Linear regression

STATS – Modelação Estatística

PhD Programme in Health Data Science

Cristina Costa Santos & Andreia Teixeira

Statistical Modeling Curricular Unit

This unit aim to empower the students with skills to interpret **statistical modeling** results**.**

After this course unit the students should also develop the adequate skills in order to apply and correctly interpret the studied statistical methodologies using the appropriate software (preferably the R)**.**

Statistical Modeling Curricular Unit

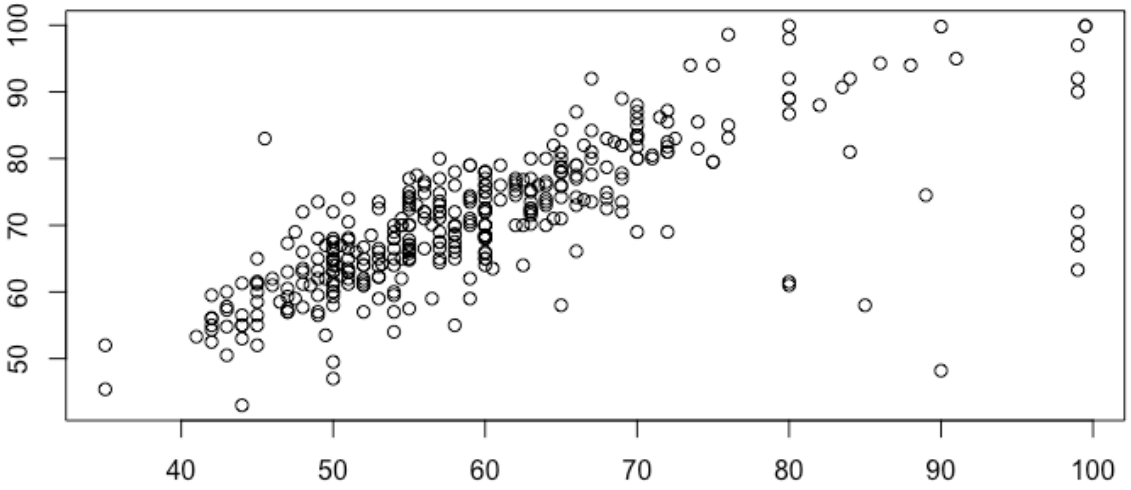
- Linear regression
- Logistic regression
- Log linear regression
- GLM
- Poisson regression
- Survival analysis (COX regression).
- Longitudinal and repeated data analysis

Statistical Modeling Curricular Unit

- **Assessment Components:**
  - **Exam** (50% of the final grade):1st of July
  - **Pratical Assignment** (50% of the final grade): goups of 3 –will be presented on the 3rd of June

Linear regression

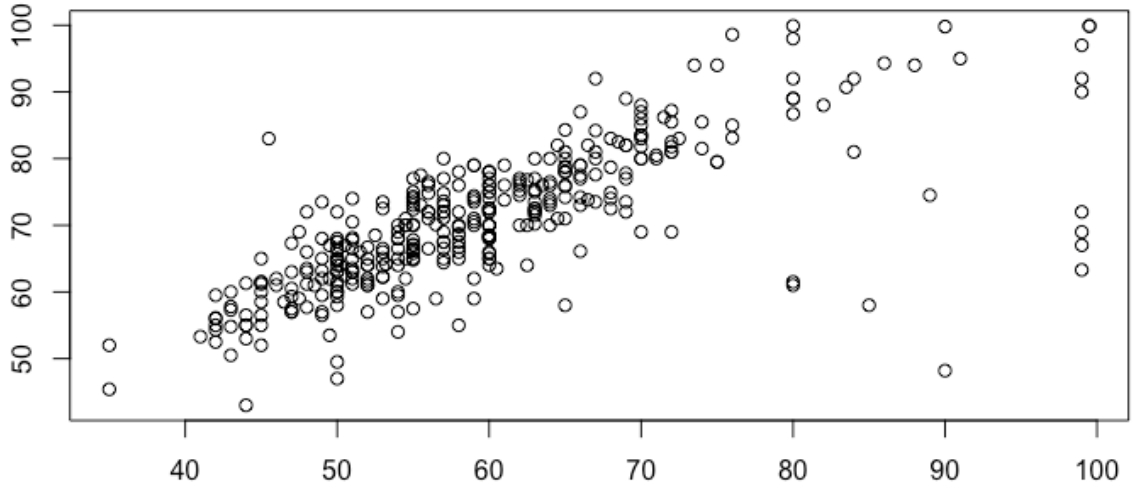
A scatterplot is a graphical representation of the association between two variables.



here each dot represents a woman and her weight at admission for labor (y-axis) and before pregnancy (x-axis).

