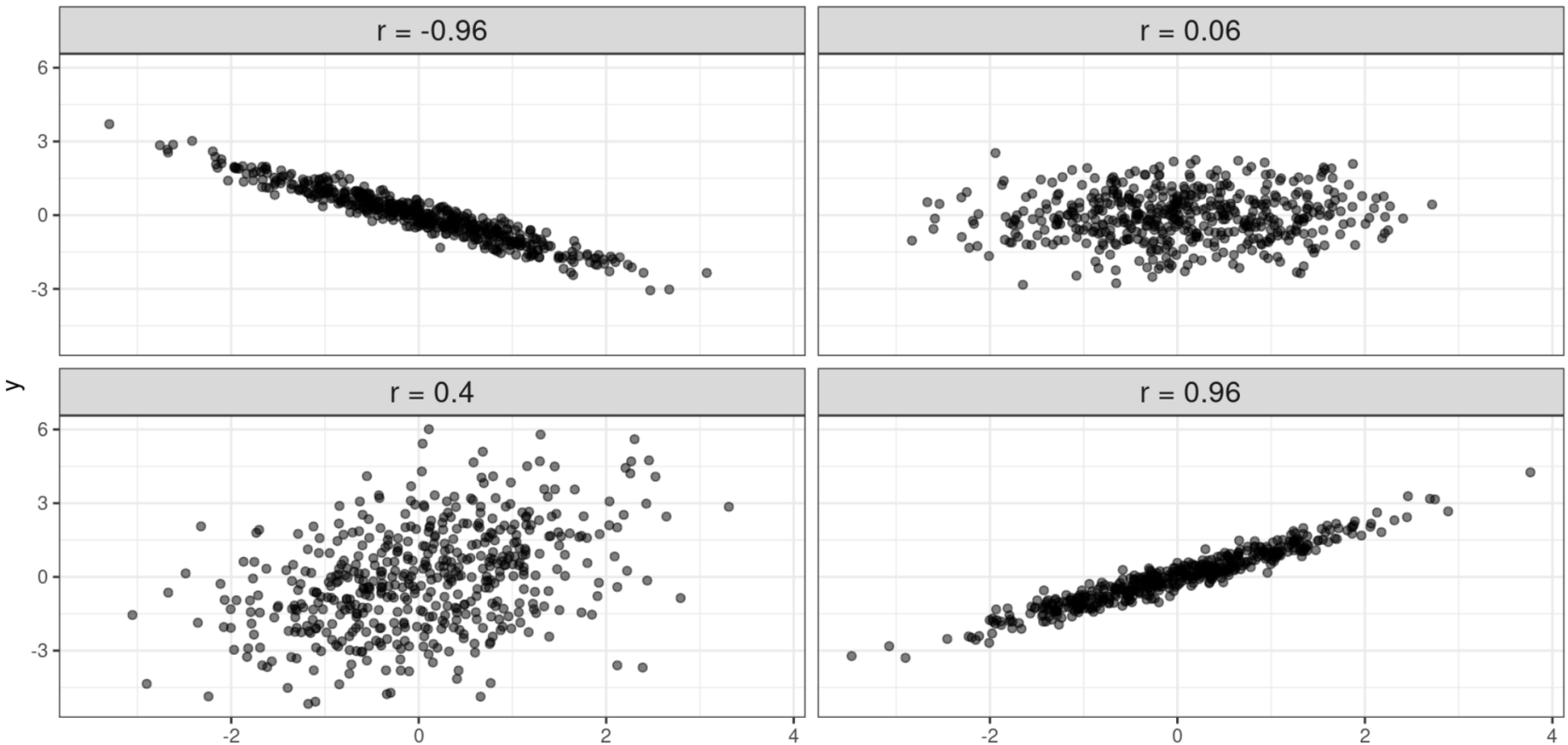
Correlation

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)
- **standardize covariance through dividing by product of standard deviations of the two variables,** $r_{xy} = \frac{covariance(XY)}{S_X S_Y}$
- ranges between -1 and 1, with 0=no association, the larger the absolute value, the stronger the association

Correlation

What is correlation?



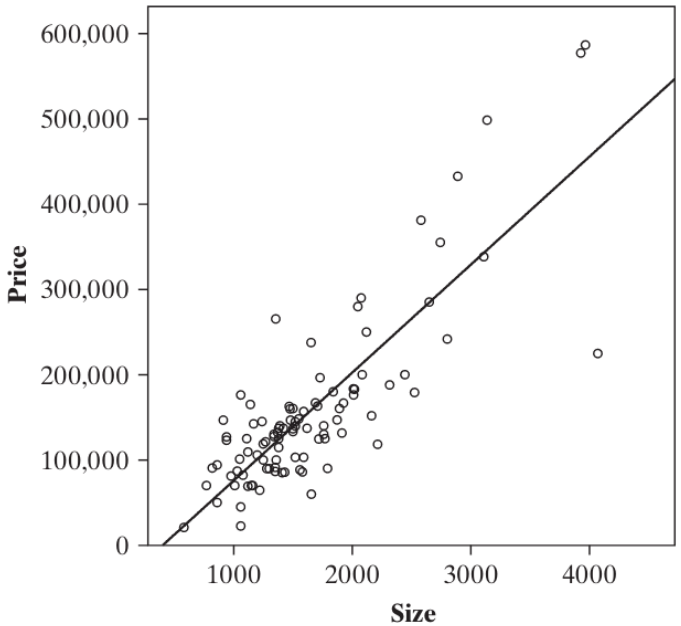
Correlation

How can we test the statistical significance of correlation?

- Null and alternative hypotheses:
  - there is no association between X and Y, $\rho_{xy} = 0$ (H_0)
  - there is an association between X and Y, $\rho_{xy} \neq 0$ (H_a)
- Test statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (in R)
- Test statistic: $t = \frac{r}{\sqrt{1-r^2/n-2}}$ (in Agresti 2018)

Correlation

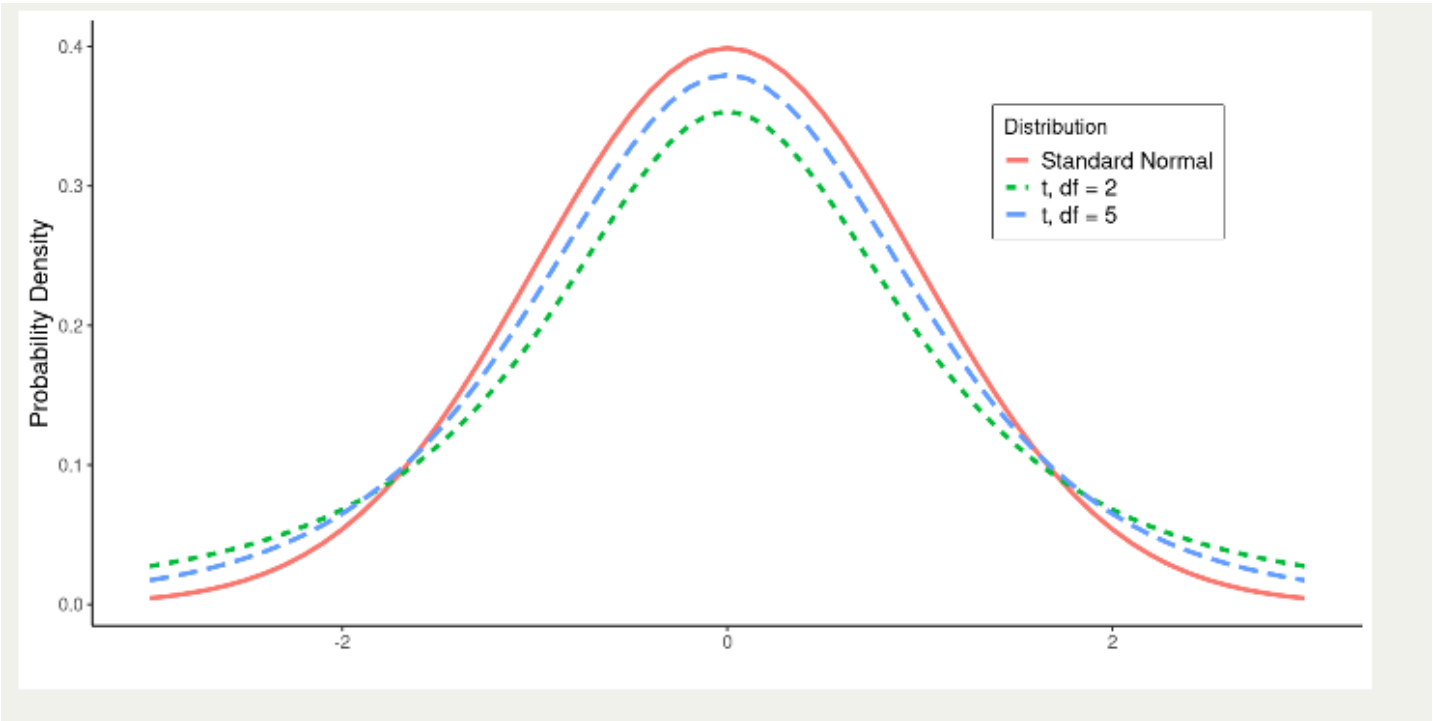
How can we test the statistical significance of correlation?



