E0410 Fundamentals of Statistics for Scientific Data Using R

by Daria Sapunova, PhD student, RECETOX

daria.sapunova@recetox.muni.cz

Bohunice, D29, room 123

Theoretical part

We are approaching the end of the course

Date	Content
25/04	Correlation
02/05	Linear regression
09/05	Linear + multiple regression
16/05	Dean’s holiday
23/05	Wrap up or linear regression if we didn’t cover it up, questions

Final assignment

Working on a dataset provided **by a student or the teacher**.
The results should be presented as a **pdf file** consisting of:

Context	Words
a title	-
a short introduction regarding the topic with references	approx. 200 words
data description	approx. 100 words
statistical method description	approx. 200 words
results with data visualization and interpretation of the results	approx. 200 words
a short conclusion	approx. 100 words
Appendix with **the R script**	-

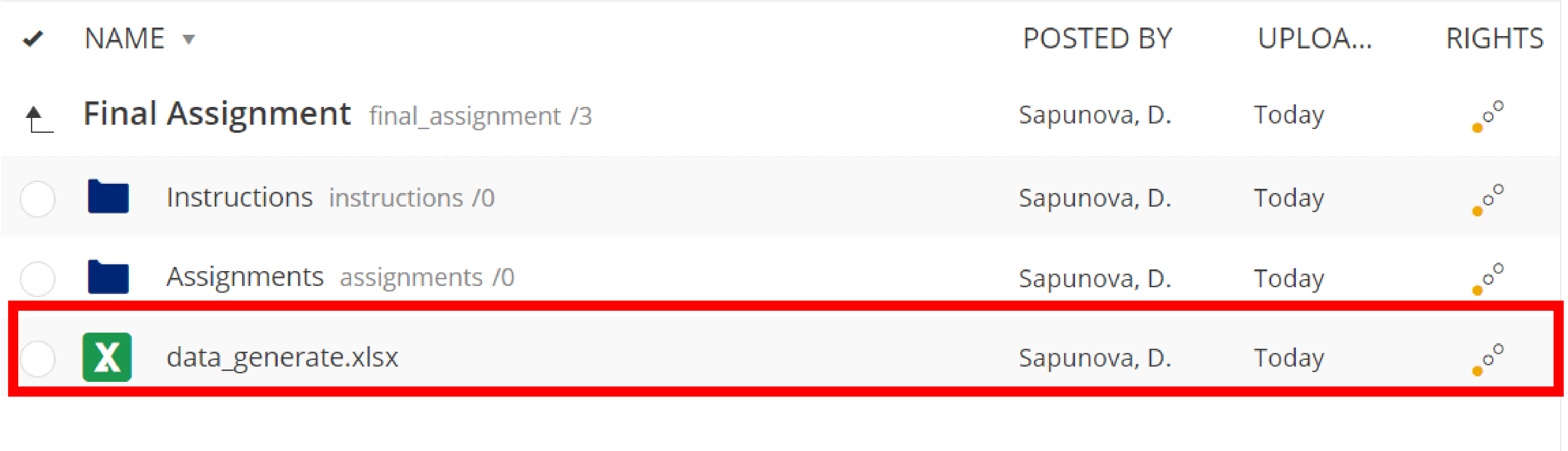
Final assignment

- ✓ **The examination period** is from 27.05.24 to 04.07.24
- ✓ **Please, upload pdf files with your assignment till 01.07.24 included**
- ✓ **The instruction with the example** will be uploaded **18.05.-19.05.24,** prepare **your questions by 23.05**. if you will have any
- ✓ **The generated data** will be sent **18.05.-19.05.24**

Who needs data?

✓**Till 03.05.24 included fill up the table**

Study materials -> Learning materials -> Final Assignment -> data_generate.xlsx



✓	NAME ▾	POSTED BY	UPLOA...	RIGHTS
	Final Assignment final_assignment /3	Sapunova, D.	Today	
○	Instructions instructions /0	Sapunova, D.	Today	
○	Assignments assignments /0	Sapunova, D.	Today	
○	data_generate.xlsx	Sapunova, D.	Today	

Check list

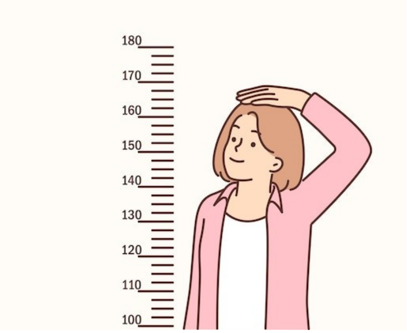
Deadline	Action
03/05	Fill up **data_generate.xlsx**
18/05-19/05	**The instruction** with the example and **the generated data will be uploaded** by the teacher
23/05	**Questions** regarding the instruction and data
01/07	Upload the final assignment

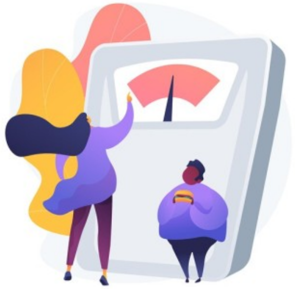
If you have questions - daria.sapunova@recetox.muni.cz