Linear regression

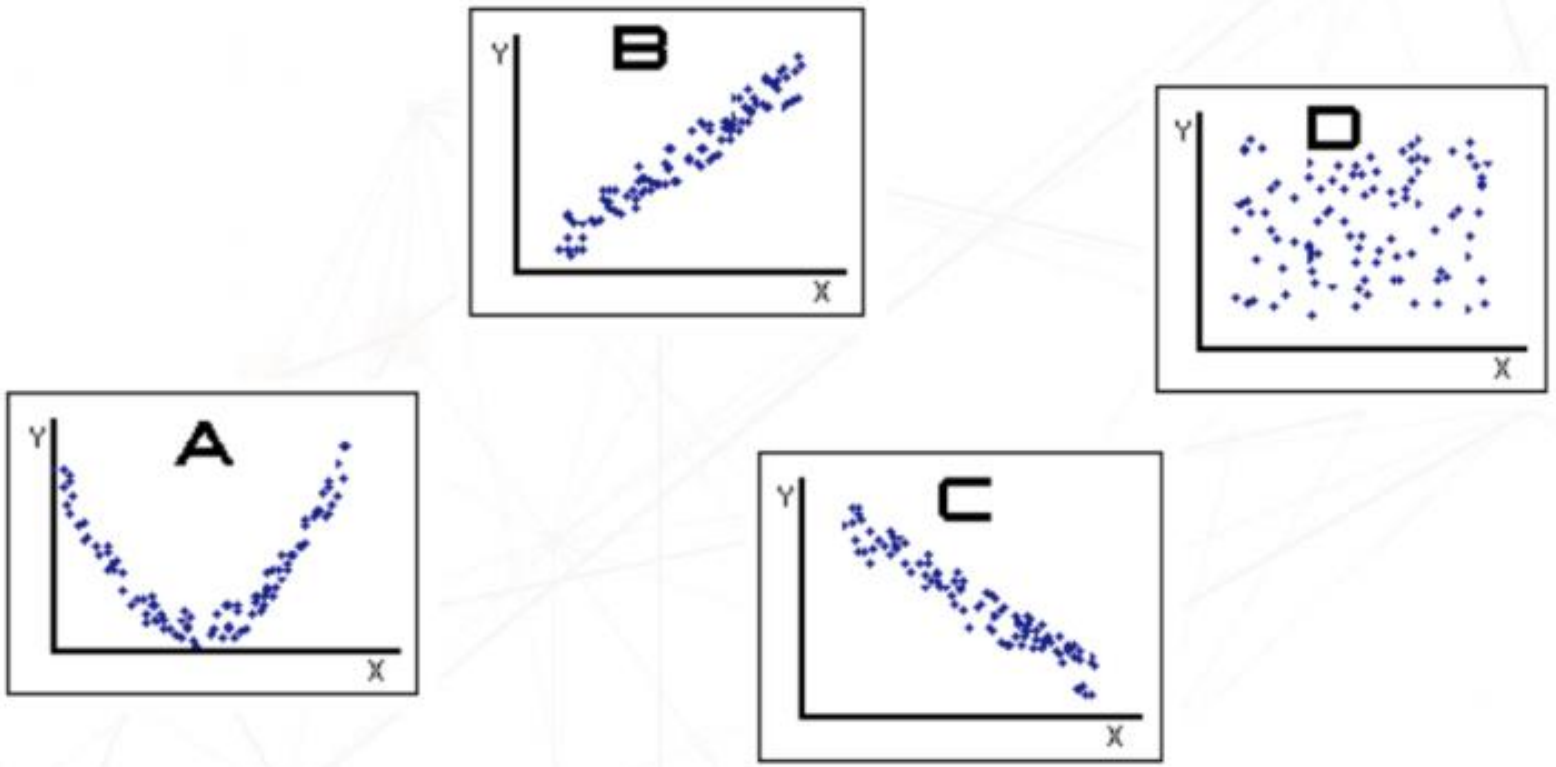
A scatterplot is a graphical representation of the association between two variables.