- Is there an association between house selling price and size (Agresti 2018, 278–283)? $r = 0.83378$
- $t = \frac{r}{\sqrt{1-r^2/n-2}} = \frac{0.834}{\sqrt{(1-0.695)/98}} = 14.95$
- 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? $\rightarrow$ Probability distribution

Correlation

How can we test the statistical significance of correlation?



What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0% (p=0.0001). $\rightarrow$ There is an association between house selling price and size

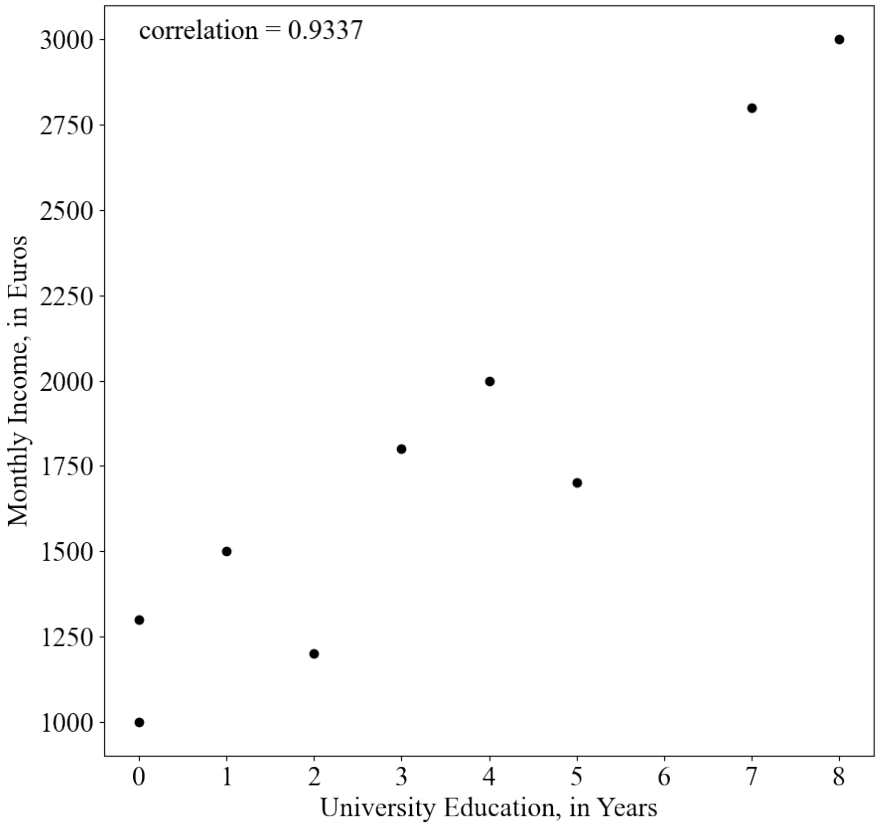
Shortcomings of correlation analysis

- 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



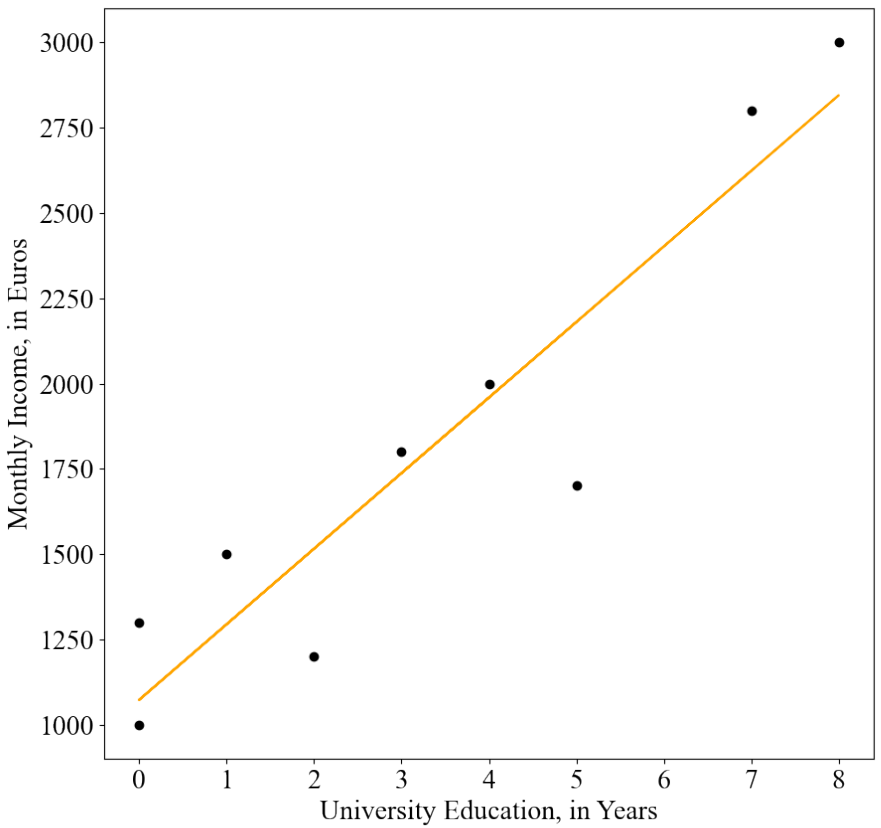
Just by looking at the plot, can you identify the straight line which best describes the joint variation between X and Y?

*This is fictional data.

Regression analysis

What is a linear regression model?

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



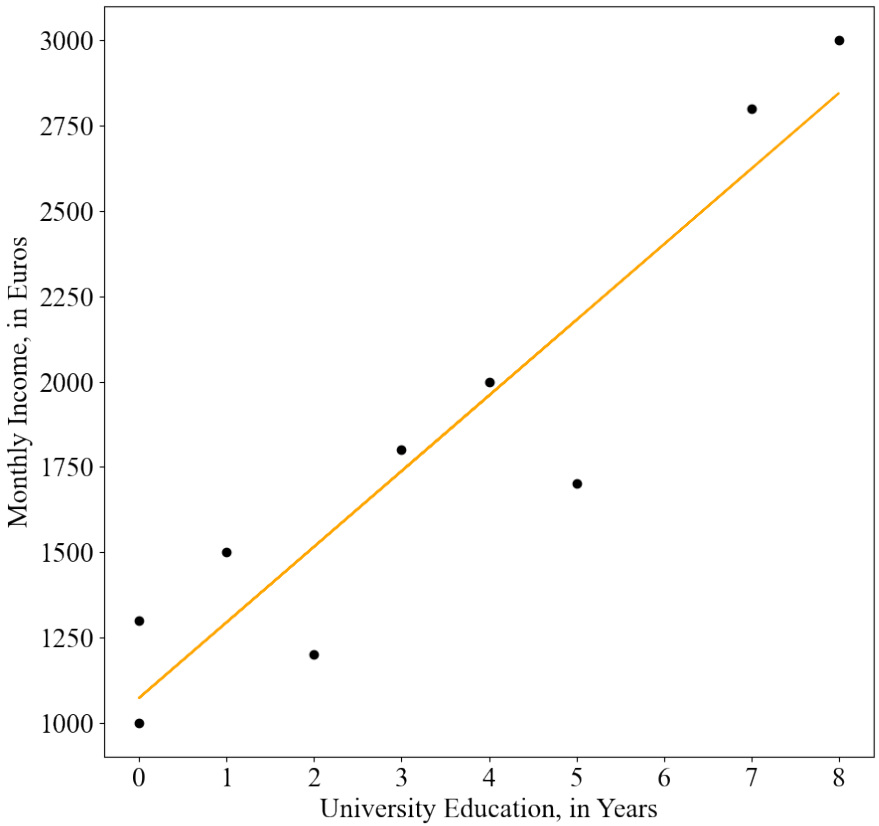
Regression analysis

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?