Repetition

Parametric and non-parametric: association between two numerical variables

We have: two numerical continuous variables





Height
Weight

association

Salary
Productivity

association

Employment
Inflation

association

Correlation

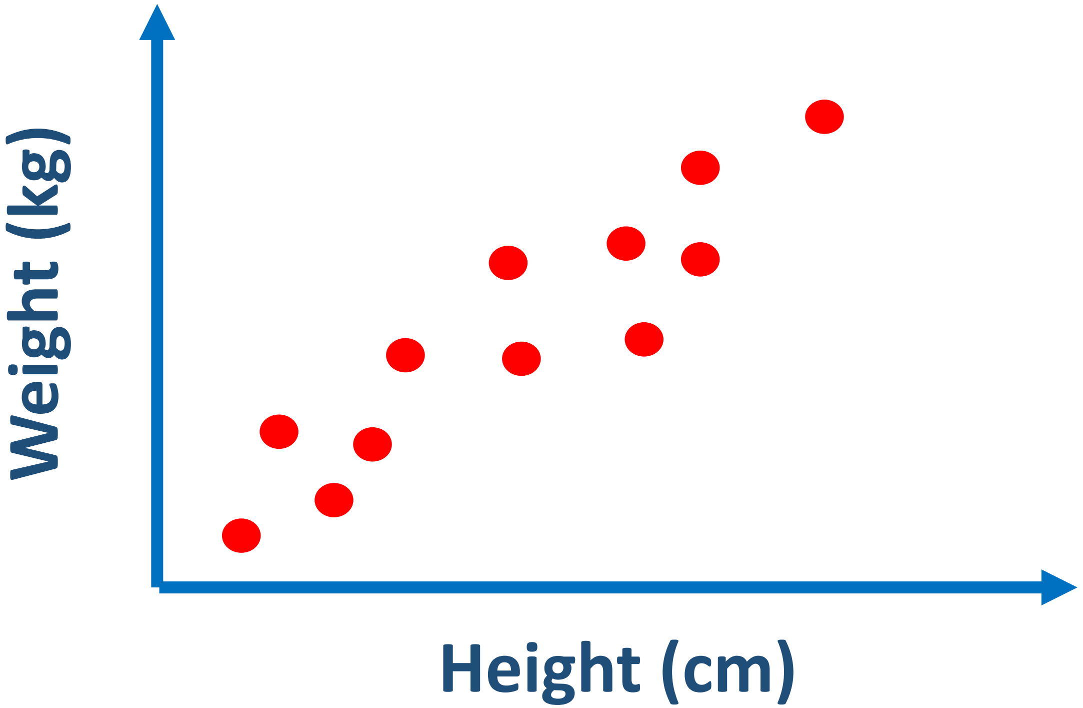
is a measure of **the strength** of the **linear association**, and it **indicates the direction of the linear association.**