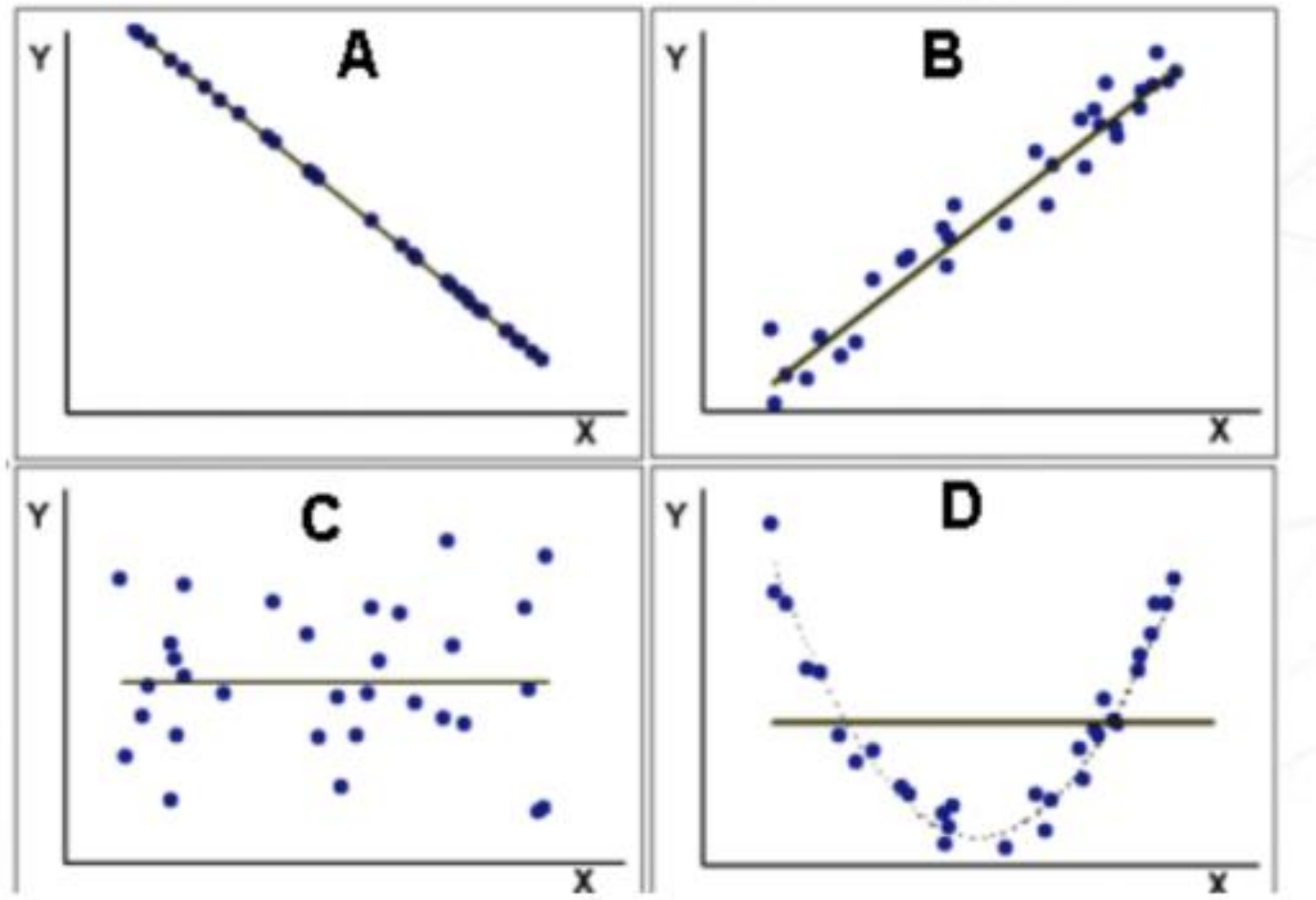
here each dot represents a woman and her weight at admission for labor (y-axis) and before pregnancy (x-axis).

The diagram suggests the phenomenon we expected:
women with higher pre-pregnancy weight tend to be heavier on admission for labor
The association between weights seems to be linear.

Is there any linear relationship?

