Data Analytics in R
Session 13

Maria Kunevich

Assignments

Assignment	Date of assignment	Deadline (midnight 23:59)
HW1	22 Sept 2022	28 Sept 2022
HW2	29 Sept 2022	5 Oct 2022
HW3	6 Oct 2022	12 Oct 2022
HW4	13 Oct 2022	19 Oct 2022
HW5	20 Oct 2022	2 Nov 2022
Paper summary	20 Oct 2022	20 Nov 2022
HW6	3 Nov 2022	9 Nov 2022
HW7	17 Nov 2022	27 Nov 2022
HW8	1 Dec 2022	11 Dec 2022
Project	20 Oct 2022	
Interim pitch		17 Nov 2022
Project Presentations		15 Dec 2022
Submission of Final report		29 Dec 2022

Comments for the assignment and your projects

- Example reports (for your information and example of code)

  Your research questions are important! (Are they worth answering? Is it possible to break down the questions to tailor them to the kind of answers you can get with statistical tests? Think carefully especially for correlation tests, are these relationships worth investigating? Remember about **spurious correlations**

- Project guidelines, any questions I should include in particular?

- A link to Google Presentation for next session:

- https://docs.google.com/presentation/d/1lsoJqmbWrztm-IJb3HZOJLEeeeZrGr7-iGLUKwxc0sc/edit?usp=sharing

Types of statistical tests

- Parametric
- Non-Parametric

Forms of statistical tests

- Tests of Comparison
- Tests of Correlation
- Tests of Prediction (regression tests)

Simple Linear Regression

- A simple linear regression is **a statistical model** that analyses the relationship between two quantitative variables: *a response variable* (often called y) and *an explanatory variable* (often called x)
- Linear regression models are a useful tool **for predicting a quantitative response**
- Linear regression assumes that there exists **a linear relationship** between the response variable and the explanatory variables, i.e. you can fit a line between the two

Source: https://www.datacamp.com/tutorial/linear-regression-R

Examples of linear regression model

If you are interested in the relationship between income and satisfaction with life, you can collect data from people (income in euros) and ask them to rate their satisfaction with life on a 10.point scale

Independent variable (explanatory): income

Dependent variable (response): life satisfaction

RQ: What is the relationship between people's income and satisfaction with life?

What type of variables do we have? Quantitative: continuous and discrete

What kind of statistical test can you run? Linear regression model

Examples of linear regression model

You are interested if there is a relationship between the amount of time spent on homework assignment and the final score the student gets. You can collect data from students (time in hours) and the final score.

Independent variable (explanatory, x): hours

Dependent variable (response, y): score

RQ: is there a relationship between the amount of hours students spend on their homework assignments and the final score they get?

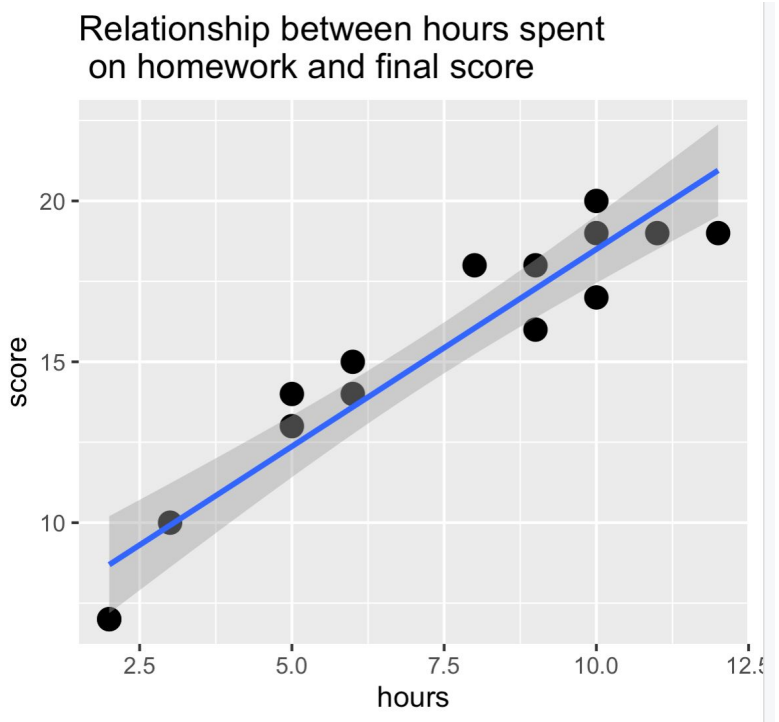
What type of variables do we have? Quantitative: continuous and discrete

