MASTER MEIM 2023

BIG DATA ANALYTICS

EXPLORATORY DATA ANALYSIS, DATA VISUALIZATION, PREDICTION

Giovanni De Luca
Professor of Economic Statistics at Parthenope University
giovanni.deluca@uniparthenope.it

1

Introduction

- Today, providing (a sufficient level of) data literacy is almost compelling.
- MIT Sloan School of Management:
- «Data literacy is an in-demand skill for today's workforce.»
- «Data literacy means the ability to read data, work with data, acquire, clean and analyze data, and communicate what data is telling you».

2

2

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Exploratory data analysis

- The collection and storage of data are not exhaustive in themselves.
- The processing step is fundamental: it allows the achieving of the goal of supporting business decisions.
- The purpose of collecting and processing large volumes of complex data is to understand the trends of the phenomena of interest, uncover hidden trends, detect anomalies, etc., to make data-driven decisions.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Data matrix

- Structured data can be arranged in a data matrix.
- The data matrix is a two-dimensional table whose rows are associated with statistical units and columns are associated with variables.
- The statistical units can represent the entire population or constitute a (representative) sample of the population.
- We denote by n the number of statistical units and by p the number of statistical variables ($n > p$).
- The data matrix is the starting point for data analysis.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Variables

- The number of variables, even if high, has to be lower than the number of statistical units.
- There are two main types of variables:

1. numerical variables
2. categorical variables

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Numerical variables

- Numerical variables are quantitative and are classified into:
  - discrete numerical variables (derive from a counting process, they take integer values, e.g. the number of cigarettes a person smokes a day).
  - continuous numerical variables (can take any value such as heights if measured with enough precision, e.g. 68.1 or 68.09 or 68.092 inches).

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Categorical variables

- Categorical variables are presented in non-numerical form (categories), and do not allow any metric statement on the differences between categories.
- They can be:
  - Ordinal categorical variables (spiciness can be mild, medium, or hot. Even if they are not numbers per se, they can still be ordered)
  - Non-ordinal categorical variables (sex at birth, or regions of a country)

7

7

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Data matrix

- Data matrix can contain numerical as well as categorical variables.
- Sometimes, categorical variables are translated into numerical variables.

8

8

- Example: dataset *Diamonds* describes almost 54,000 diamonds using numerical and categorical variables. The data matrix has n = 53,940 rows and p = 10 columns.

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46

9

- Example: dataset *mtcars* describes the performance of 32 cars based on 11 variables. The data matrix has n = 32 rows and p = 11 columns.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

10

10

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Explorative data analysis and data visualization

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Explorative data analysis
Numerical variables

- After getting data, an explorative data analysis (or preliminary data analysis) is carried out to grasp its main characteristics.
- The analysis includes the calculation of simple statistics (summary measures) and the visualization of the data with the most appropriate graphics.
- This step is also known as pre-processing.

- The **median** (Me) is the central value (preceded by 50% of the data and followed by the remaining 50% of the data).
- R → median(x)

- The **first quartile** (Q_1) is the value preceded by 25% of the data and followed by the remaining 75% of the data.
- The first quartile equals the 25th percentile,

$$Q_1 = P_{25}$$

- R → quantile(x, 0.25)
- The **third quartile** (Q_3) is the value preceded by 75% of the data and followed by the remaining 25% of the data.
- The third quartile equals the 75th percentile,

$$Q_3 = P_{75}$$

- R → quantile(x, 0.75)

- The median is also the second quartile.
- R → quantile(x, 0.50)
- In general, the *p*-percentile is the value preceded by p% of the data.
- R → quantile(x, p/100)

15

- The **mode** is the most frequent number. It makes sense when we have discrete numerical variables.
- To find the mode, we first have to organize the data in a table.

16

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Data visualization

- Sometimes, extracting information just by looking at the numbers is quite difficult.
- Data visualization provides a powerful way to communicate a data-driven finding.
- Perhaps, data visualization is the strongest tool for explorative data analysis ("A picture is worth a thousand words“).
- “The greatest value of a picture is when it forces us to notice what we never expected to see.” (Tukey).
- Histogram.
- Box plot.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