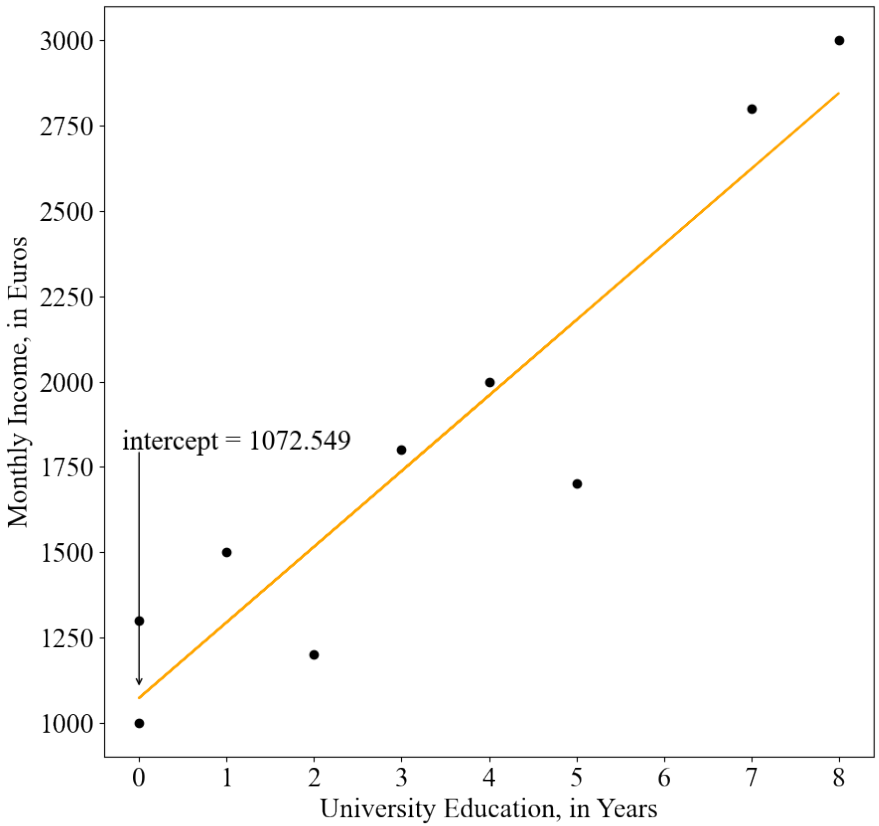


$$income = \alpha + \beta * education$$

$$income = 1072.5490 + 221.5686 * education$$

Regression analysis

What interpretations can we make? (intercept)

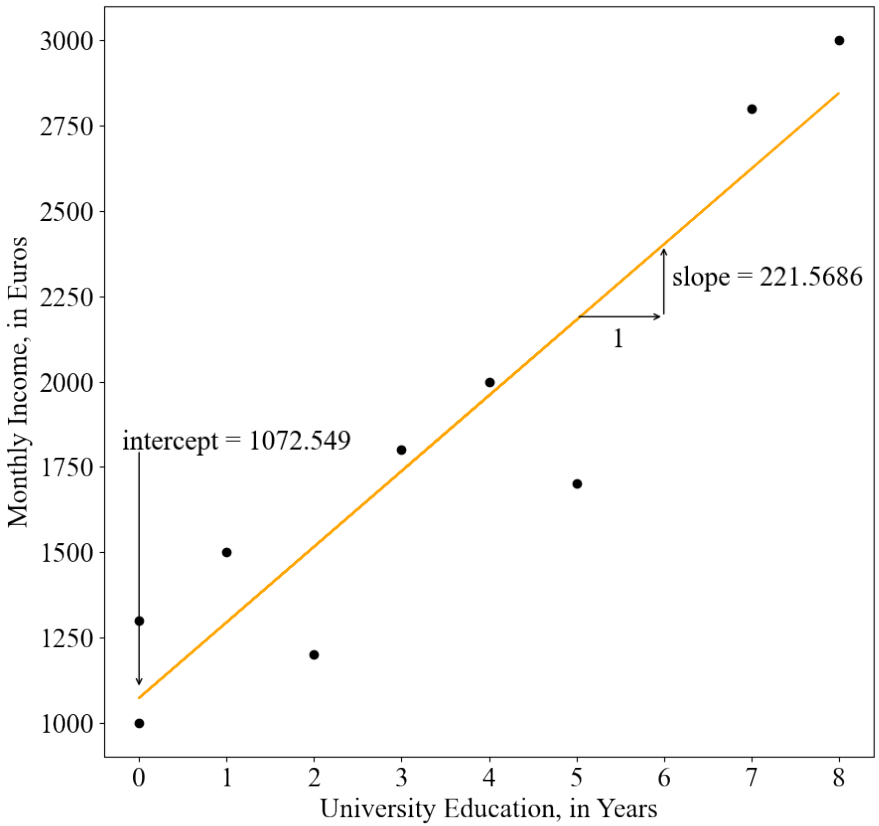


If an individual has a university education of 0 years, what income would we expect for that person?

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

Regression analysis

What interpretations can we make? (slope)