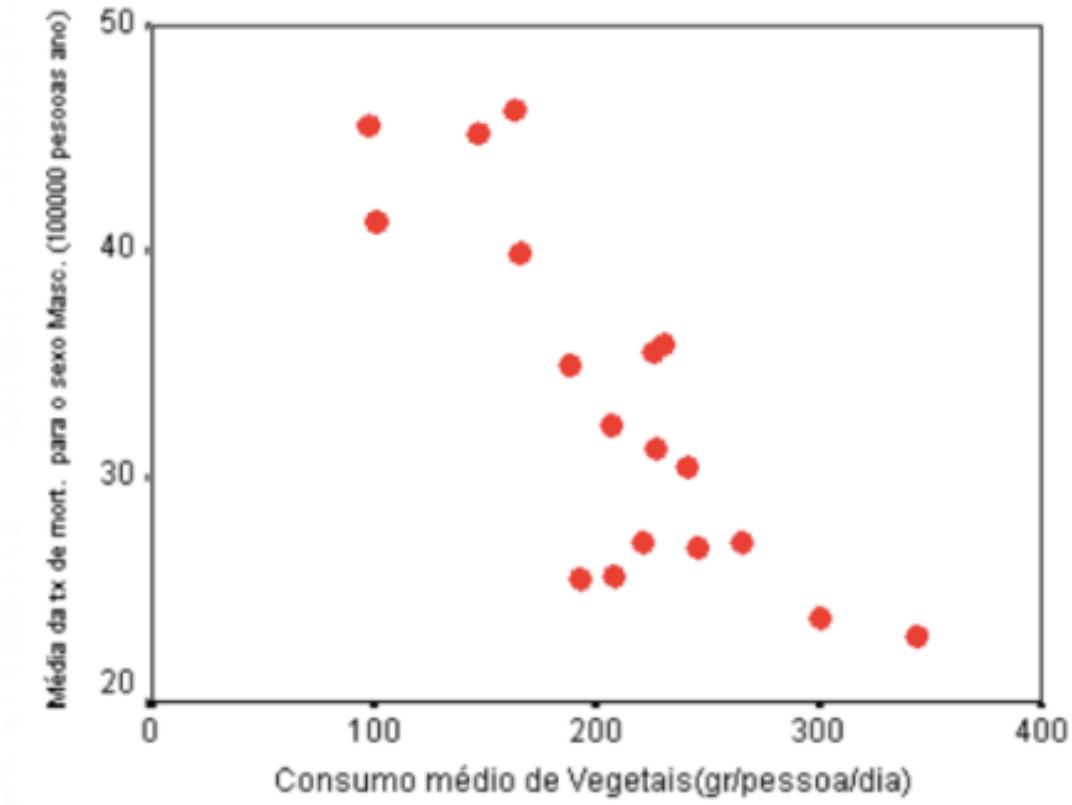


Is there any linear relationship?



Linear relationship?

In a study carried out in the 18 districts of Portugal:



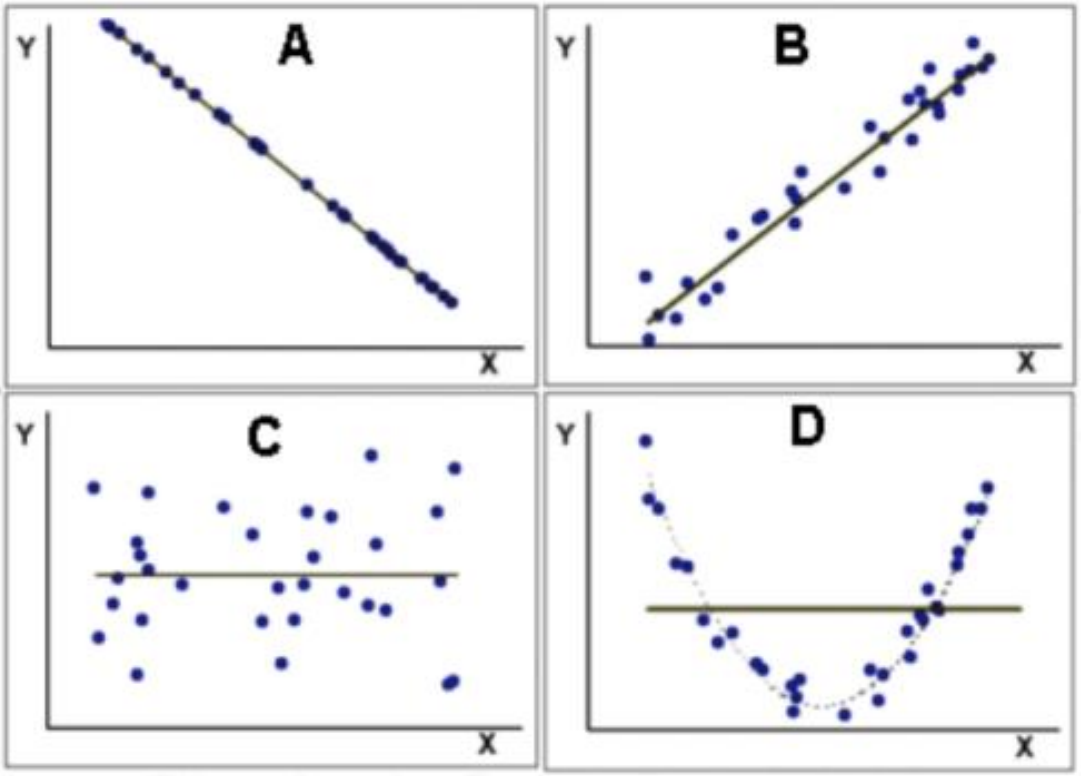
Is there any association between the average consumption of vegetables and the average stomach cancer mortality rate for males?