✓**Strength**

correlation coefficient (r/ρ) from -1 to 1

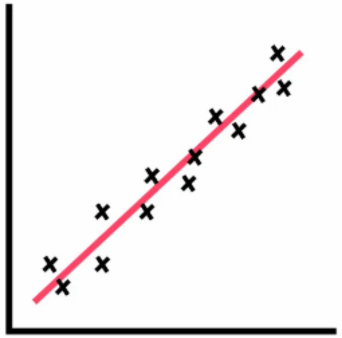
✓**Direction**

Positive or negative

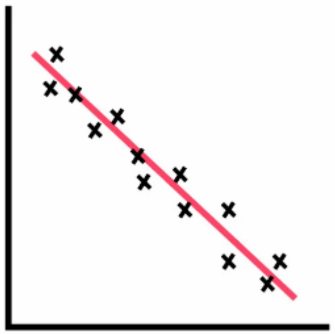


✓Direction

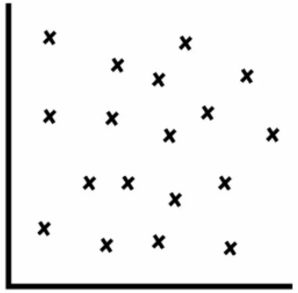
Positive or negative



Positive Correlation



Negative Correlation

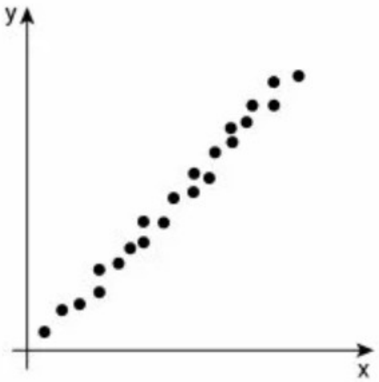


No Correlation

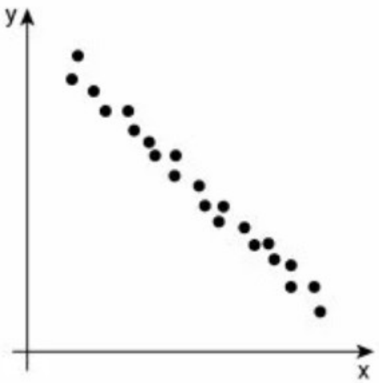
✓Strength

(r) from -1 to 1

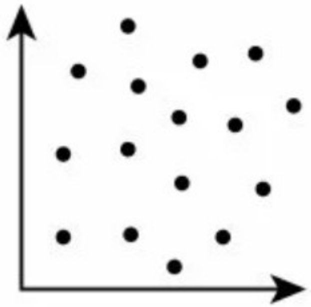
High Degree of Positive Correlation



High Degree of Negative Correlation



No Correlation



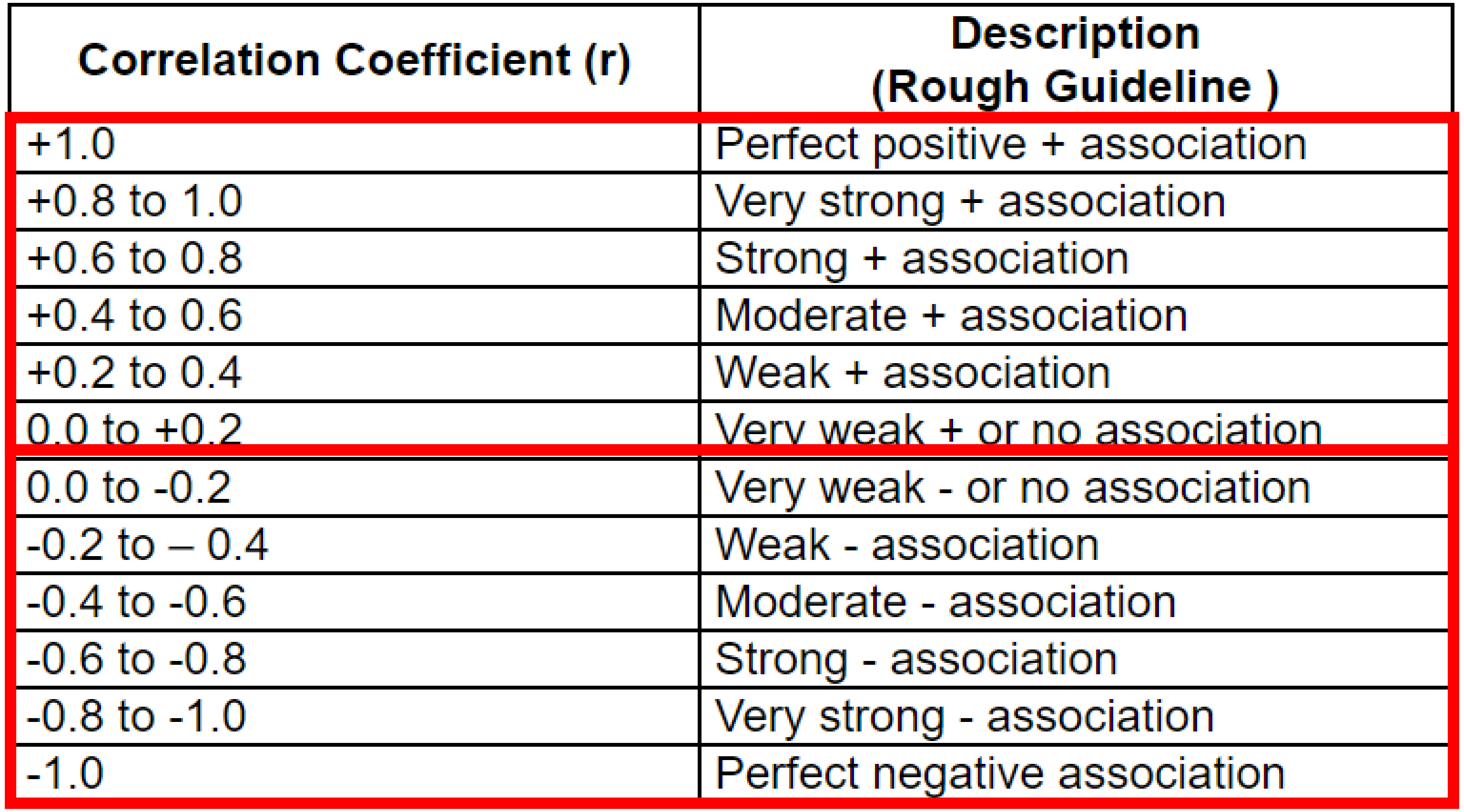
Correlation

✓**Direction**

Positive or negative

✓**Strength**

(r) from -1 to 1



Correlation Coefficient (r)	Description (Rough Guideline)
+1.0	Perfect positive + association
+0.8 to 1.0	Very strong + association
+0.6 to 0.8	Strong + association
+0.4 to 0.6	Moderate + association
+0.2 to 0.4	Weak + association
0.0 to +0.2	Very weak + or no association
0.0 to -0.2	Very weak - or no association
-0.2 to – 0.4	Weak - association
-0.4 to -0.6	Moderate - association
-0.6 to -0.8	Strong - association
-0.8 to -1.0	Very strong - association
-1.0	Perfect negative association