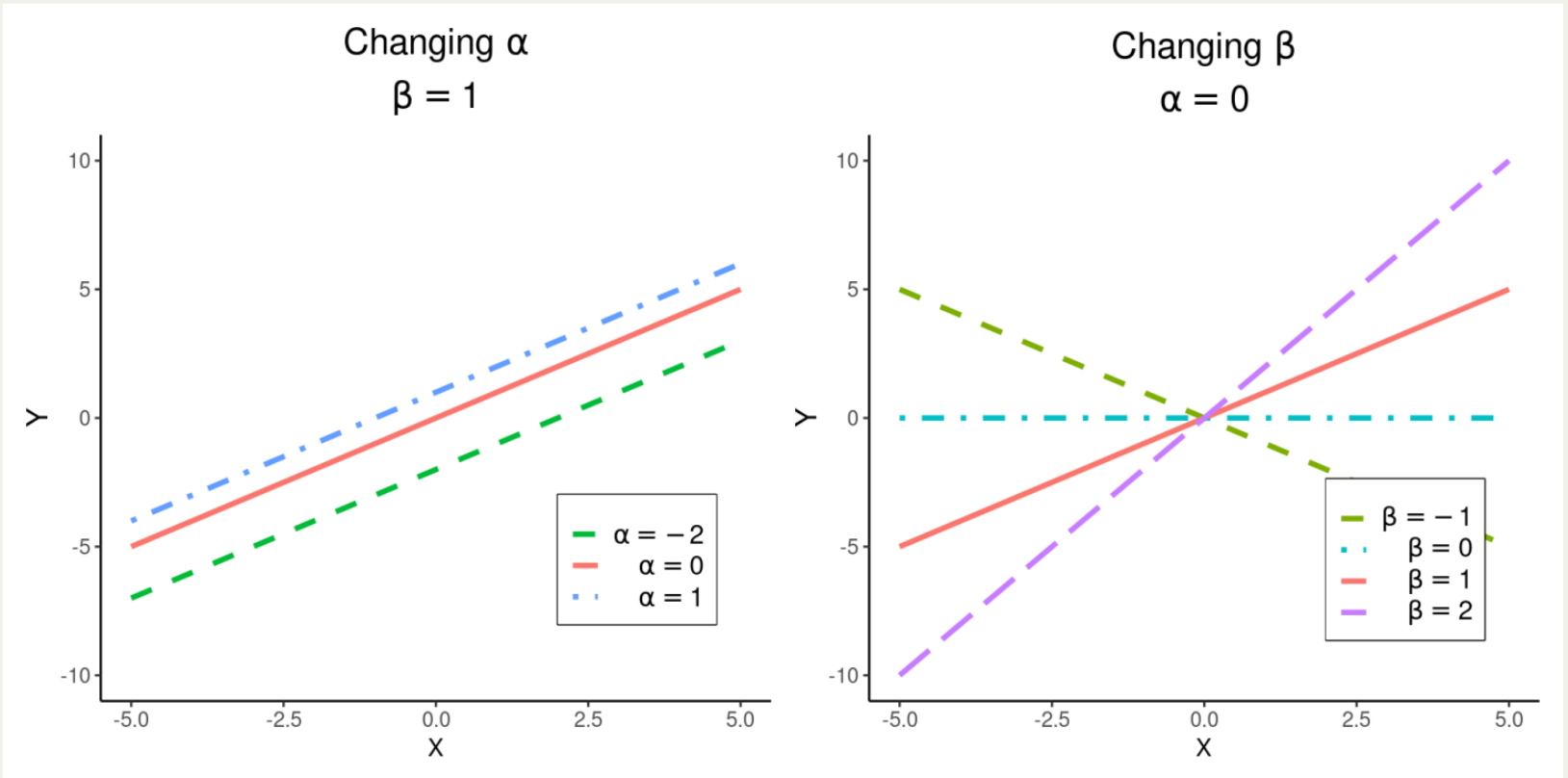


If the university education increases by one year, how much more Euros would we expect an individual to earn? $income = 1072.5490 + 221.5686 * 1 = 1294.1176$

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

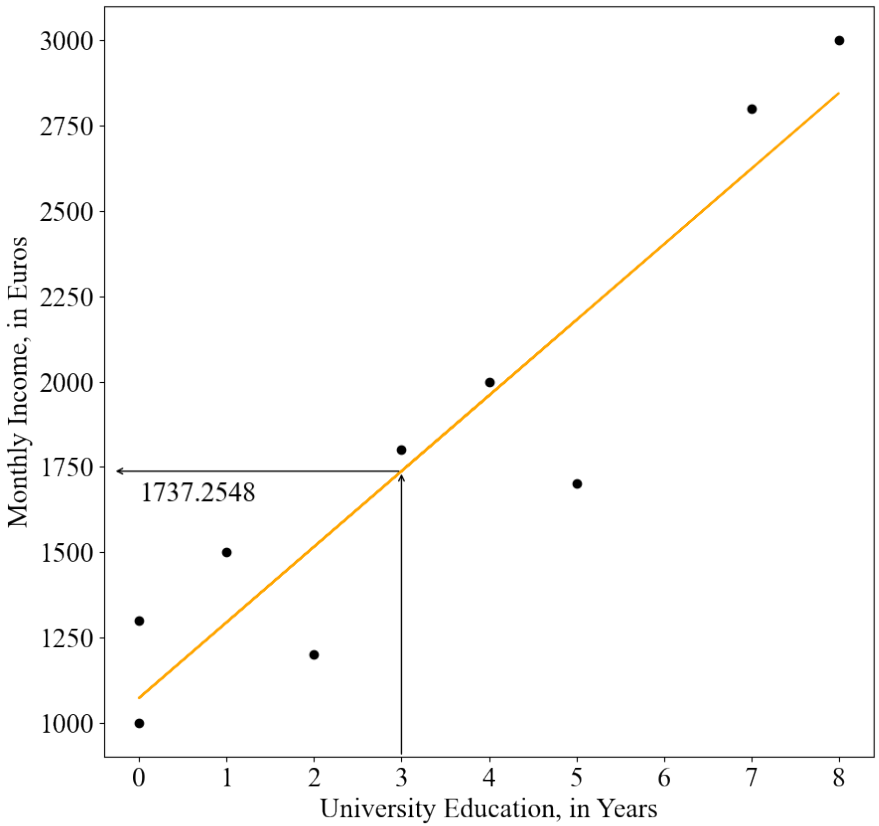
Regression analysis

Varieties of linear relationships



Regression analysis

What interpretations can we make? (expected value)

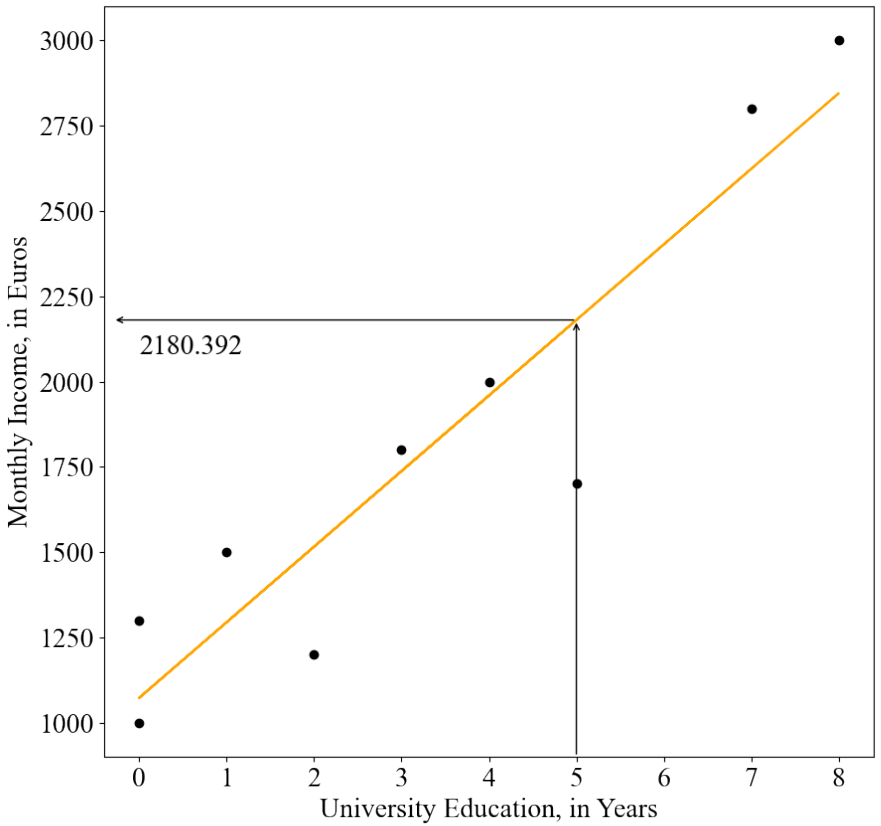


If an individual has 3 university education years, what income would we expect for that person?