What kind of statistical test can we run? Linear regression model

Descriptive statistics

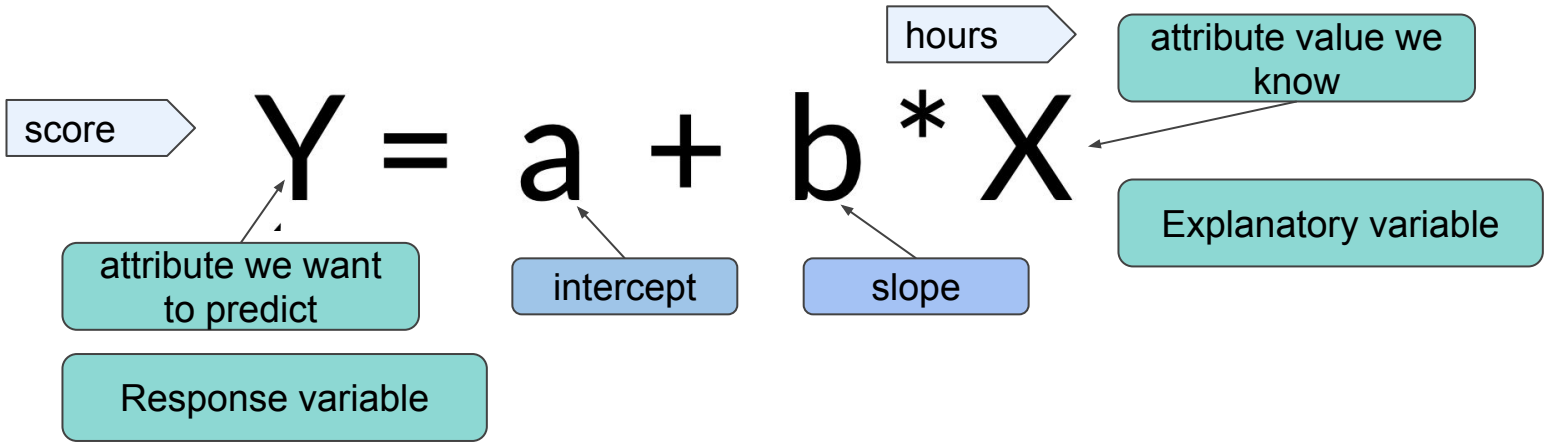
Relationship?

r =0.95

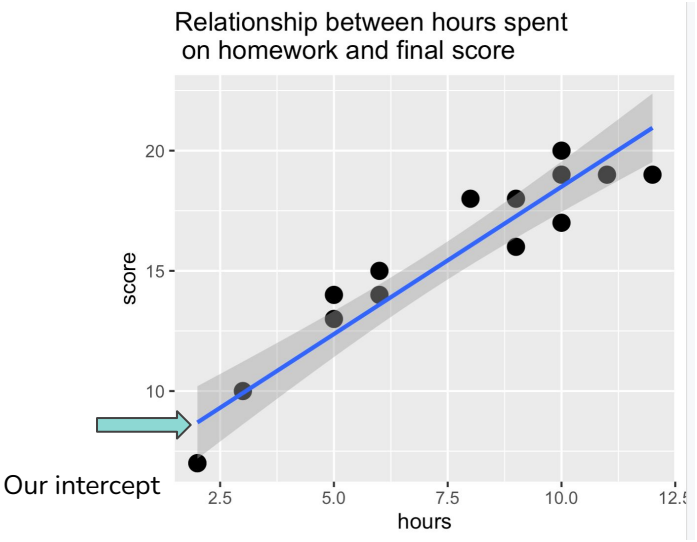


Simple linear regression model: example

Can we predict the final score if someone tells us about the amount of time they spent on homework?