Pearson correlation coefficient

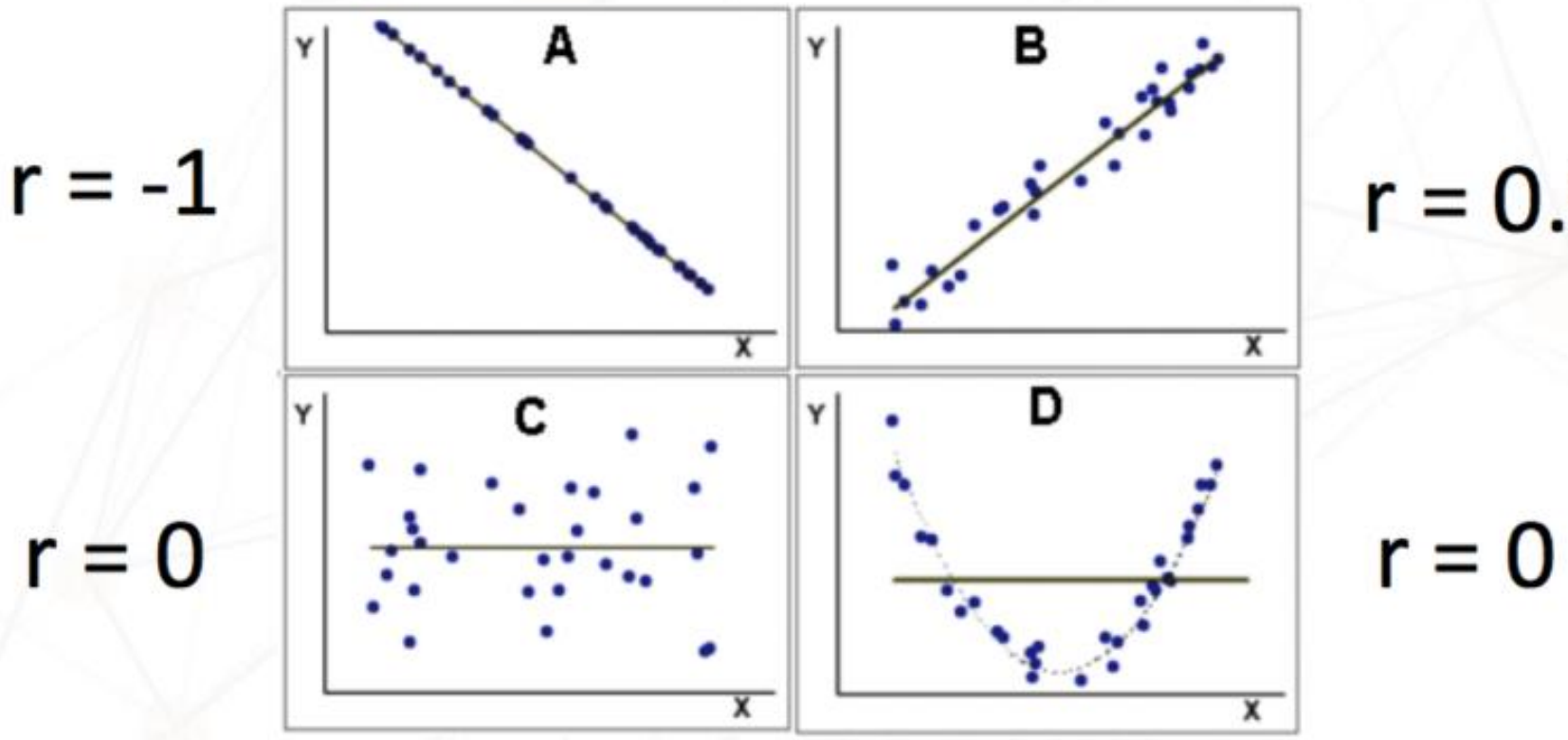
Pearson's correlation coefficient is a measure of the 'quality' of the approximation of the relationship between two variables by a straight line, that is, it measures the linear association between two variables.

$$r = \frac{cov(x,y)}{\sqrt{var(x)v\ ar(y)}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Pearson correlation coefficient