https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-

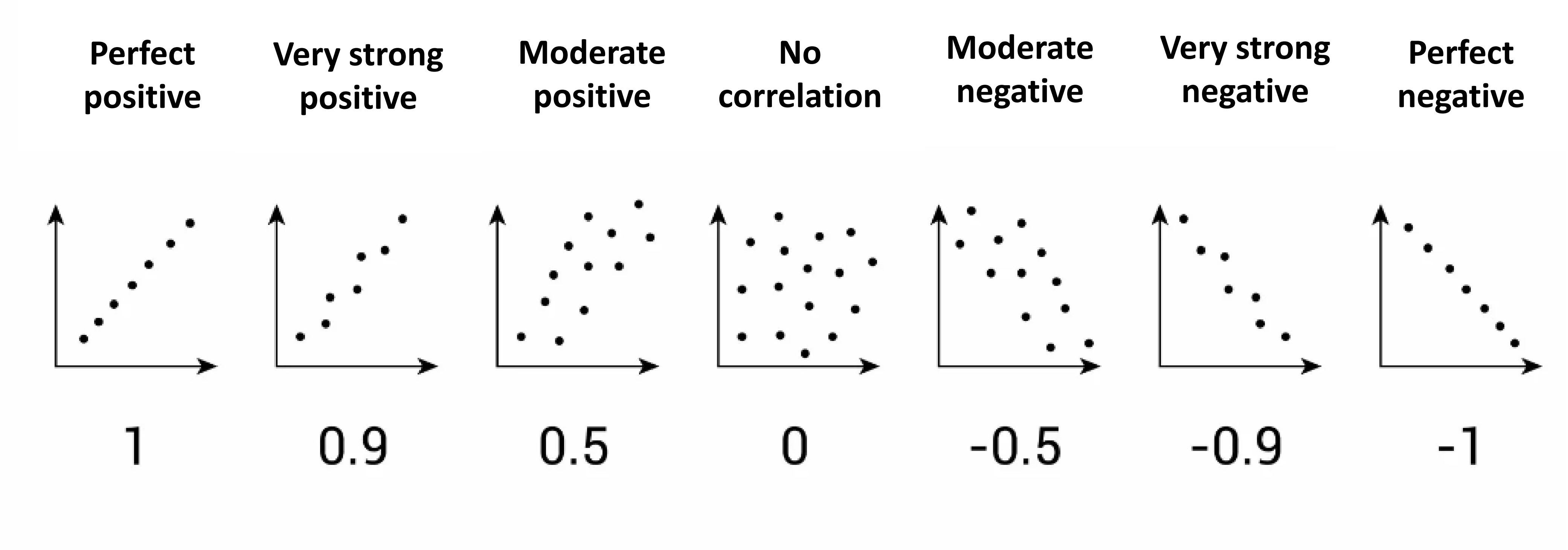
Correlation

✓Direction

Positive or negative

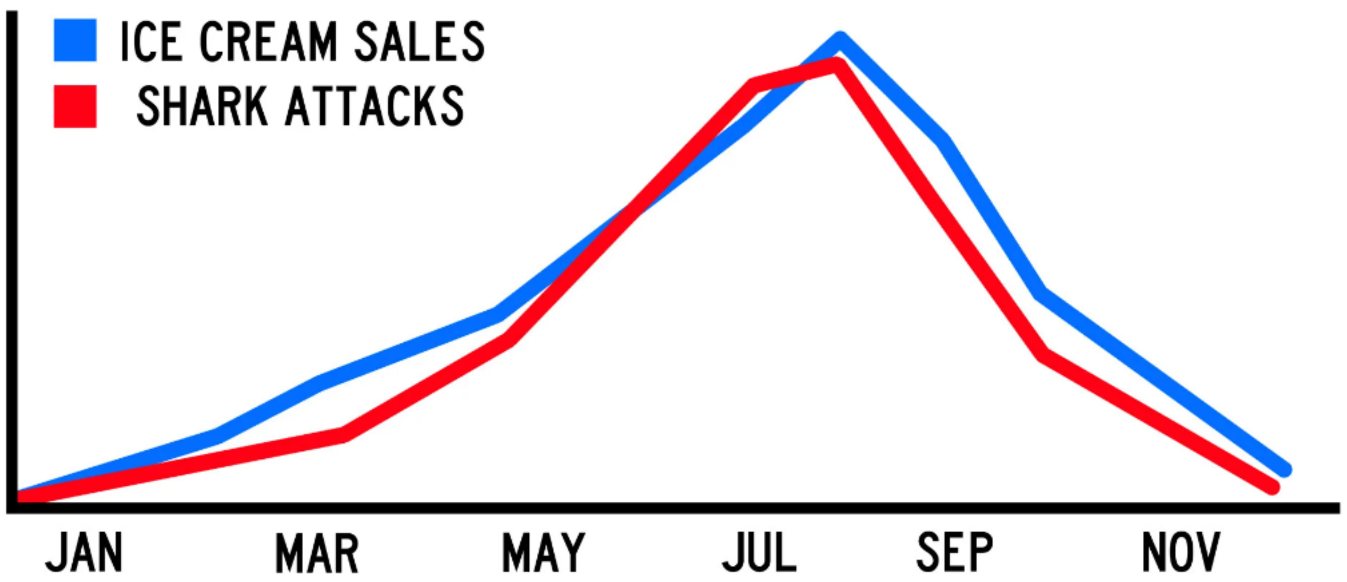
✓Strength

(r) from -1 to 1



Correlation

CORRELATION IS NOT CAUSATION!