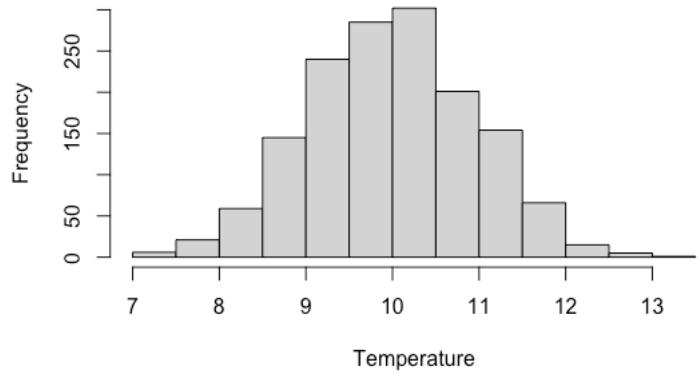
Histogram

- The **histogram** is a representation of the distribution of data.
- It is obtained by dividing the entire range of values into a series of intervals and counting how many values fall into each interval.
- The bins are non-overlapping.
- R → hist(x)

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Histogram (example 1)



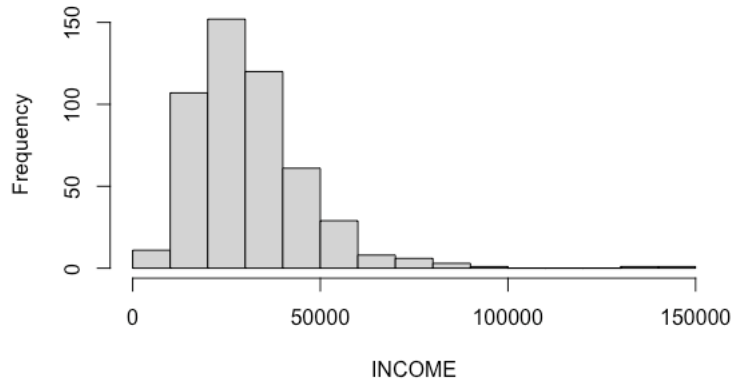
19

19

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Histogram (example 2)



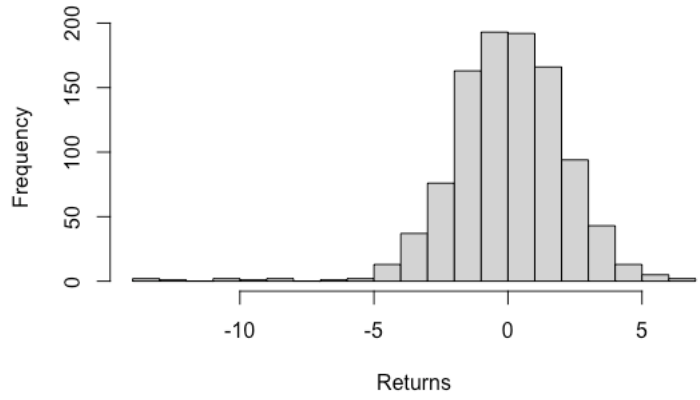
20

20

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Histogram (example 3)



MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

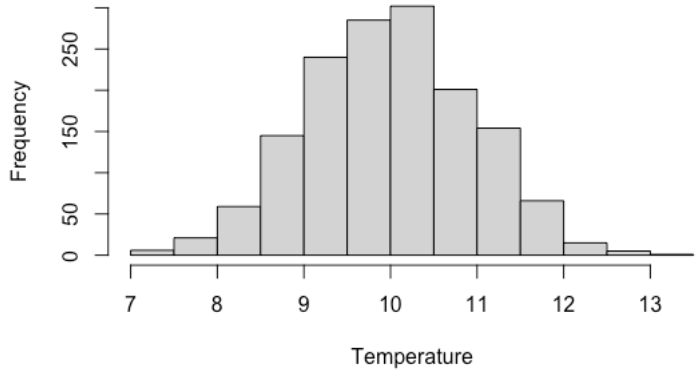
Histogram with density

- We can draw the histogram using frequency densities (on the y-axis) instead of frequencies.
- The frequency densities are computed in such a way that all histogram area is equal to 1.
- We can interpret the bins of the histogram in terms of proportions.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

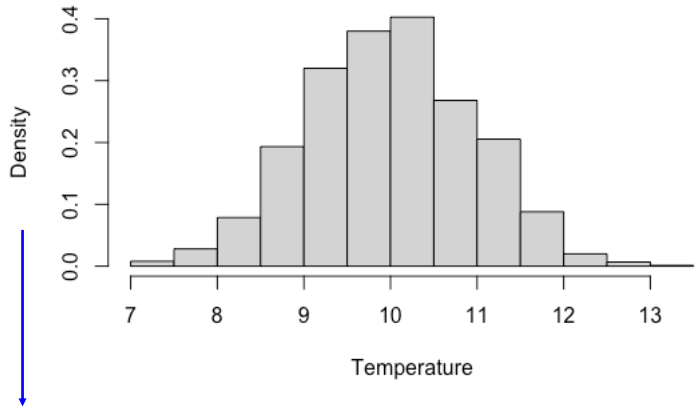
Histogram (example 1)



MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Histogram (example 1)



The area of the bins sums to 1.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

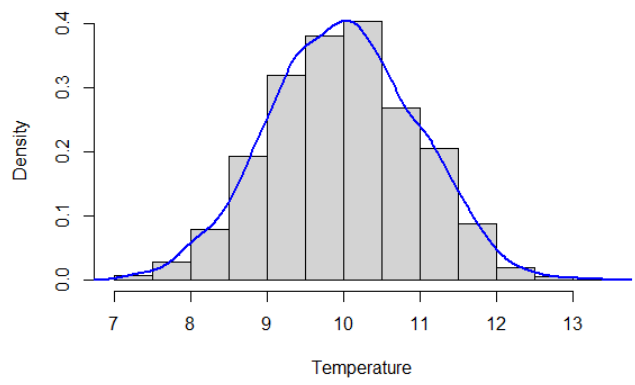
Smooth density plot

- When the histogram is represented with densities, we can draw a smooth density plot.
- Smooth density plots are similar to histograms but are aesthetically more appealing.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

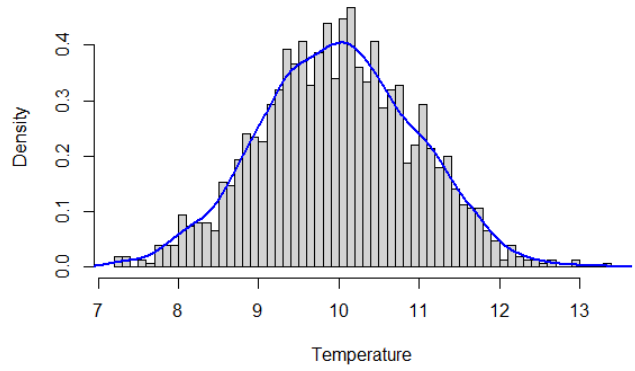
Density plot (example 1)



MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Density plot (example 1)



MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

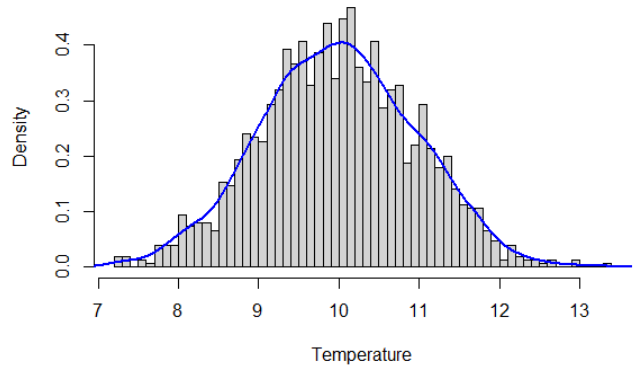
Smooth density plot

- Instead of making a histogram with very small bins, we can draw this smooth curve.
- Note that “smooth” is a relative term. We can control the degrees of smoothness of the curve.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Density plot (example 1)



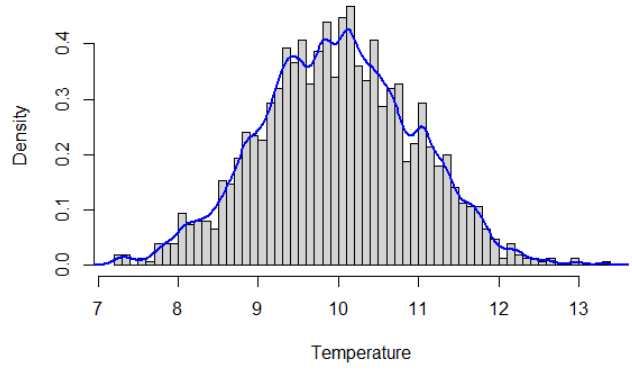
29

29

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Density plot (example 1)



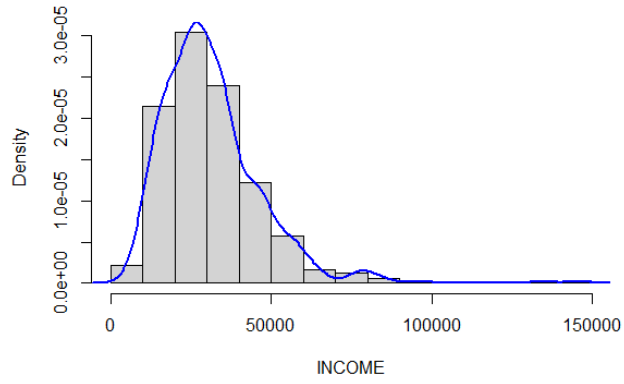
30

30

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Density plot (example 2)