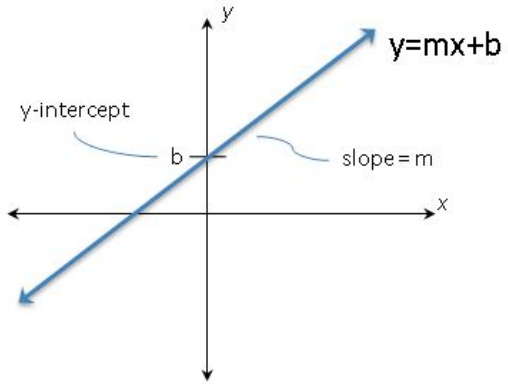


The Slope of the regression line



Steepness of **the line of best fit**

Interpretation: it will tell you how dependent variable change when you change the independent variable



The intercept

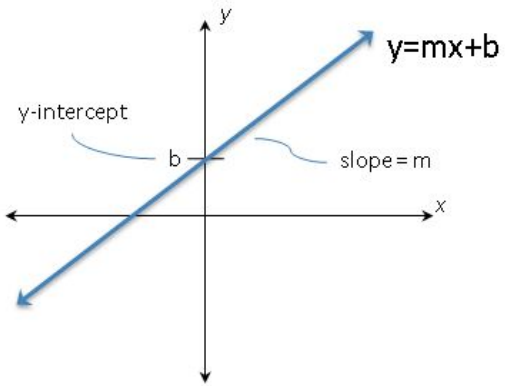
The expected **mean value** of *y* when *x*=o

Our example: intercept 6.24, interpretation: the mean score is 6.24 when 0 hours are spent on homework

=> intercepts do not always make sense to interpret (weight / height)

the intercept term in the model is important as **it helps to use the model for making predictions**

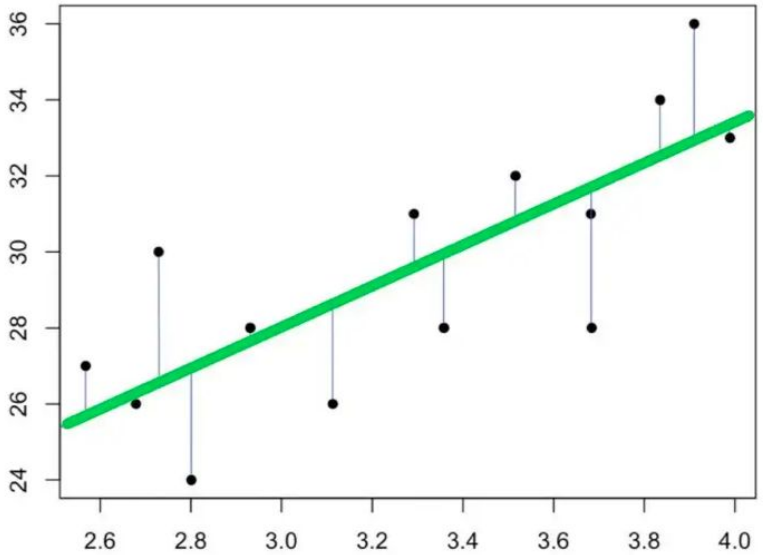
BUT - it is possible for the intercept to have no meaningful interpretation for a certain model



Estimates

We don’t have enough information or data to know the exact equation = > generate **estimates** for both **the slope** and **the intercept**

Estimates are generated through **the ordinary least squares method**, i.e. the regression model finds the line that fits the points in such a way that it minimizes the distance between each point and the line (minimizes the sum of the squared differences between the actual values and the predicted values)



Source: https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3

Assumptions for linear regression

Simple linear regression is **a parametric test** -> assumptions:

1. **The relationship** between the independent and dependent variable **is linear**: the line of best fit through the data points is a straight line
2. Homogeneity of variance (**homoscedasticity**): the size of the error in our prediction doesn't change significantly across the values of the independent variable, i.e. errors must have constant variance
3. **Independence of observations**: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations
4. **Normality:** the data follows a normal distribution, but we are more interested in errors: **errors must be normally distributed**

Source: https://www.scribbr.com/statistics/simple-linear-regression/

Steps for linear regression analysis

Step 1: preparing the dataset and EDA

Step 2: visualisation (outliers, normal distribution)

Step 3: performing simple linear regression + interpretation

Step 4: creating residual plots

Step 5: reporting

In R:

lmer package

lm [target/response]~[predictor]
model <- lm(y~x, data=df)

Call the model:

summary(model)