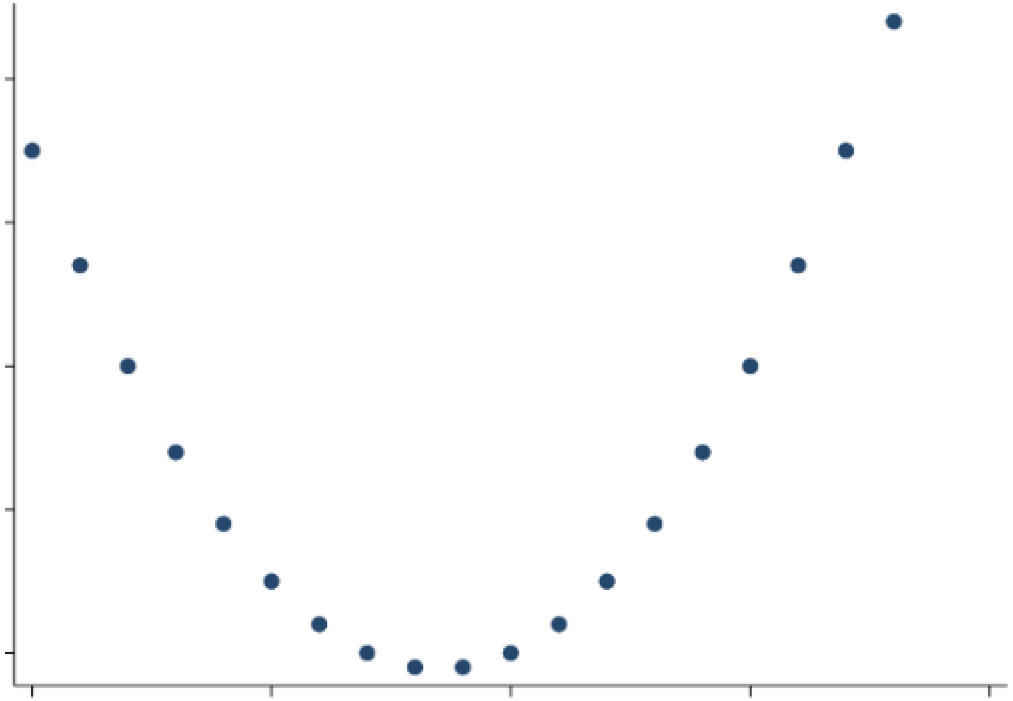


Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

https://www.simplypsychology.org/correlation.html

Correlation

is a measure of **the strength** of the **linear association**



https://www.coursera.org/learn/linear-regression-r-public-health

Parametric and non-parametric: association between two numerical variables

We have: two numerical continuous variables

Parametric

↓

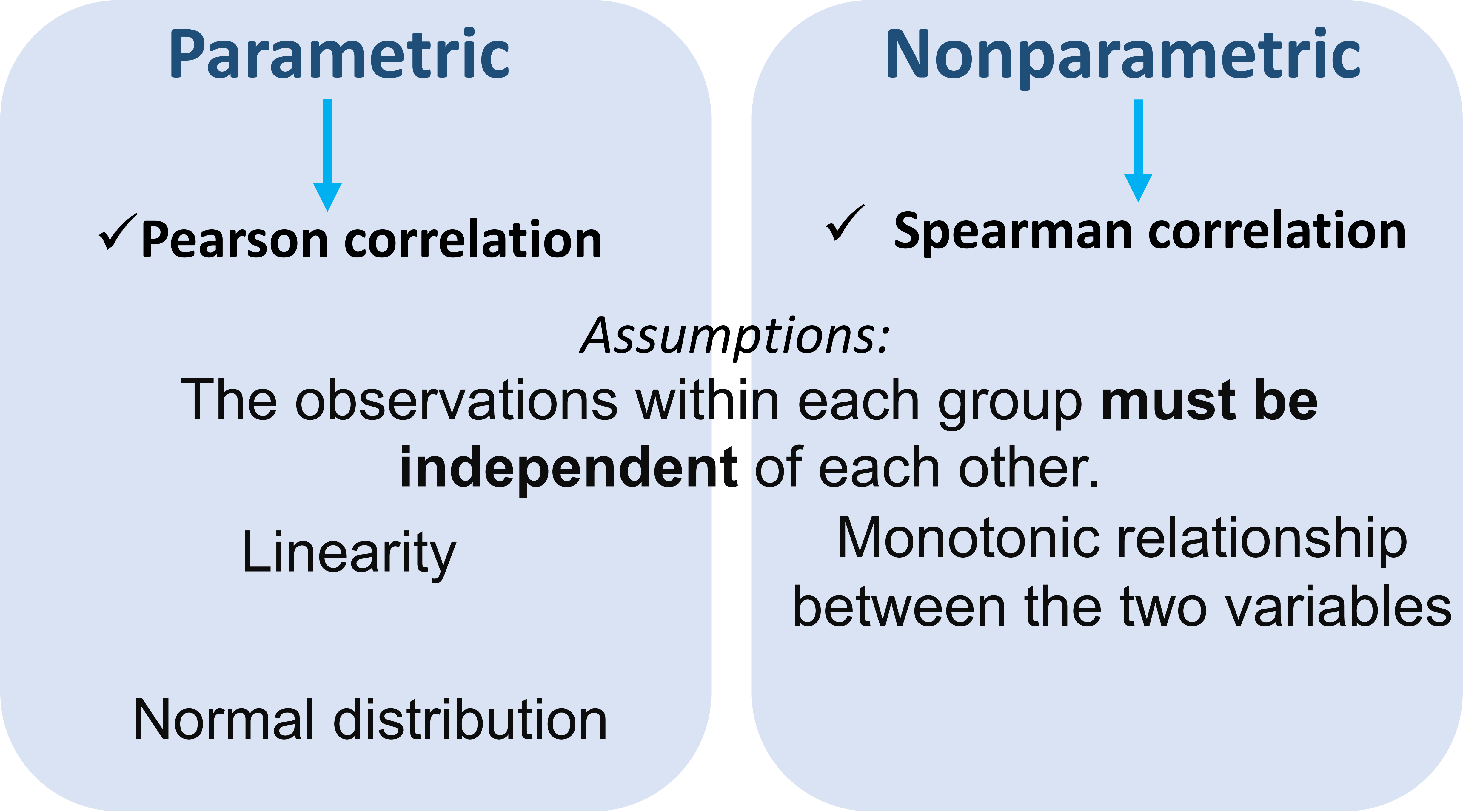