Pearson correlation coefficient



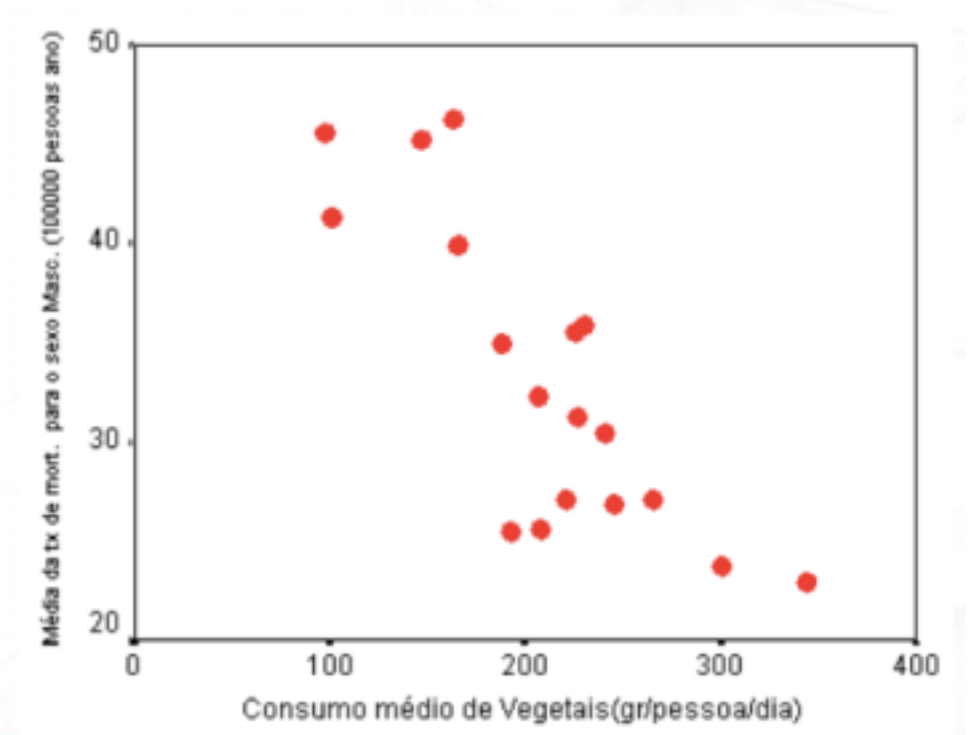
Pearson correlation coefficient

```
> cor.test(alcohol$mwt0, alcohol$mwtadm)

        Pearson's product-moment correlation

data:  alcohol$mwt0 and alcohol$mwtadm
t = 20.135, df = 363, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6739595 0.7714865
sample estimates:
      cor
0.7263589
```

The hypothesis test associated with the correlation is not very informative

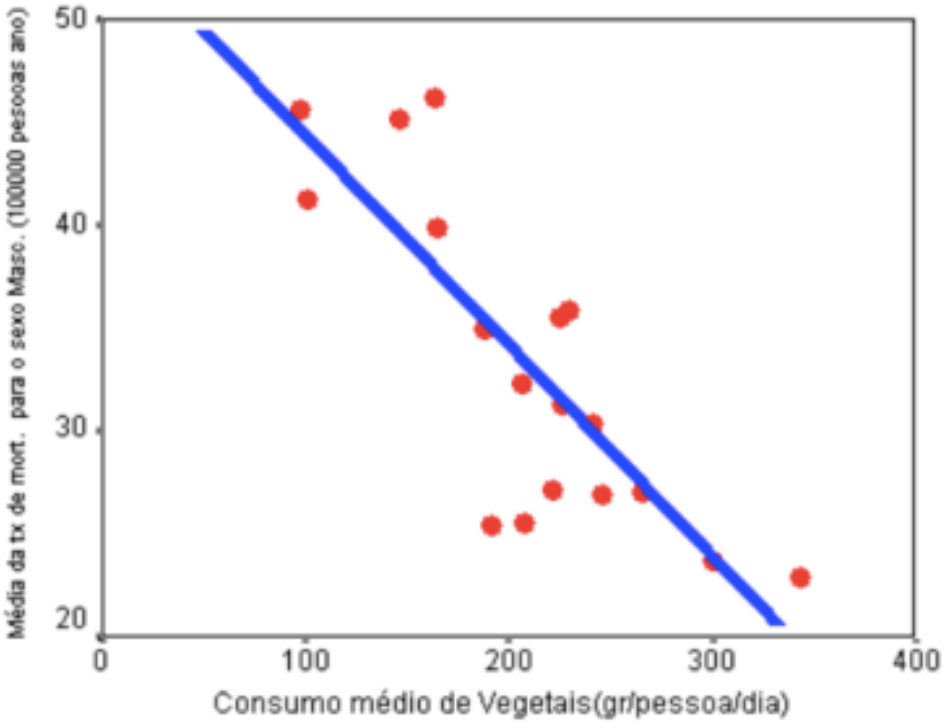
Normally we are more interested in the strength of the linear association and not in whether it is significantly different from zero