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

Regression analysis

What interpretations can we make? (residual)



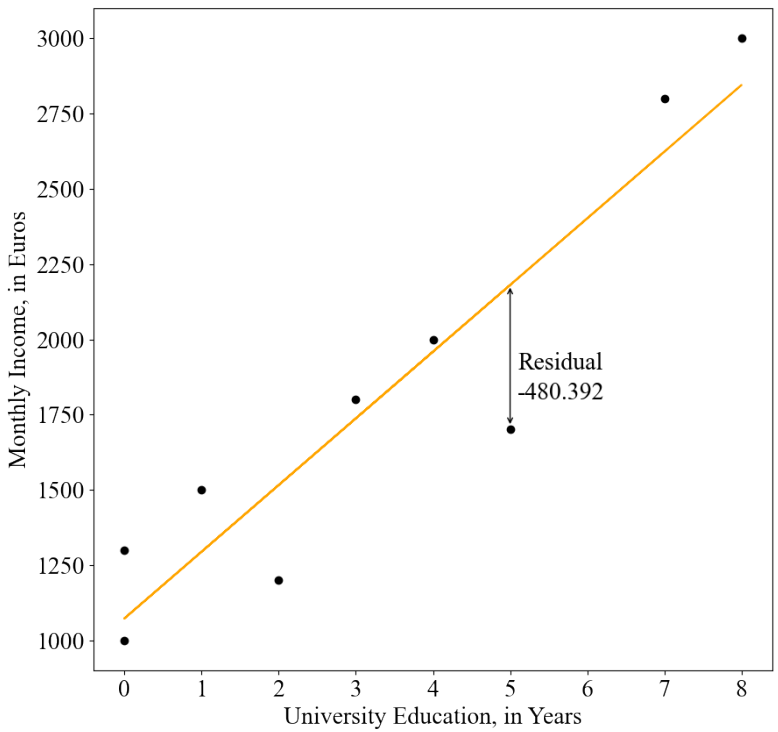
$income = 1072.5490 + 221.5686 * 5 = 2180.392$

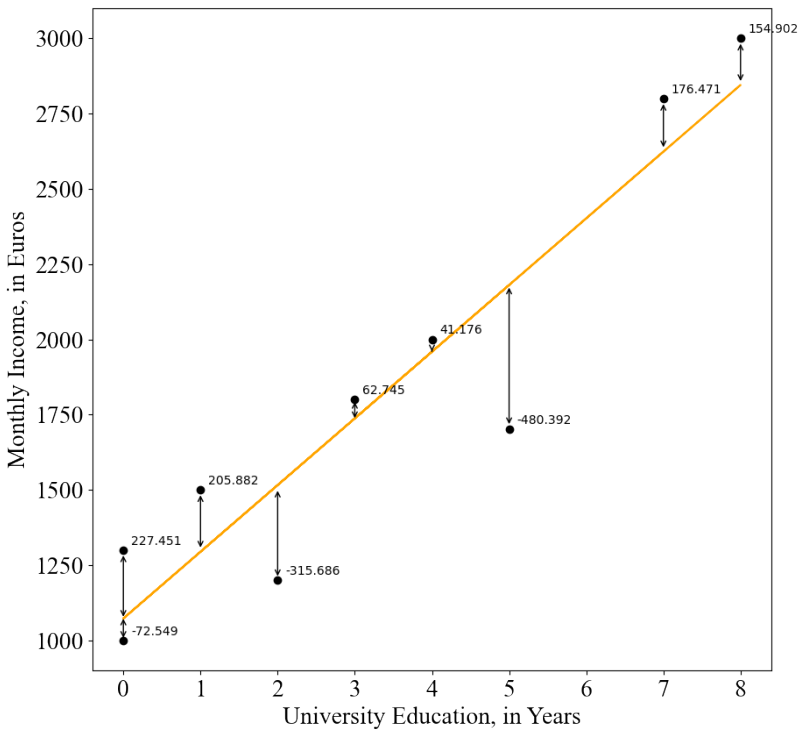
Residual = Actual – Predicted

Residual = 1700 – 2180.392 = –480.392

Regression analysis

What interpretations can we make? (residuals)





Ordinary least squares (OLS)

How are intercept and slope estimated?

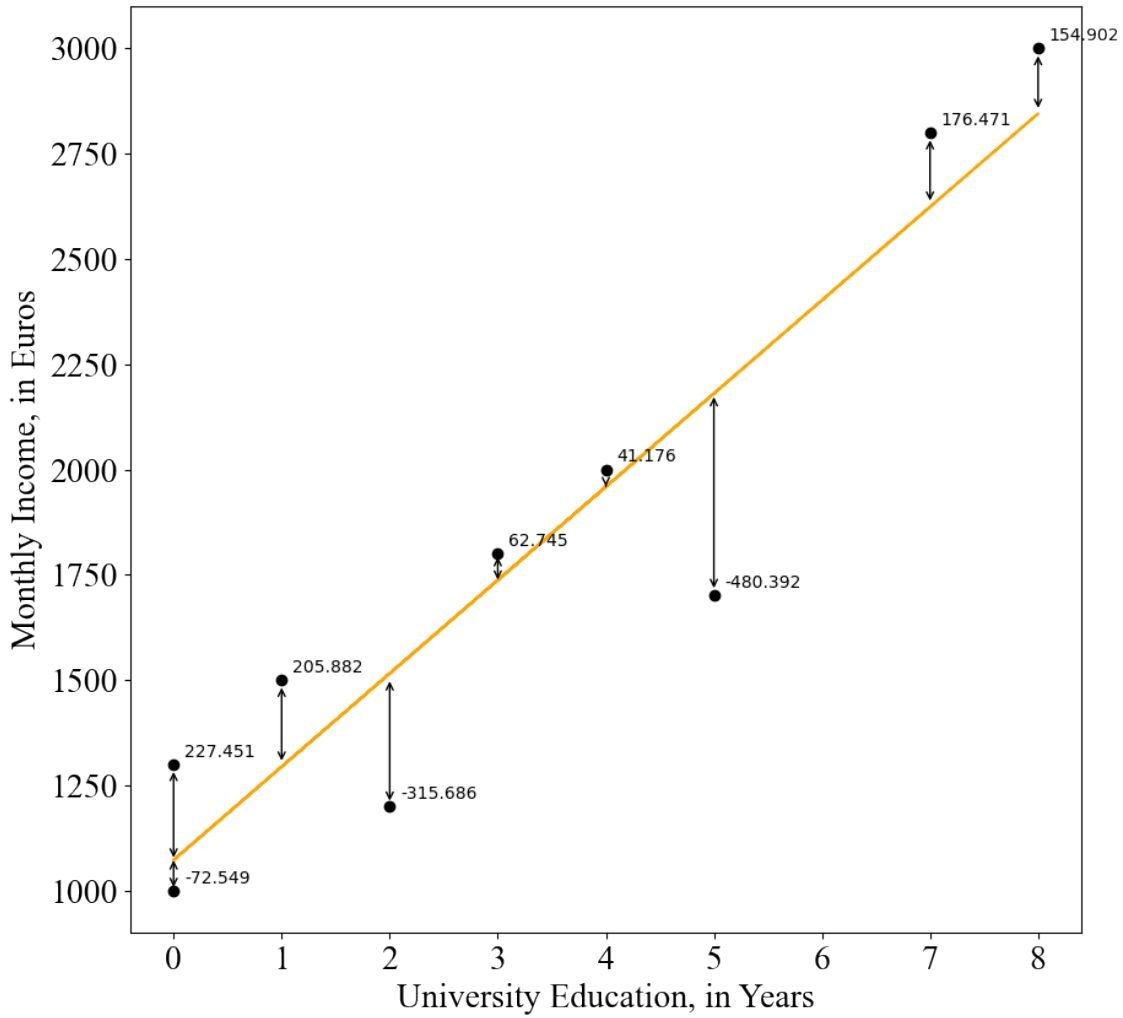
Ordinary least squares (OLS)

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.
- $SSE = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{\alpha} - \hat{\beta}X_i))^2$ $\rightarrow$ minimize this!