✓Pearson correlation

! extremely sensitive to sample size, approx.< 30

Nonparametric

↓

✓ Spearman correlation



Parametric

✓ **Pearson correlation**

Nonparametric

✓ **Spearman correlation**

Assumptions:

The observations within each group **must be independent** of each other.

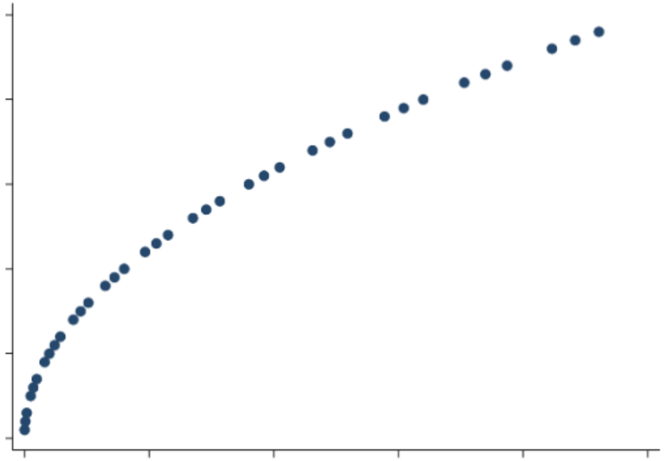
Linearity

Normal distribution

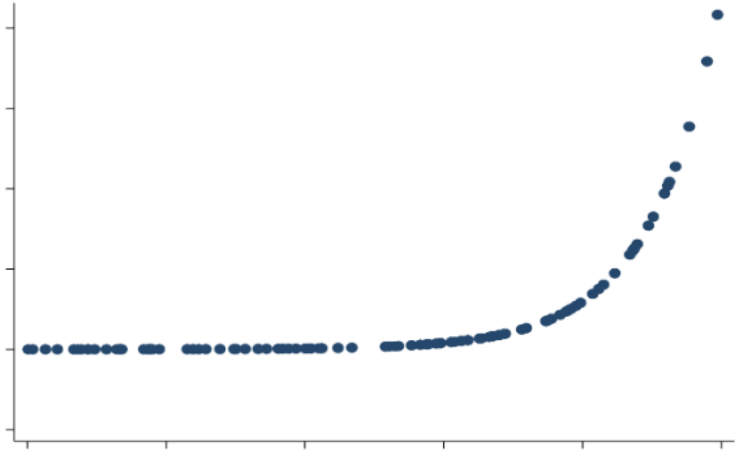
Monotonic relationship between the two variables