Pearson correlation coefficient



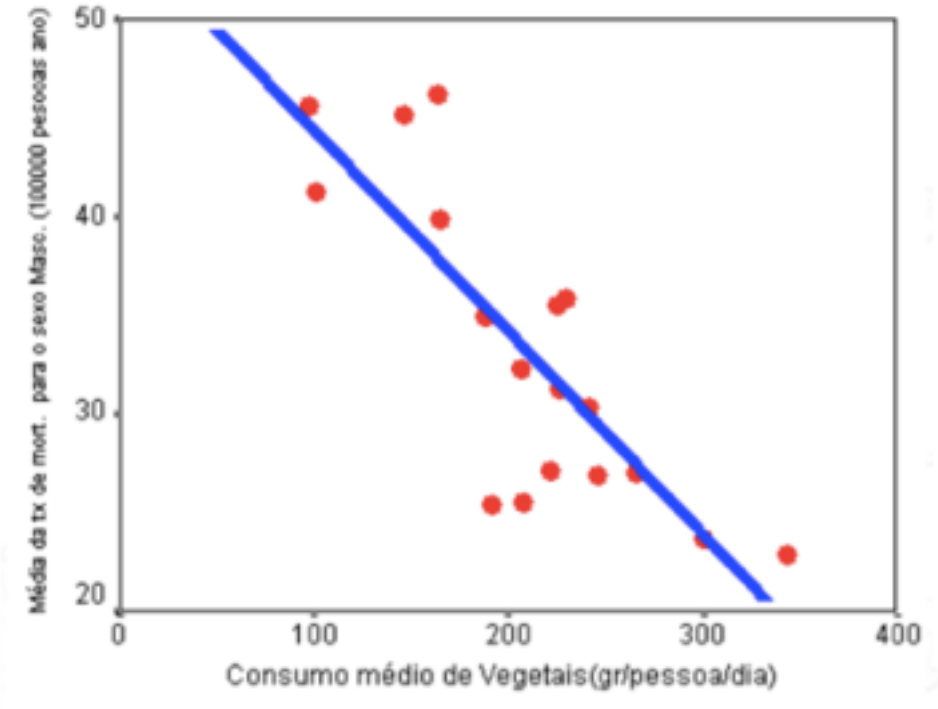
What will be the correlation coefficient between vegetable consumption and stomach cancer mortality?

Pearson correlation coefficient



In the study about the relationship between the consumption of vegetables and the mortality rate from stomach cancer in men, a correlation of -0.814 was obtained

Determination coefficient



r^2 is interpreted as the amount of variation of one variable that can be explained by the other

