Multivariate Descriptive Statistics: Multiway tables and dependence structures

Rosario Barone

Tor Vergata University of Rome

Statistical tools for decision making

Undergraduate Degree in Global Governance

A.Y. 2023/2024

Multidimensionality

- So far we have only considered univariate tools, which looks at just one variable.
- **Multivariate Statistics**: simultaneous observation and analysis of more than one variable.
- Multivariate tools helps in evaluating the relationship between two or more variables.

Summary

We presents methos basing on the nature of the variables:

- Categorical variables: Contigency (Multiway) tables and measures of association:
  - Risk Ratio;
  - Odds Ratio
  - Relative Risk.
- Quantitative variables: Multivariate distributions and measures of association:
  - Covariance;
  - Correlation
  - linear regression.

Contingency tables

The joint distribution between two categorical variables determines their relationship. This distribution also determines the marginal and conditional distributions.

Let X and Y denote two categorical response variables, X with I categories and Y with J categories. Classifications of subjects on both variables have IJ possible combinations. The responses (X, Y) of a subject chosen randomly from some population have a probability distribution.

We define as *contingency table* a rectangular tables with I rows and J columns, containing the frequencies of the outcome for each of the variables.

Notation tables

- **Joint Absolute Frequency**: how many times a combination of two conditions happens together.

	1	**2**	**I**
1	f_{11}	f_{12}	f_{1+}
2	f_{21}	f_{22}	f_{2+}
J	f_{+1}	f_{+2}	N

- **Relative Frequencies**...

	1	**2**	**I**
1	n_{11} $(n_{1\|1})$	n_{12} $(n_{2\|1})$	n_{1+} (1)
2	n_{21} $(n_{1\|2})$	n_{22} $(n_{2\|2})$	n_{2+} (1)
J	n_{+1}	n_{+2}	1

Contingency tables

Table cells at the intersections of rows and columns indicate frequencies of both events coinciding.

For example, the table below shows the preferred financial asset by a group of 223 investors .

	Bonds	**Equity**	**Row Tot**
Male	66	40	106
Female	30	87	117
Col Tot	96	127	223

Contingency tables

Contigency tables helps in calculating probabilities:

	Bonds	**Equity**	**Row Tot**
Male	66	40	106
Female	30	87	117
Col Tot	96	127	223

1. Joint relative Frequency: $n_{ij} = \frac{f_{ij}}{\sum_i \sum_j f_{ij}}$:

$$(\text{Female and Bond}) \equiv (\text{Female} \cap \text{Bond}) = 30/223 = 0.135$$

2. Marginal relative Frequency: $n_i = \frac{\sum_j f_{ij}}{\sum_i \sum_j f_{ij}}$

$$n_{Bond} = (\text{Bond}) = 96/223 = 0.431$$

Contingency tables

- Conditional relative frequency: $n_{i|j} = n_{ij}/n_j$, where n_{ij} is calculated as in (1) and n_j calculated as in (2).

$$n_{Female|Bond} = (\text{Female}|\text{Bond}) = \frac{(\text{Female} \cap \text{Bond})}{(\text{Bond})} = \frac{0.135}{0.431} = 0.313$$

However, via contingency tables:

	Bonds	**Equity**	**Row Tot**
Male	66	40	106
Female	**30**	87	117
Col Tot	**96**	127	223

$$n_{Female|Bond} = (\text{Female}|\text{Bond}) = 30/96 = 0.313$$

Comparing two proportions

- Relative risk: simply the ratio of proportions n_1/n_2.
- Odds-ratio: the ratio of odds $\left(\frac{n_1/(1-n_1)}{n_2/(1-n_2)}\right)$.
- Risk difference: difference of proportions.

Let suppose to observe two groups of individuals, respectively G_T (Treated group) and G_C (Control Group).

	Treatment (T)	**Control (C)**
Event (E)	TE	CE
Non-event (N)	TN	CN

Relative Risk

Relative risk:

$$RR = (TE/(TE + TN))/(CE/(CE + CN))$$

$$RR = n_T/n_C$$

- $RR = 1$ the treatment is not associated with the outcome;
- $RR < 1$ the risk of the outcome might be decreased by the treatment, which is a "protective factor";
- $RR > 1$ the risk of the outcome might be increased by the treatment, which is a "risk factor".

Odds Ratio

Odds Ratio:

$$OR = (n_T/(1 - n_T))/(n_C/(1 - n_C))$$

- $OR = 1$ the treatment is not associated with the outcome;
- $OR < 1$ the treatment might be a "protective factor" against the outcome;
- $OR > 1$ the treatment might be a "risk factor" for the outcome.