Monotonic relationship between the two variables

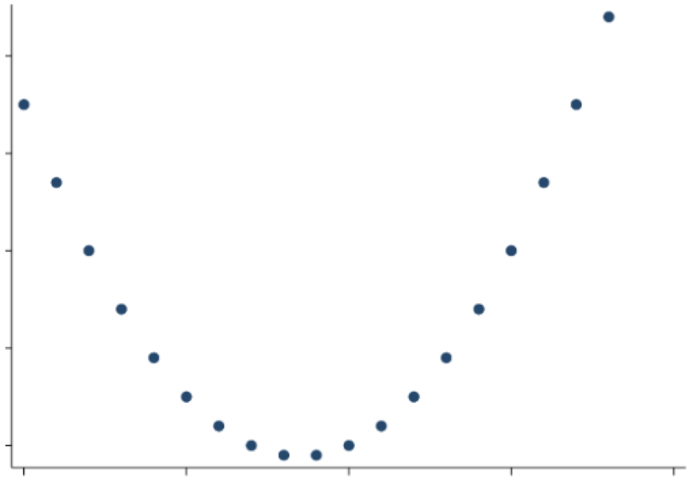
Monotonicity means that as the value of **one variable increases**, the value of the **other variable either consistently increases or consistently decreases** (but not necessarily at a constant rate). This assumption means that the relationship between the variables **doesn't have to be strictly linear.**