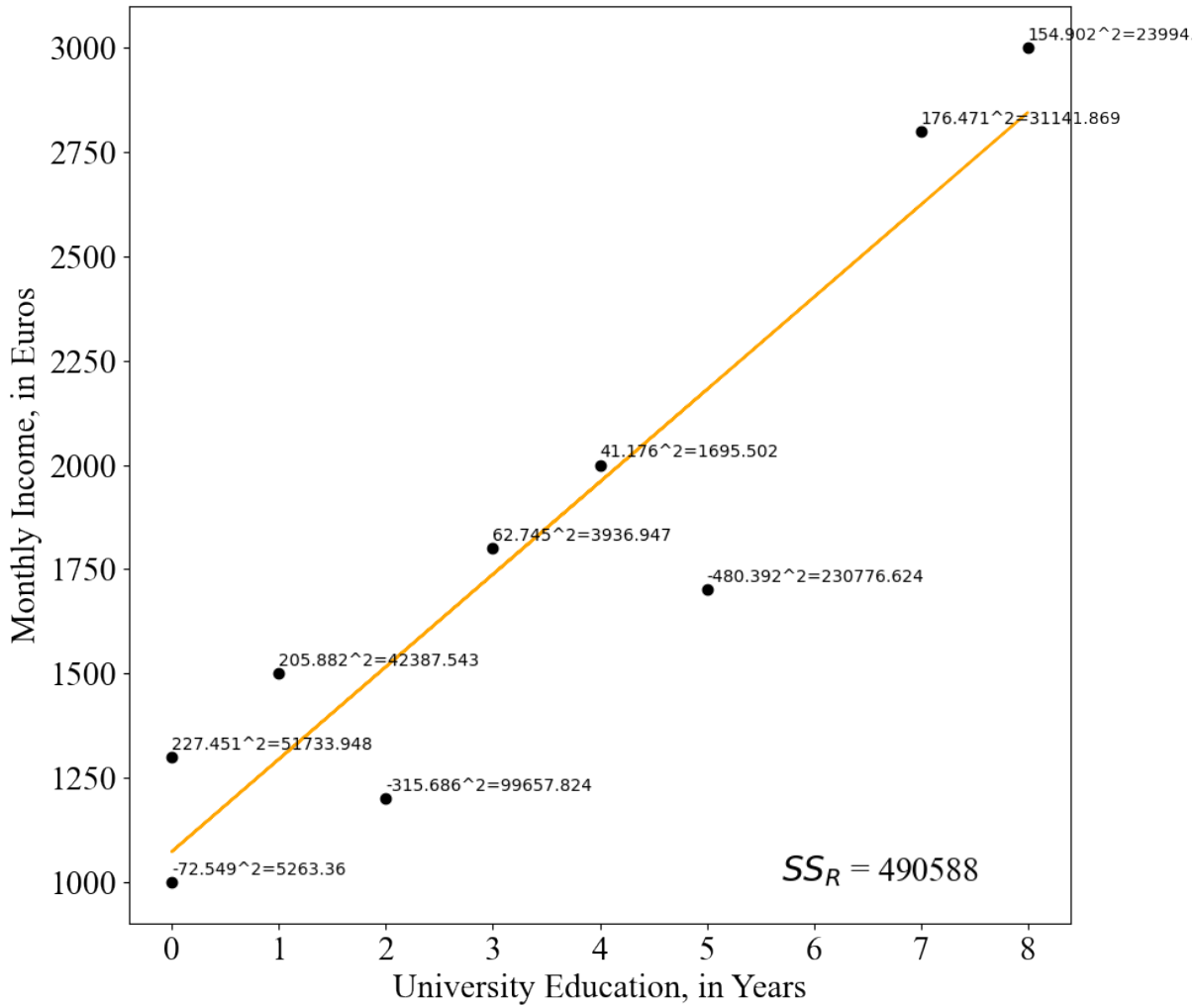
Ordinary least squares (OLS)

How are intercept and slope estimated?



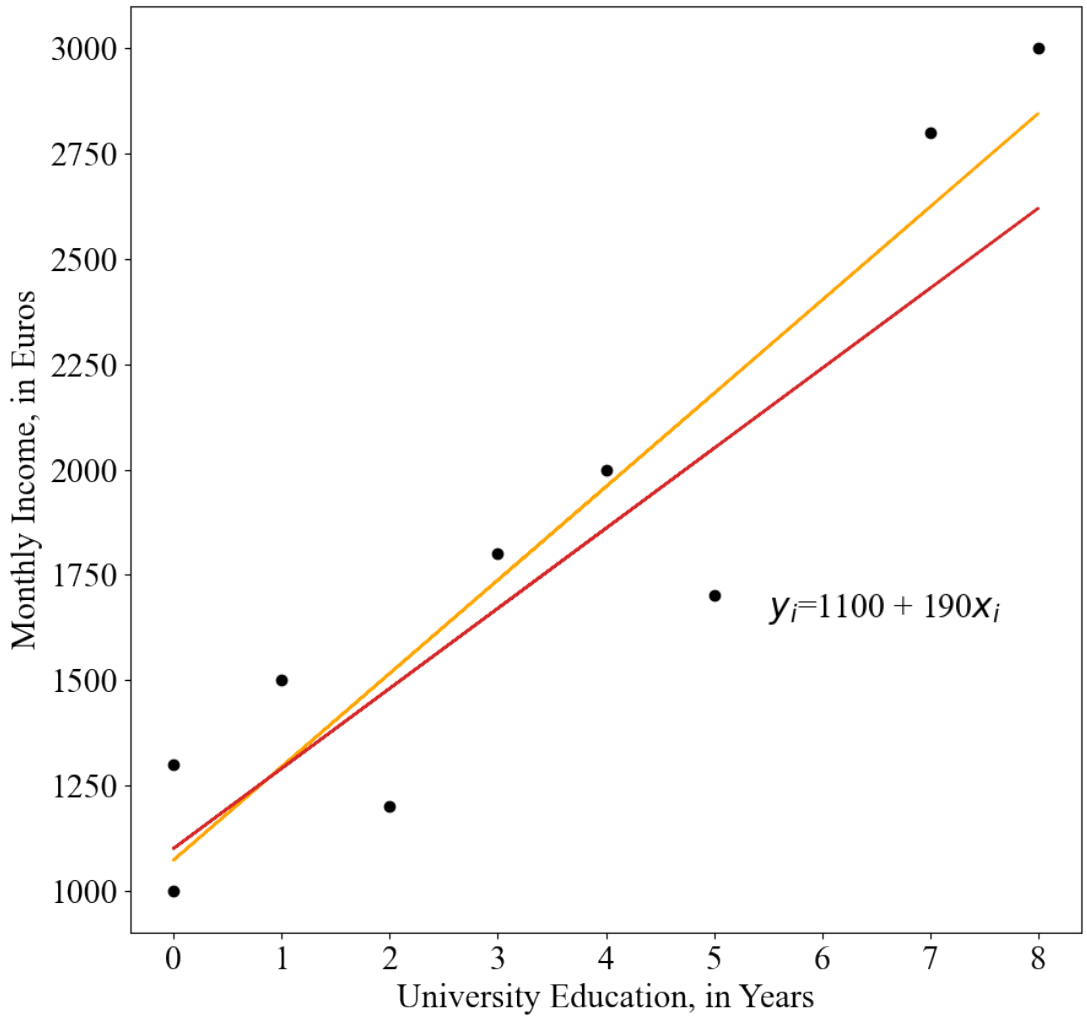
Ordinary least squares (OLS)

How are intercept and slope estimated?