Risk difference

Relative risk:

$$RD = (TE/(TE + TN)) - (CE/(CE + CN))$$

$$RD = n_T - n_C$$

- $RD = 0$ the treatment is not associated with the outcome;
- $RD < 0$ the risk of the outcome might be decreased by the treatment, which is a "protective factor";
- $RD > 0$ the risk of the outcome might be increased by the treatment, which is a "risk factor".

RR and OR

- RR more interpretable
- OR can always be computed while RR and RD only when outcomes are not fixed ("prospective" studies)
- The OR asymptotically approaches the RR for small frequencies of outcomes.
- Precisely,

$$OR = RR\frac{1 - n_C}{1 - n_T}$$

Example: Policies in the management of prisoners

We analyze the data deriving from 3 Italian prisons:

- Prison A is particularly active in the work of social reintegration of prisoners.
- Prison B does not implement any particular type of policy.

Prison B will be our Control Group.

Example: Policies in the management of prisoners

	Prison A	**Prison B**
Recidivism	3	26
Non-Recidivism	25	10

$$RR = n_T/n_C = \frac{\frac{3}{3+25}}{\frac{26}{26+10}} = \frac{0.11}{0.72} = 0.15$$

$$OR = (\frac{n_T}{1-n_T})/\frac{n_C}{1-n_C} = \frac{\frac{0.11}{1-0.11}}{\frac{0.72}{1-0.72}} = 0.05$$

$$RD = n_T - n_C = 0.11 - 0.72 = -0.61$$

Example: Prohibition policy

We analyze the data deriving from 2 Countries:

- Country A has introduced strong laws against the consuption of the alcohol.
- Prison B does not implement any particular type of policy.

Prison B will be our Control Group.

Example: Does the Prohibition works?

	Country A	**Country B**
Addicted to alcohol	35687	65592
Non-Addicted to alcohol	2057659	55346543

$$RR = n_T/n_C = \frac{\frac{35687}{35687+2057659}}{\frac{65592}{65592+55346543}} = \frac{0.0018}{0.0012} = 1.5$$

$$OR = (\frac{n_T}{1-n_T})/\frac{n_C}{1-n_C} = \frac{\frac{0.0018}{1-0.0018}}{\frac{0.0012}{1-0.0012}} = 1.5$$

$$RD = n_T - n_C = 0.0018 - 0.0012 = 0.0006$$

Note: RR and OR have the same value. Why?

Multivariate Distributions

Suppose we have p variables $X_1, ..., X_j, ..., X_p$ observed on the same n units. Let's consider the vectors $(n \times 1)$ associated with them and denote them by $x(1), ..., x(j), ..., x(p)$. The generic variable vector is

$$x(j) = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

With these vectors it is possible to construct a matrix X with p columns (the vectors of the variables) and n rows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Covariance

Covariance

The covariance between two distributions, denoted as $\text{cov}(x(j), x(h))$ or $\sigma_{x(j)x(h)}$, measures the linear dependence of two distributions and it is defined as the mean value of the product of the deviations of two disrtibutions from their respective means

$$\text{cov}(x(j), x(h)) = \frac{1}{n}\sum_{i=1}^{n}(x(j)_i - M_{x(j)})(x(h)_i - M_{x(h)}).$$

where:

- $x(j)_i$ and $x(h)_i$ are individual data points of variables $x(j)$ and $x(h)$,
- $M_{x(j)}$ and $M_{x(h)}$ are the means (averages) of the distributions $x(j)$ and $x(h)$, respectively, and n is the number of data points.

Covariance

- The covariance between two variables measures the degree to which they vary together.
- A positive covariance indicates that two phenomena tend to increase or decrease together.
- A negative covariance suggests an inverse relationship.

Covariance Matrix

- A covariance matrix is a symmetric matrix that provides information about the covariance between multiple variables.

- The covariance matrix allows us to examine the linear relationships and dependencies between all pairs of variables in a dataset simultaneously. It is an essential tool for understanding the multivariate distribution of data.

Covariance Matrix

1. Calculate the mean (average) of each variable $M(x(j))$ for $j = 1, \ldots, p$.
2. For each pair of variables $(x(j), x(h))$, calculate the covariance using the following formula:

$$\text{cov}(x(j), x(h)) = \frac{1}{n} \sum_{i=1}^{n} (x(j)_i - M_{x(j)})(x(h)_i - M_{x(h)})$$

   where $\text{cov}(x(j), x(h)$ is the covariance between $x(j)$ and $x(h)$ and n is the number of observations.
3. Assemble the computed covariances into a matrix. The diagonal elements of the matrix represent the **variances** of individual variables, while the off-diagonal elements represent the **covariances** between pairs of variables.