Monotonic relationship example1

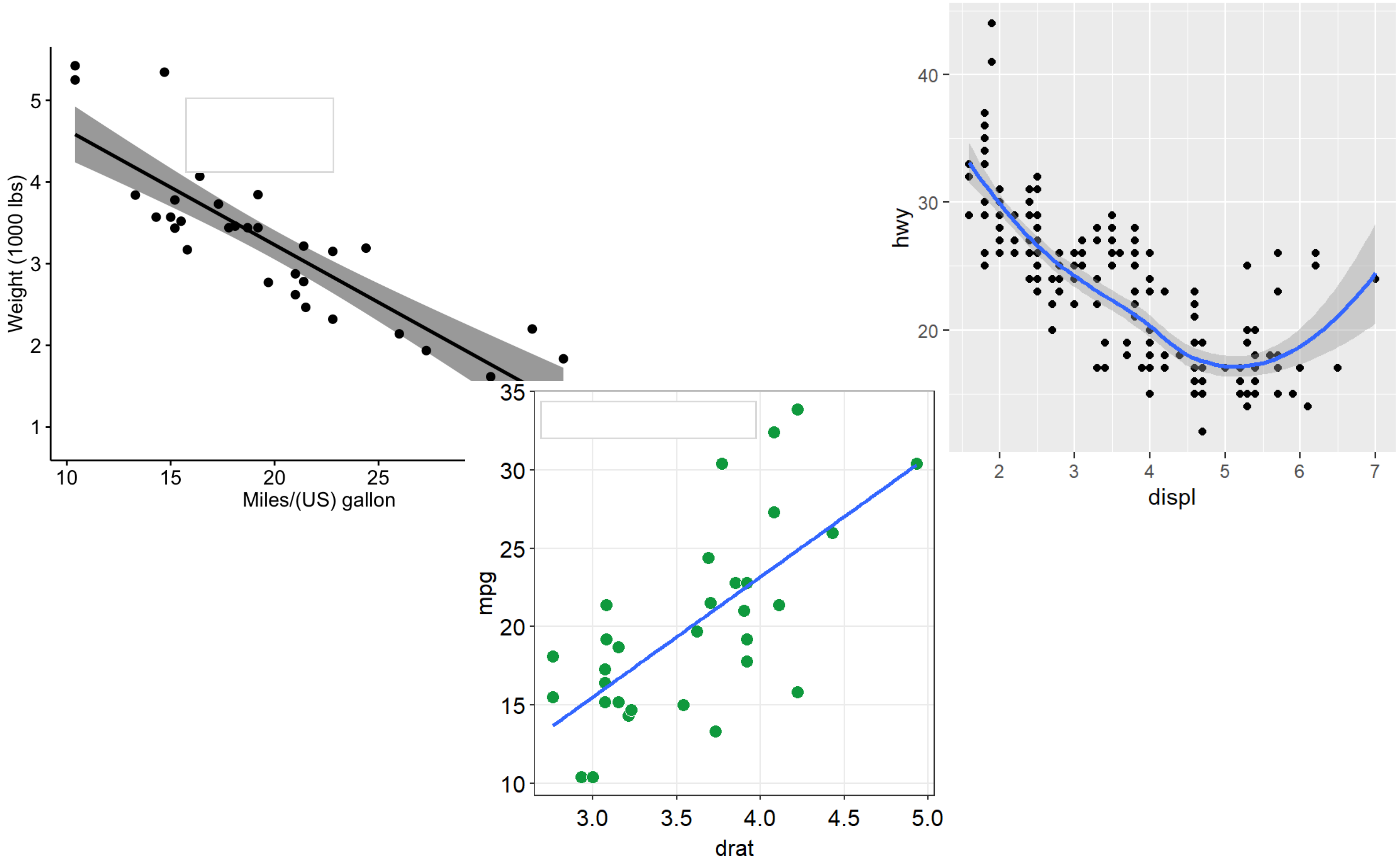


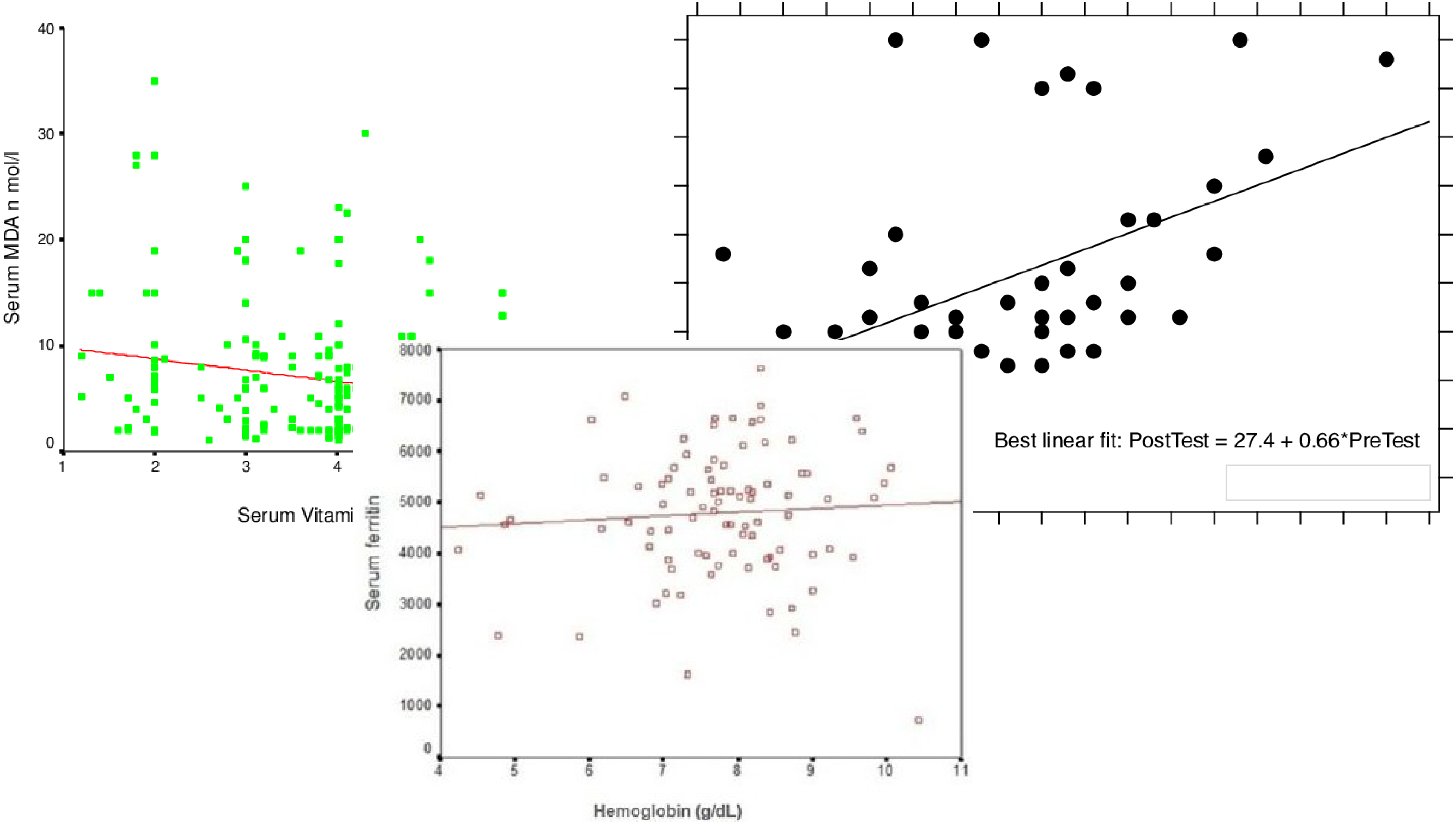
Monotonic relationship example2



https://www.coursera.org/learn/linear-regression-r-public-health

Task!


Weight (1000 lbs)
1
2
3
4
5
10
15
20
25
Miles/(US) gallon
mpg
10
15
20
25
30
35
3.0
3.5
4.0
4.5
5.0
drat
hwy
20
30
40
2
3
4
5
6
7
displ


Serum MDA n mol/l
40
30
20
10
0
1
2
3
4
Serum Vitami
Best linear fit: PostTest = 27.4 + 0.66*PreTest
Serum ferritin
8000
7000
6000
5000
4000
3000
2000
1000
0
4
5
6
7
8
9
10
11
Hemoglobin (g/dL)

Practical part