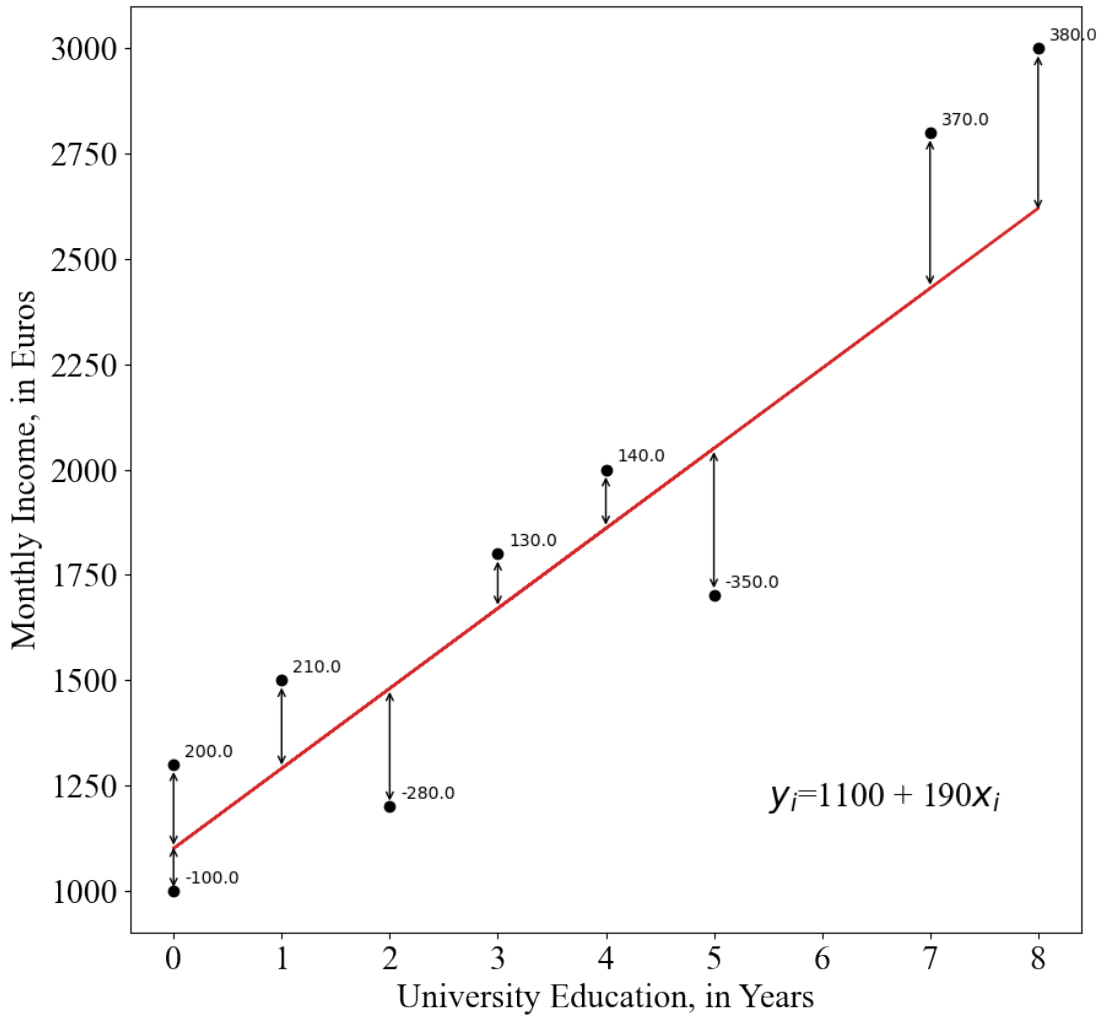
Ordinary least squares (OLS)

How are intercept and slope estimated?



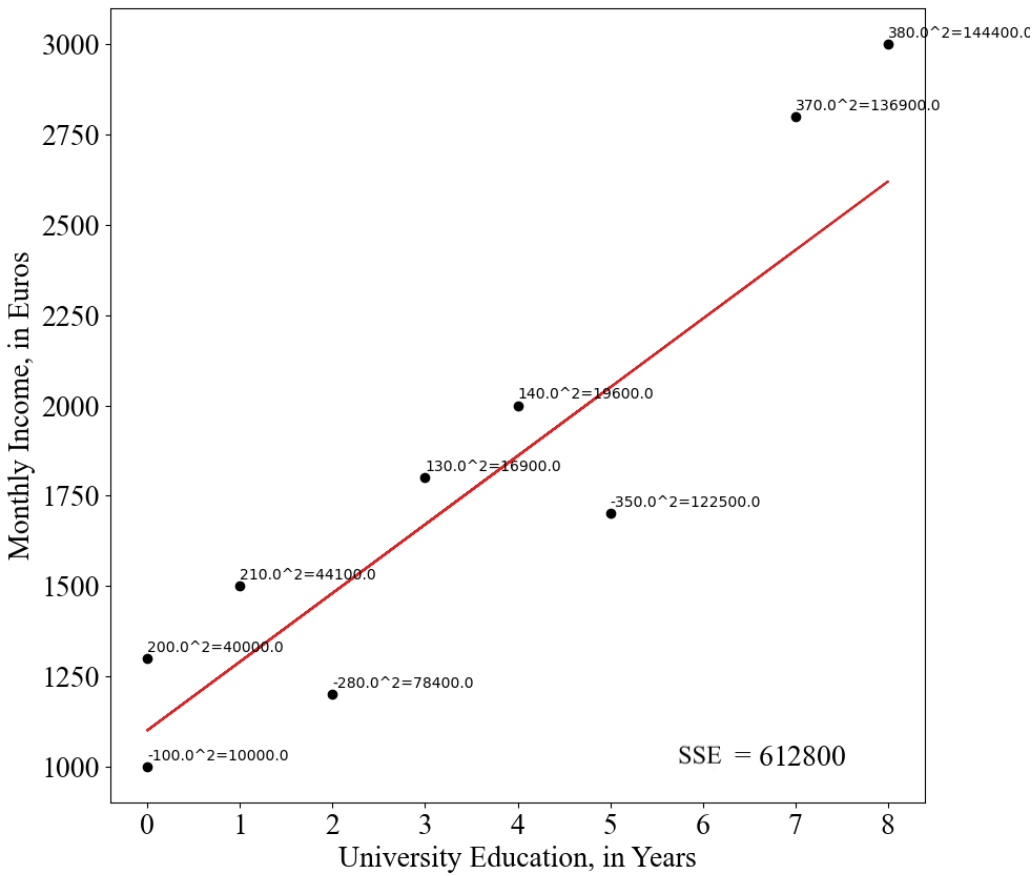
Ordinary least squares (OLS)

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

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

* ϵ_i is normally distributed $\rightarrow$ needed for inference
* $E(\epsilon_i) = 0$, no bias $\rightarrow$ violated if error is not random, but correlated with omitted variable
* ϵ_i has constant variance σ^2 (Homoscedasticity $\leftrightarrow$ Heteroscedasticity)
* No autocorrelation, "Autocorrelation occurs when the stochastic terms for any two or more cases are systematically related to each other".
* X values are measured without error

(Kellstedt and Whitten 2018, 190–194)

Assumptions of linear regression

Assumptions about the model specification, $Y_i = \alpha + \beta X_i + \epsilon_i$

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

* 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)

Inference about the slope

What is the t-test for the slope of a regression line?

Inference about the slope

What is the t-test for the slope of a regression line?

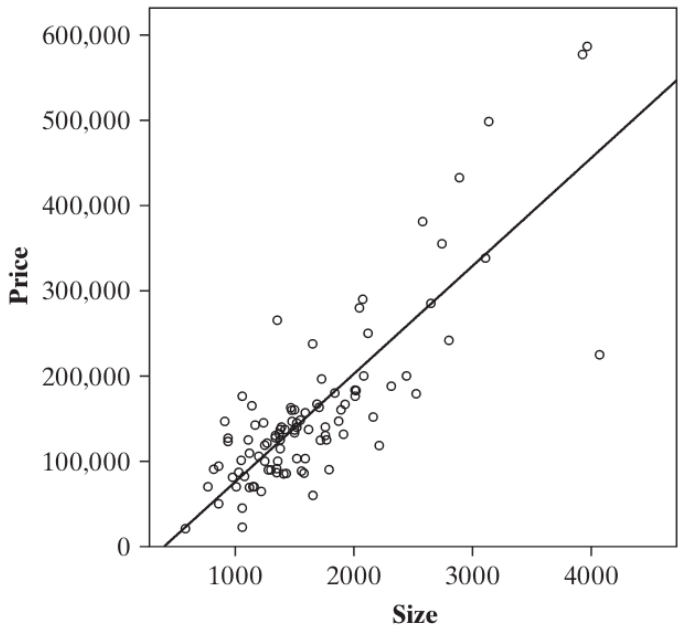
- Null and alternative hypotheses:
  - there is no association between X and Y, $\beta = 0$ (H_0)
  - there is an association between X and Y, $\beta \neq 0$ (H_a)
- Test statistic: "measures the number of standard errors between the estimate and the H_0 value" (Agresti 2018, 192).