Covariance Matrix

$$\mathbf{X} = \begin{bmatrix} \sigma^2_{x(1)} & cov(x(1), x(2)) & \cdots & cov(x(1), x(p)) \\ cov(x(2), x(1)) & \sigma^2_{x(2)} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ cov(x(p), x(1)) & cov(x(p), x(2)) & \cdots & \sigma^2_{x(p)} \end{bmatrix}$$

Note that: $cov(x(2), x(1)) = cov(x(1), x(2))$, $cov(x(p), x(1)) = cov(x(1), x(p))$ and more generally $cov(x(h), x(j)) = cov(x(j), x(h))$, by implying the **simmetry** of the matrix!

Example

Let's consider two distributions, $x(j)$ and $x(h)$, with the following data points: $x(j) : 2, 3, 5, 7, 8$ and $x(h) : 1, 2, 4, 6, 6$. Calculate the covariance $\text{cov}(x(j), x(h))$:

- we first need to find the means:

$$M_{(x(j))} = 5 \text{ and } M_{(x(h))} = 3.8$$

- Calculate the covariance:

$$\text{cov}(x(j), x(h)) = 4.48$$

- Is this result informative on the relationship between the considered distributions?

Example

Let's consider two distributions, $x(j)$ and $x(h)$, with the following data points: $x(j) : 2, 3, 5, 7, 8$ and $x(h) : 1, 2, 4, 6, 6$. Calculate the covariance $\text{cov}(x(j), x(h))$:

- we first need to find the means:

$$M_{(x(j))} = 5 \text{ and } M_{(x(h))} = 3.8$$

- Calculate the covariance:

$$\text{cov}(x(j), x(h)) = 4.48$$

- Is this result informative on the relationship between the considered distributions?
  Covariance values are not bounded: therefore it may be wrong to conclude that there might be a high relationship between variables when the covariance is high.

Correlation

Correlation

The correlation between two random variables, denoted as $\rho(x(j), x(h))$ or $r_{x(j),x(h)}$, measures the strength and direction of their linear relationship. It quantifies how well the values of one variable can be predicted from the values of another. The correlation coefficient is defined as:

$$\rho(x(j), x(h)) = \frac{\text{cov}(x(j), x(h))}{\sigma_{x(j)}\sigma_{x(h)}}$$

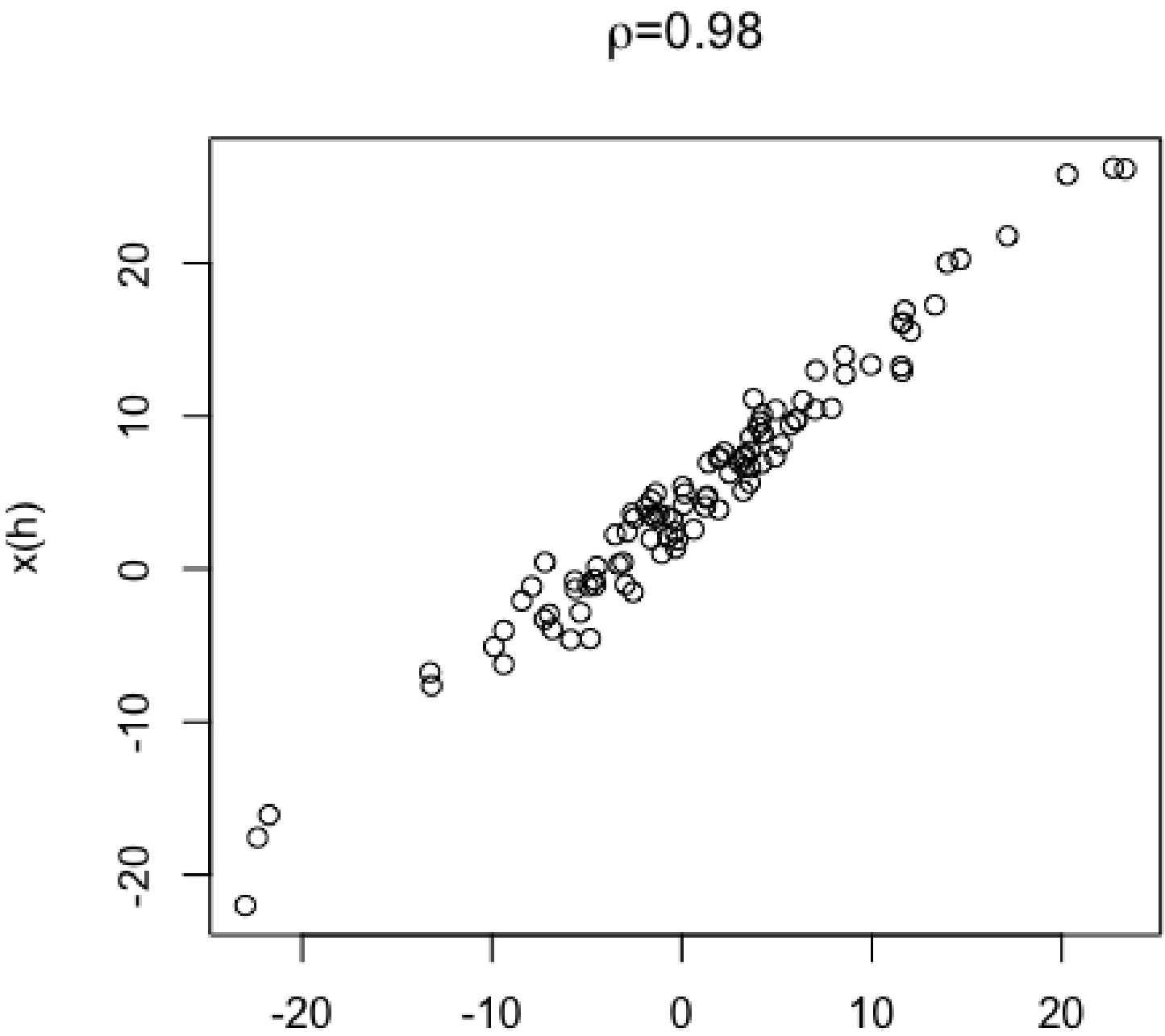
Where:

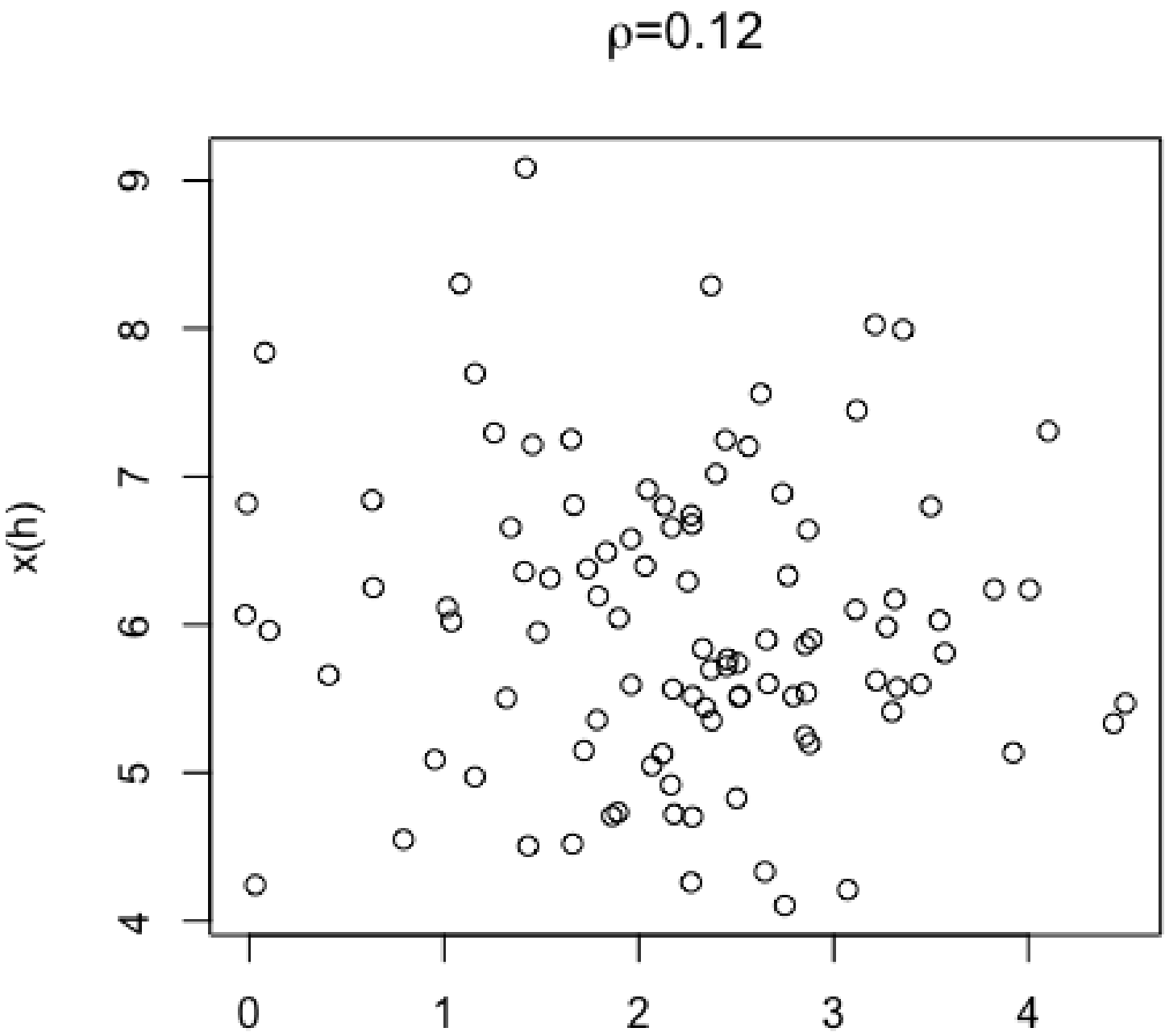
$$\begin{aligned}
\rho(x(j), x(h)) &: \text{Correlation coefficient between variables } x(j) \text{ and } x(h) \\
\text{cov}(x(j), x(h)) &: \text{Covariance between variables } x(j) \text{ and } x(h) \\
\sigma_{x(j)} &: \text{Standard deviation of variable } x(j) \\
\sigma_{x(h)} &: \text{Standard deviation of variable } x(h)
\end{aligned}$$