Interpretation - model fit

```
Call:
lm(formula = score ~ hours, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9501 -1.3858  0.4042  1.0656  1.9528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.2415     0.9015   6.924 1.05e-05 ***
hours         1.2257     0.1145  10.706 8.11e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.413 on 13 degrees of freedom
Multiple R-squared:  0.8981,    Adjusted R-squared:  0.8903
F-statistic: 114.6 on 1 and 13 DF,  p-value: 8.107e-08
```

1. The model
2. Residuals
3. Coefficients

Residuals - model fit

Residuals report the difference between the actual observed response values (score for homework) and the response values that the model predicted

The Residuals section of the model consists of 5 summary points: we are interested in a symmetrical distribution across these points on the mean value zero (0). If these distributions we can assess how well the model fits the data: the more symmetrical, the better the fit

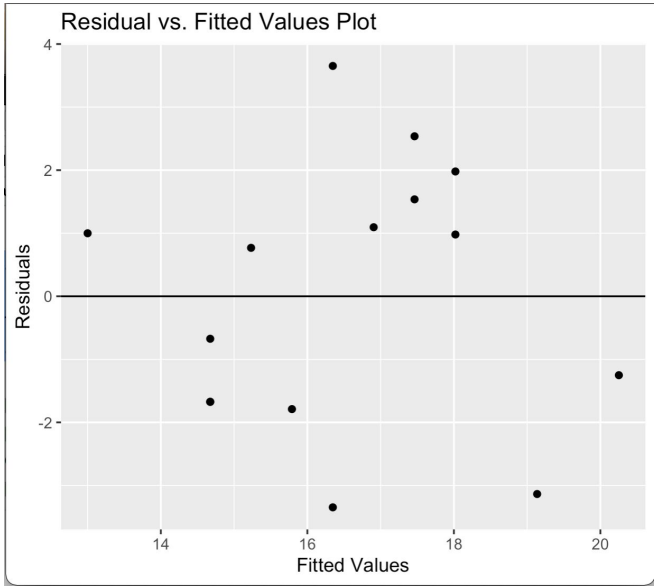
```
Residuals:
    Min      1Q  Median      3Q     Max
-1.9501 -1.3858  0.4042  1.0656  1.9528
```

Creating a Residuals plot

Residual plots are used to assess whether or not the residuals in a regression model are **normally distributed**

The x-axis displays **the fitted values** (a fitted value is a statistical model's prediction of the mean response value when you input the values of the predictor)
The y-axis displays **the residuals**

In our dataset the residuals appear to be randomly scattered around zero with no clear pattern, which indicates that **the assumption of homoscedasticity is met**.
In other words, the coefficients of the regression model should be trustworthy and we don't need to perform a transformation on the data



Creating a quantile-quantile plot (QQ-plot) for residuals

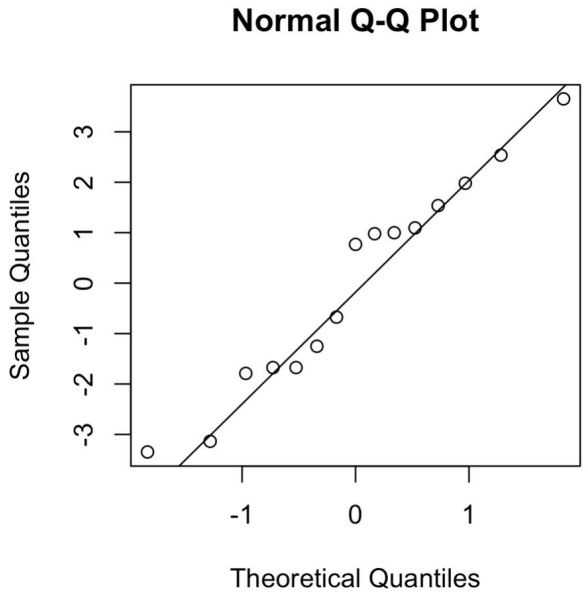
Normal Q-Q Plot: this plot is used to assess if the residuals **are normally distributed**

We are checking if the data points are closely following the straight line at a 45% angle upwards (left to right)

In out dataset the points approximately fall on the line

=> the relationship between the theoretical percentiles and the sample percentiles is approximately linear

The normal probability plot of the residuals suggests that the error terms are normally distributed for our dataset



Coefficients

The coefficient Estimate contains two rows: the first row is the intercept

The value for **the intercept** term in our model is 6.24. This means the average score is 6.245 when the number of hours studied is equal to zero. Remember that intercepts are not always meaningful!