$$t = \frac{\textit{Estimate of parameter} - \textit{Null hypothesis value of parameter}}{\textit{Standard error of estimate}}$$

$t = \frac{\hat{\beta} - \beta_{H_0}}{se_{\hat{\beta}}} = \frac{\hat{\beta}}{se_{\hat{\beta}}}$, H_0 assumes $\beta = 0$

Inference about the slope

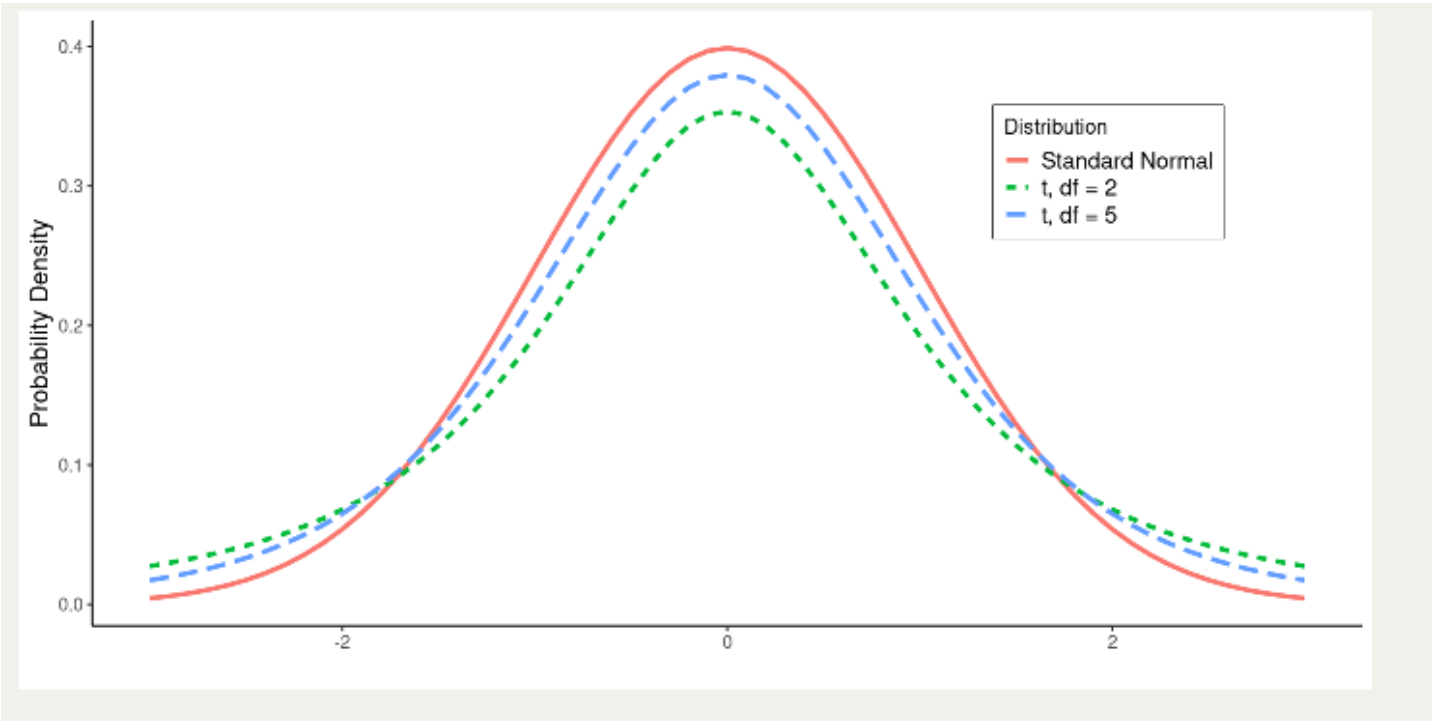
What is the t-test for the slope of a regression line?



- Is there an association between house selling price and size (Agresti 2018, 278–280)? $Price = 50,926.2 + 126.6 * Size$
- $t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} = \frac{126.6}{8.47} = 14.95$
- 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? $\rightarrow$ Probability distribution

Inference about the slope

What is the t-test for the slope of a regression line?

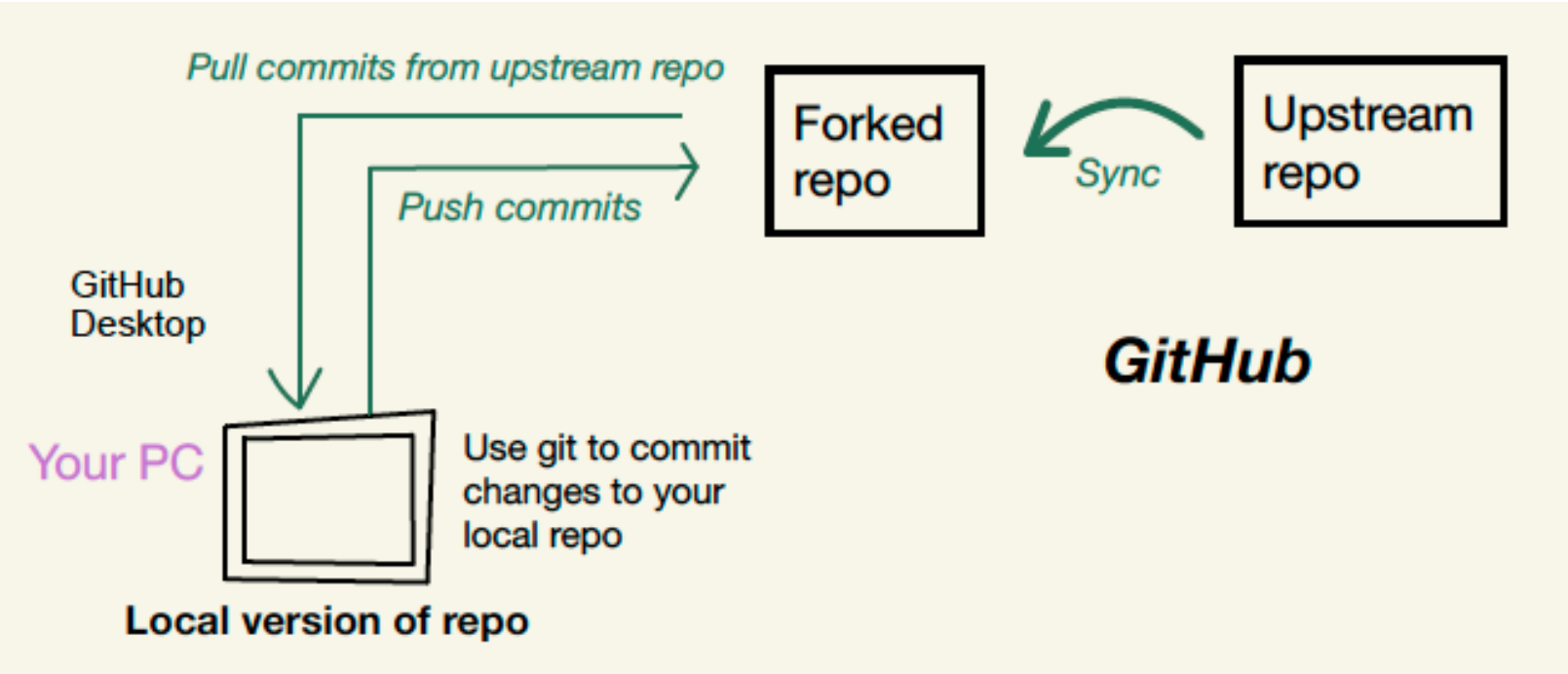


What is the conclusion? P-value < 0.05, We can reject H_0 with an error probability (p-value) of essentially 0% (< 0.0001). $\rightarrow$ There is an association between house selling price and size

Software check

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

Software check



1. Synchronize fork
2. Fetch origin

References I

Agresti, Alan. 2018. *Statistical methods for the social sciences.* Harlow: Pearson.

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.