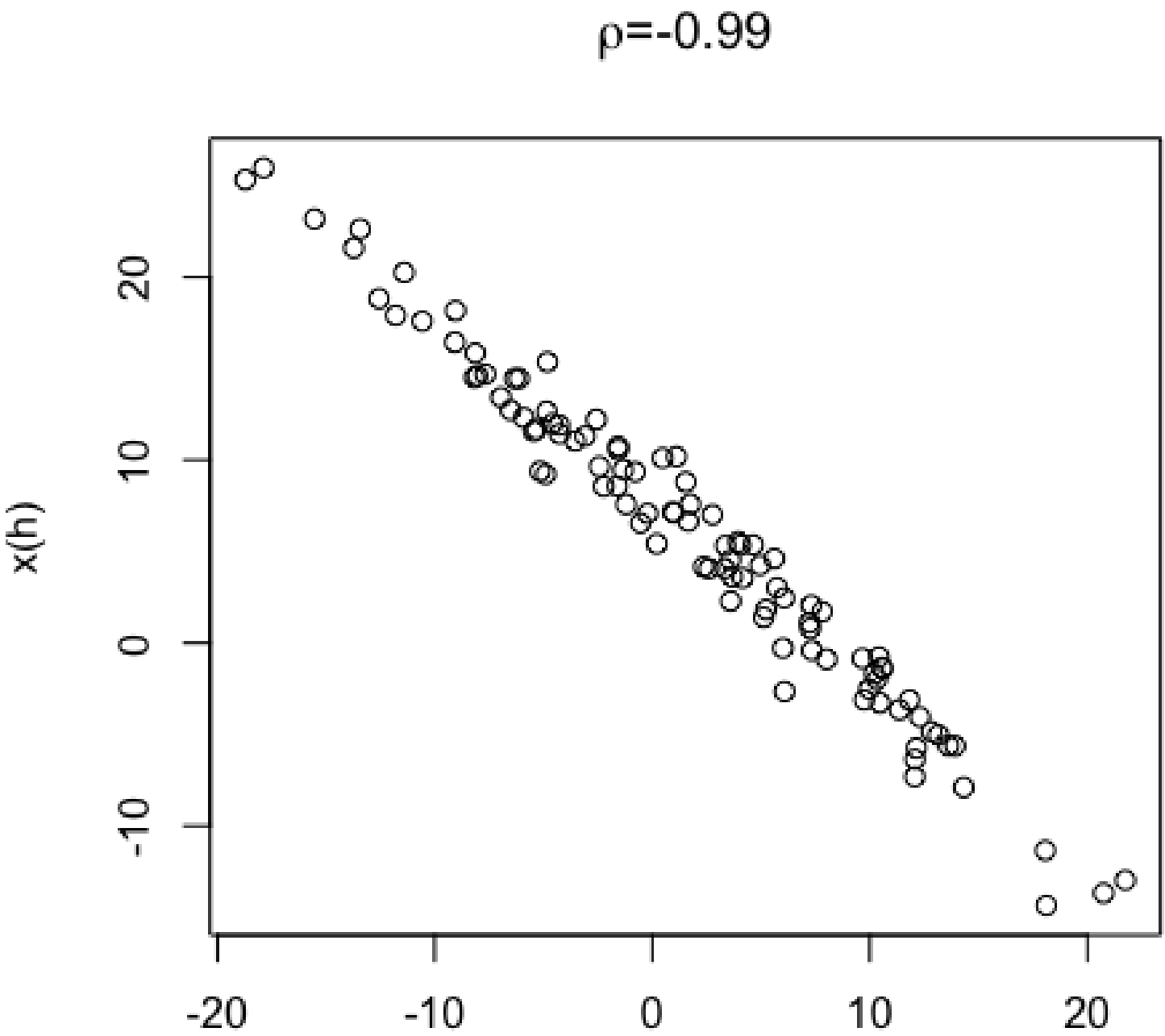
Correlation

The correlation coefficient ranges from -1 to 1:

- $\rho = 1$ indicates a perfect positive linear relationship.
- $\rho = -1$ indicates a perfect negative linear relationship.
- $\rho = 0$ indicates no linear relationship (uncorrelated).


ρ=0.98
x(h)
20
10
0
-10
-20
-20
-10
0
10
20


ρ=0.12
x(h)
4
5
6
7
8
9
0
1
2
3
4


ρ=-0.99
x(h)
20
10
0
-10
-20
-10
0
10
20

Correlation Matrix

- A correlation matrix is a symmetric matrix that provides information about the intensity and the directions of the linear relationships of multiple variables.

- The correlation matrix, denoted as R or $\boldsymbol{\rho}$, is calculated as:

$$R = \begin{bmatrix} \rho(x(1), x(1)) & \rho(x(1), x(2)) & \dots & \rho(x(1), x(p)) \\ \rho(x(2), x(1)) & \rho(x(2), x(2)) & \dots & \rho(x(2), x(p)) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(x(p), x(1)) & \rho(x(p), x(2)) & \dots & \rho(x(p), x(p)) \end{bmatrix}$$

  Where:

  - $\rho(x(j), x(j)) = 1$ for all $j = 1, \dots, p$
  - $\rho(x(j), x(h)) = \rho(x(h), x(j))$for all $j, h = 1, \dots, p$

Example

Let's consider two distributions, $x(j)$ and $x(h)$, with the following data points: $x(j) : 2, 3, 5, 7, 8$ and $x(h) : 1, 2, 4, 6, 6$. Calculate the covariance $\text{cov}(x(j), x(h))$:

- we first need to find the means:

$$M_{(x(j))} = 5 \text{ and } M_{(x(h))} = 3.8$$

- Calculate the covariance:

$$\text{cov}(x(j), x(h)) = 4.48$$

- Calculate the standard deviations

$$\sigma_{x(j)} = 2.55 \text{ and } \sigma_{x(h)} = 2.28$$

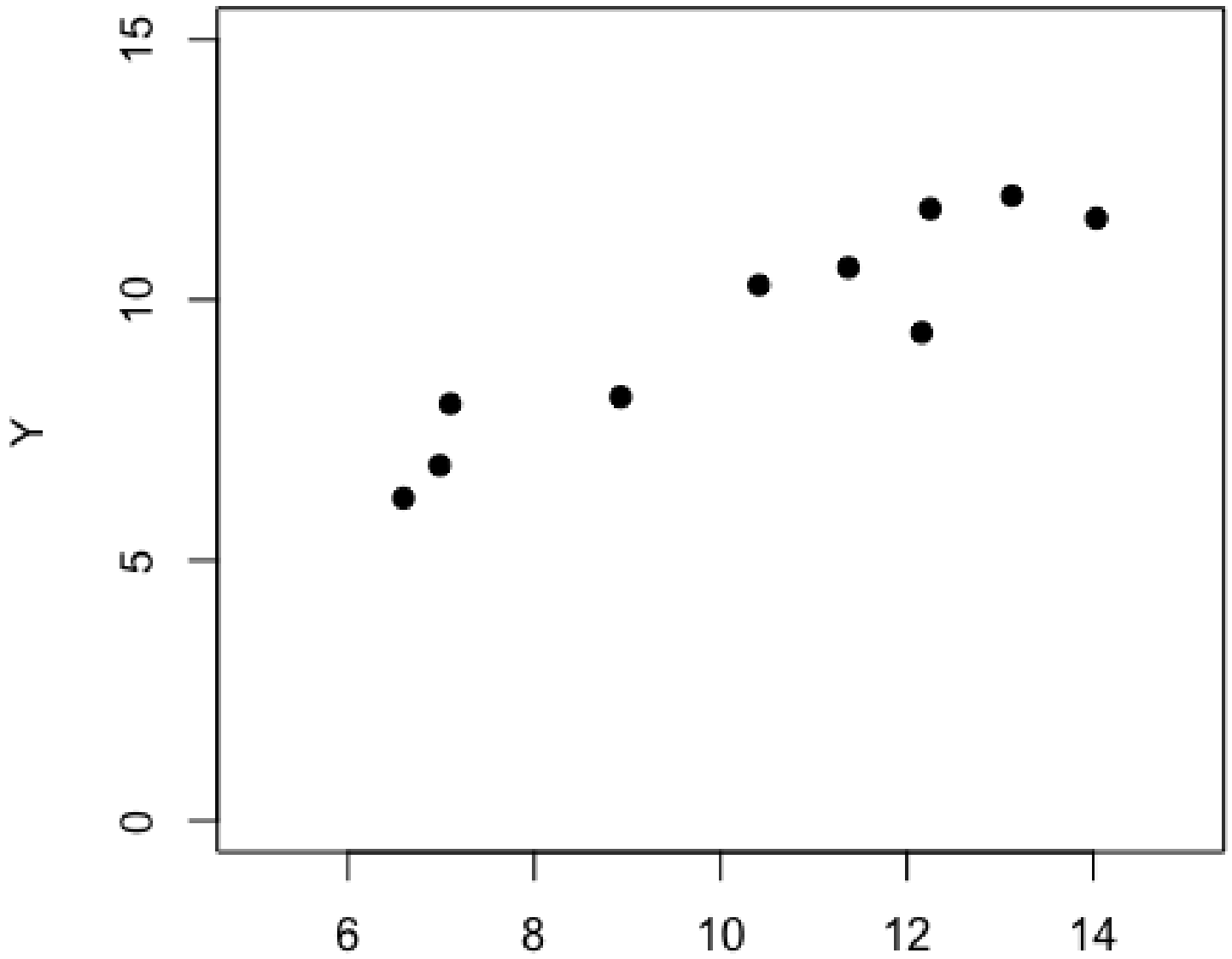
- Calculate the correlation coefficient:

$$\rho(x(j), x(j)) = \frac{4.48}{2.55 \cdot 2.28} = 0.77$$

Statistical relationship

Suppose to observe two variables X and Y (change of notation):

	X	Y
1	11.37	10.62
2	8.93	8.14
3	6.99	6.82
4	10.41	10.28
5	14.03	11.57
6	7.10	8.01
7	13.13	12.00
8	6.60	6.19
9	12.25	11.75
10	12.16	9.37


Y
0
5
10
15
6
8
10
12
14

Statistical Relationship

A statistical relationship between a dependent (or response) variable Y and an independent variable (or regressor) X is described by:

$$Y = f(X) + \epsilon,$$

where:

- $f(X)$ represents the regression functions, namely the "contribution" of the regressor to the value of the response variable Y;
- ϵ represents the error term.

Simple linear Regression

Simple linear Regression

We define the statistical relationships with linear regression function as

$$Y = a + bX + \epsilon$$

where:

- b is the "slope" of the regression line, calculated as

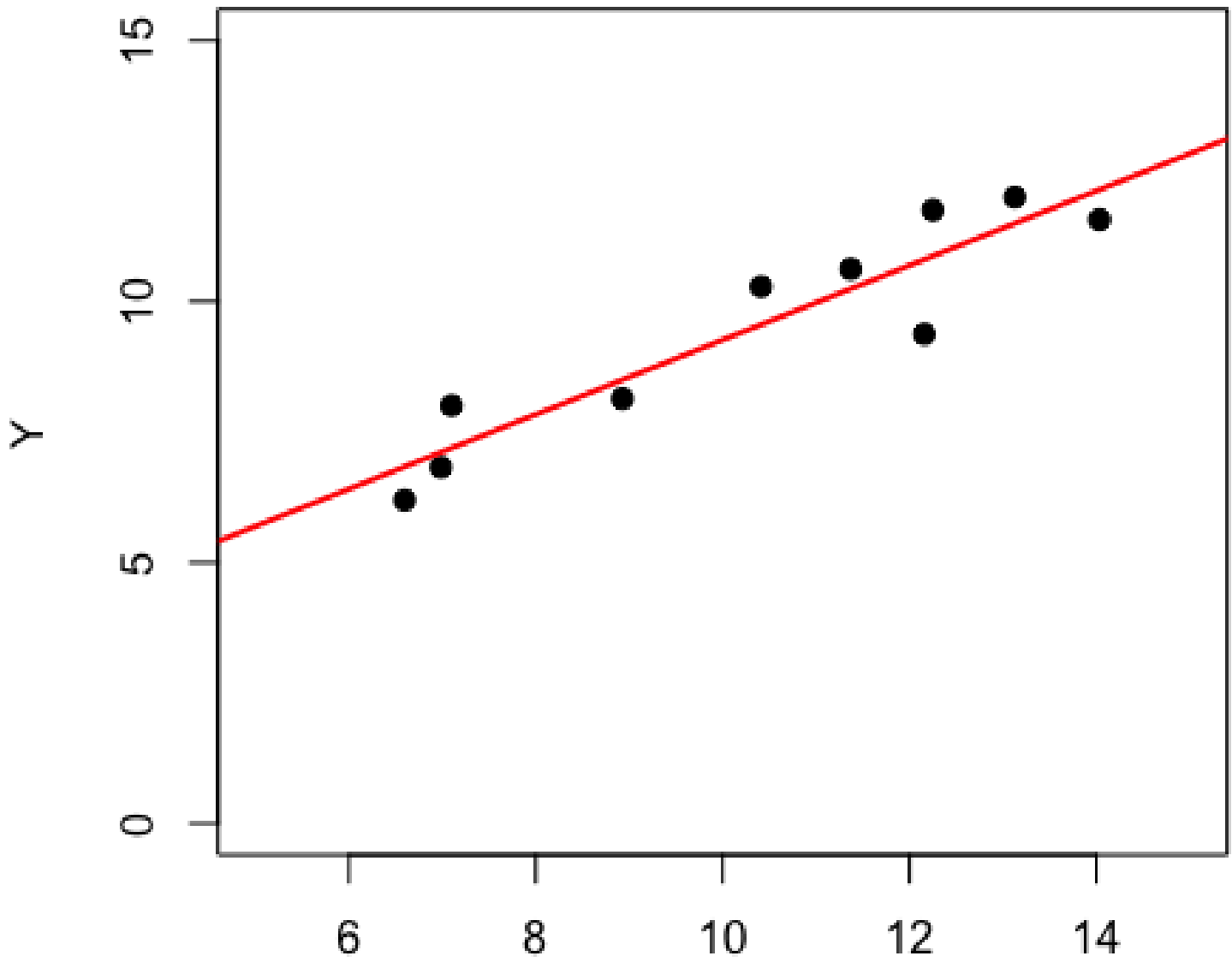
$$b = \frac{\sigma_{XY}}{\sigma_X^2}$$

.

- a is the "intercept", calculated as

$$a = M_a(Y) - bM_a(X)$$

.


Y
0
5
10
15
6
8
10
12
14

Simple linear Regression

Coefficient of determination

The coefficient of determination R^2 is the proportion of the variation in the dependent variable that is explained by the independent variable.

- $R^2 \in (0, 1)$
- $R^2 \to 0$, bad model
- $R^2 \to 1$, good model (if $R^2 = 1$ something went wrong...)
- In the univariate regression model: $R^2 = (\rho_{XY})^2 = \left(\frac{\sigma_{XY}}{\sigma_X \sigma_Y}\right)^2$.

Simple linear Regression

Coefficient of determination

The coefficient of determination R^2 is the proportion of the variation in the dependent variable that is explained by the independent variable.

- $R^2 \in (0, 1)$
- $R^2 \to 0$, bad model
- $R^2 \to 1$, good model (if $R^2 = 1$ something went wrong...)
- In the univariate regression model: $R^2 = (\rho_{XY})^2 = \left(\frac{\sigma_{XY}}{\sigma_X \sigma_Y}\right)^2$.

Back to our example:

- a=2.12
- b=0.71
- $\rho_{X,Y} = 0.93$
- $R^2 = 0.87$

Simple linear Regression: final remarks.

- Once we calculate the values of a and b, we may predict a future value of Y given a value of X.
- We have just introduced from a descriptive perspective the topic. By the end of the course, we will study the topic from an *inferential* perspective, emphasizing its statistical-mathematical properties and extending to the case with p regressors.
- Therefore, is this model good for evaluating how X *affects* Y? It's too early to talk about causality. Stay tuned!