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.2415     0.9015   6.924 1.05e-05 ***
hours         1.2257     0.1145  10.706 8.11e-08 ***
---
```

We can use the intercept to predict our response variable by using it in the regression equation: **y(score) = 6.24 + 1.23*(x,hours) ± 0.0078**

Coefficients

The **Estimate** column is **the estimated effect**, or the regression coefficient or r^2 value. The number in our table (1. 225) tells us that for every one unit increase (i.e. one hour spent on homework) there is a corresponding 6.24-unit increase in the final score.

The **Std. Error** column displays **the standard error** of the estimate. This number shows how much variation there is in our estimate of the relationship between time spent on homework and the final score.

The **t-value** column displays **the test statistic**. Unless you specify otherwise, the test statistic used in linear regression is the *t* value from a two-sided t-test. The larger the test statistic the less likely it is that our results occurred by chance.

The **p-value** tells us how likely we are to see the estimated effect of time spent on homework and the final score if **the null hypothesis** of no relationship were true. Because the **p value** is very low ($p < 0$), we can reject the null hypothesis and conclude that the relationship is **statistically significant**.

Source: https://www.scribbr.com/statistics/simple-linear-regression/

Reporting the results

We report the estimated effect (i.e. the regression coefficient), standard error of the estimate, the p value.

Our example: we found a significant relationship ($p < 0.001$) between hours spent on the homework assignment and the final score ($R_2 = 0.8903 \pm 0.0078$), with a 1.22-unit increase in the final score for every one hour increase in time spent on homework.

Revision: interview questions for R

- What are the different data structures in R? Briefly explain about them.
- What is Rmarkdown? How can we use it?
- What are packages and libraries in R? How do you use them?
- Name some packages in R, which you use
- What is a factor? How would you create a factor in R?
- What are the different import functions in R?
- How would you check the distribution of a categorical variable in R?
- How would you rename the columns of a dataframe?
- How would you find the number of missing values in a dataset and remove all of them?
- Can you write and explain some of the most common syntax in R?
- Name some functions available in “dplyr” package

Revision: interview questions for R

- ❏ Give examples of “rbind()” and “cbind()” functions in R
- ❏ How would you create a scatterplot using ggplot2 package?
- ❏ What are some advantages of R?
- ❏ What are the disadvantages of R?
- ❏ What are the objects you use most frequently?
- ❏ What are some of your favorite functions in R? How do you decide which package to choose to solve a problem using R?
- ❏ What are the steps to build and evaluate a linear regression model in R?

Revision: some interview questions on data analysis and statistics

- ❏ What is the difference between Descriptive and Inferential Statistics?
- ❏ What is data cleaning?
- ❏ What is data wrangling?
- ❏ What are population and sample in Inferential Statistics, and how are they different?
- ❏ What is EDA?
- ❏ What are different types of data?
- ❏ What are the different types of variables or measurement levels?
- ❏ What kind of variables are there (dependent/independent/confounding)?
- ❏ What is the difference between the long format data and wide format data?
- ❏ Mention some techniques used for sampling. What is the main advantage of sampling?

Revision: some interview questions on data analysis and statistics

- What is a normal distribution?
- What is an outlier? How can outliers be determined in a dataset?
- What is a Sampling Error and how it can be reduced?
- What are the main measures used to describe the Central Tendency of data?
- In what cases the median is a better measure when compared to the mean?
- What are the main measures used to describe the Variability of data?
- What is the meaning of standard deviation?
- What Is the Confidence Interval?
- What is the meaning of the five-number summary in Statistics?
- What is the difference between the first quartile, the second quartile, and the third quartile?

Revision: some interview questions on data analysis and statistics

- ❏ What is skewness? What are left-skewed and right-skewed distributions?
- ❏ What is a Sampling Error and how it can be reduced?
- ❏ What is correlation?
- ❏ What types of variables are used for Pearson's correlation coefficient?
- ❏ What are the steps in Hypothesis testing?
- ❏ What is an alternative hypothesis?
- ❏ What is the meaning of degrees of freedom (DF) in statistics?
- ❏ What is a p-value? How do we use it?
- ❏ What are the different types of Correlation?
- ❏ What is one sample t-test? What is a two-sample t-test?
- ❏ What is a Chi-square test?
- ❏ What do you understand about linear regression?