r^2=(-0.814)x(-0.814)=0.66

Vegetable consumption explains 66% of stomach cancer mortality in men

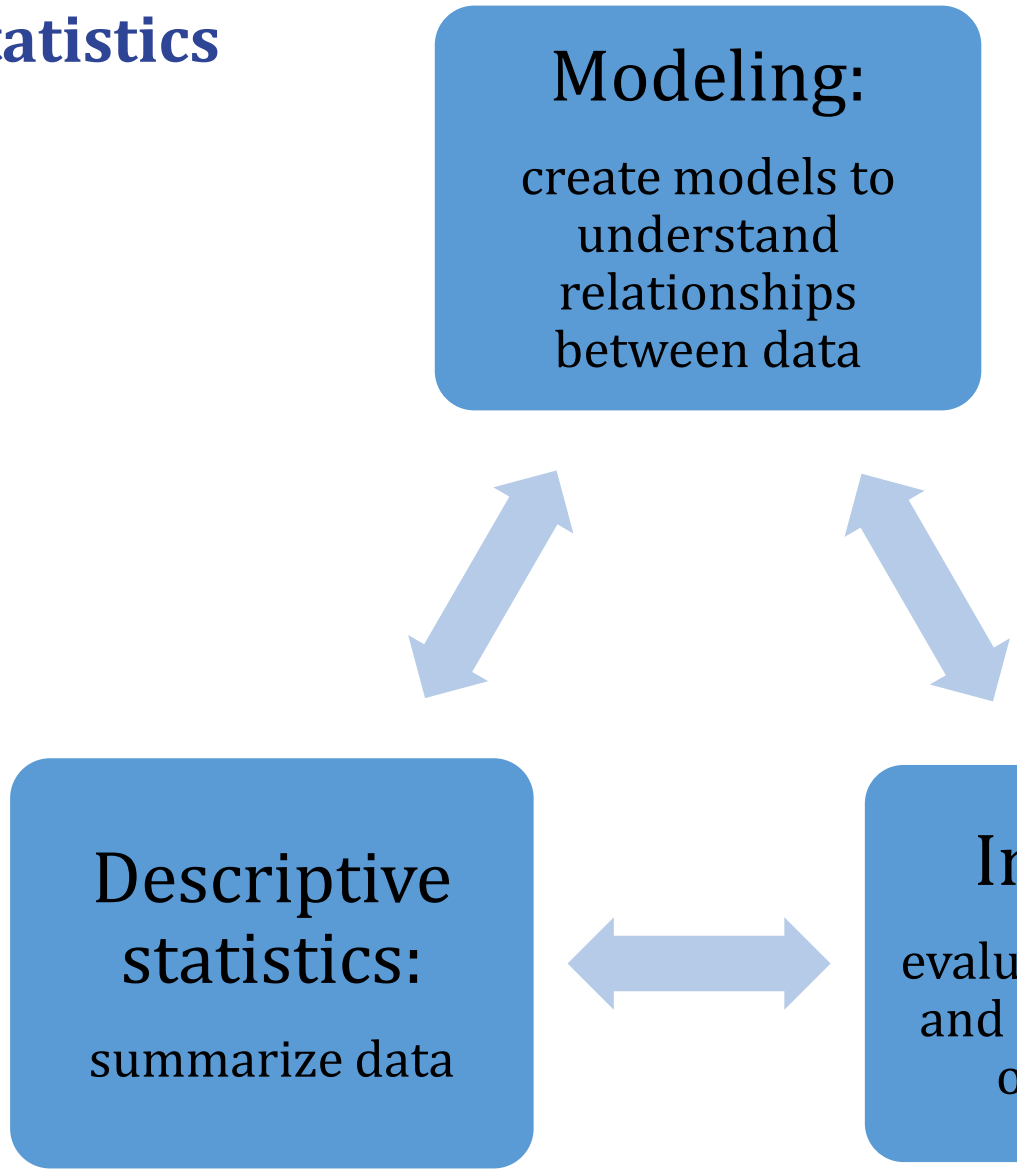
Spearman's correlation coefficient

- The Spearman correlation coefficient is the Pearson correlation coefficient applied to orderings of values instead of absolute values

- It is an alternative to Pearson's correlation coefficient when one of the variables has a skewed distribution or has outliers


r negativo

Statistics


Modeling:
create models to understand relationships between data
Descriptive statistics:
summarize data
Inferência:
evaluate the accuracy and generalizability of the results

Linear regression

- Linear regression is a mathematical model used to study the relationship between two variables.
- The model tries to predict the values of one of the variables as a function of the other.

- In the study on vegetable consumption and stomach cancer mortality rate, the most natural situation is to try to predict the mortality rate (dependent variable) for a given vegetable consumption (independent variable) and not the other way around.

Modeling

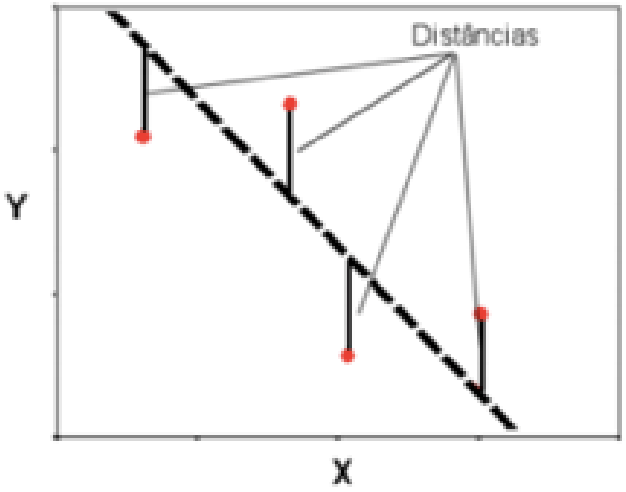
- We can use **modeling** to:
  - describe the strength of the association between outcome variables and factors of interest
  - Adjust for confounding variables
  - Identify the risk factors that affect the outcome variable
  - Prevision

Linear regression

Stomach Cancer Mortality = $b_0 + b_1$ Vegetables

The coefficients are calculated in such a way that the sum of the distances to the straight line is minimized

Least squares method



Linear regression

Stomach Cancer Mortality = 54.503 − 0.102 Vegetables

b_0 = 54.503

Predicted mortality rate with zero vegetable consumption

b_1 = -0.102

Predicted decrease (because the coefficient is negative) in the mortality rate for a one-unit increase in vegetable consumption

Linear regression

ANOVA table: variation explained by the model

In the case of the male mortality rate, the total variation is 1036.118 (685.986+350.123)

Considering the consumption of vegetables, the variation in explained mortality is 685.986

The residual (350.123) is simply the variation that remains unexplained, that is, the difference between the total variation and the explained variation.

	Df	Sum Sq	Mean Sq	F value
vegetables	1	685.986	685.986	31.348
Residuals	16	350.132	21.883	

Linear regression

The quotient of explained variation to total variation is the percentage of explained variation

$$\frac{685.986}{1036.118} = 0.66\ (66\%)$$

As expected, this value is equal to the square of the correlation coefficient (percentage of explained variation)

$r^2 = 0.8142^{\,2} = 0.66\ (66\%)$

	Df	Sum Sq	Mean Sq	F value
vegetables	1	685.986	685.986	31.348
Residuals	16	350.132	21.883	

Linear regression

Assumptions:

- linear relationship between the independent variable, x, and the dependent variable, y.
- the observations are independente
- for each fixed value of the independent variable, dependent variable follows a normal distribution, and all these normal distributions have an equal standard deviation

How to check

Residuals

Residual analysis

- The residuals are normally distributed with the mean at zero.

- The residuals and predicted values of the dependent variable calculated by the model are not related