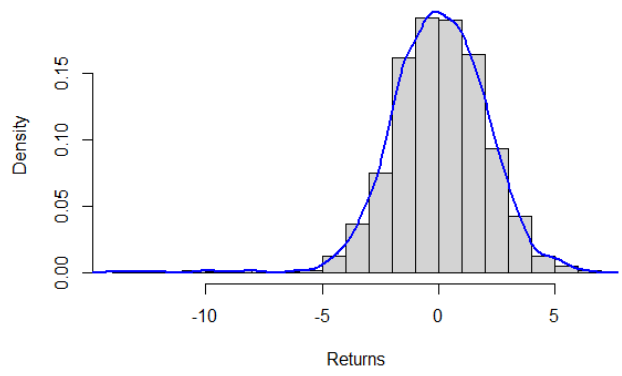
31

31

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Density plot (example 3)



32

32

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

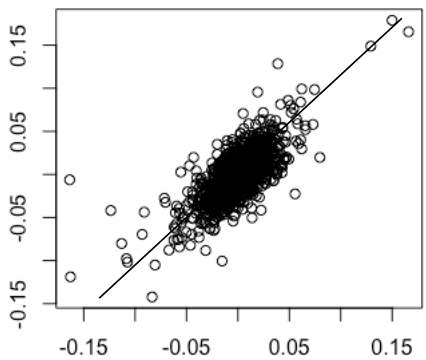
Scatterplot

- Scatterplot is one of the best known bivariate graphs.
- It highlights the (positive or negative) association between two numerical variables.
- Positive association (or positive relationship, or concordance): the two variables tend to move in the same direction and a straight line (regression line) with a positive slope can be drawn.
- Negative association (or negative relationship or discordance): the two variables tend to move in opposite directions and a straight line (regression line) with a negative slope can be drawn.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Scatterplot (example 1)

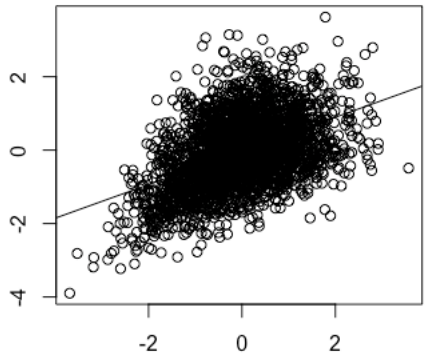


Positive association (or concordance).

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Scatterplot (example 2)



Positive association (or concordance).

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

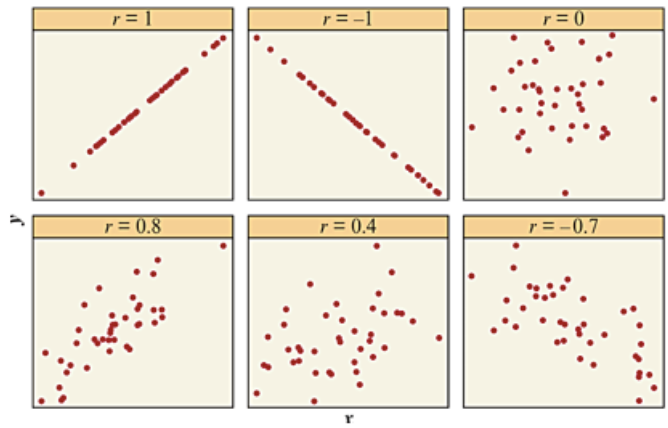
Correlation

- The correlation *r* measures the strength and the direction of the association between two numerical variables.
- Correlation always falls between -1 and +1.
- Sign of correlation denotes direction:
  - (-) indicates a negative association.
  - (+) indicates a positive association.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

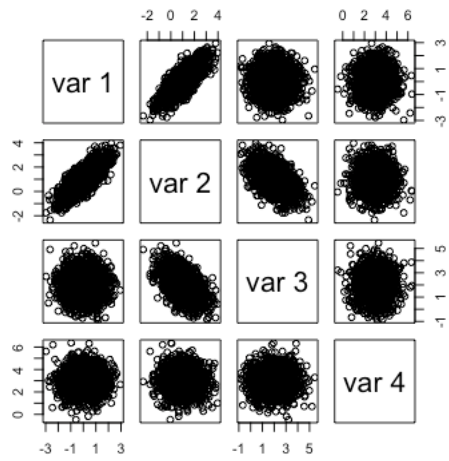
Correlation



MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Scatterplot matrix (*p*=4)

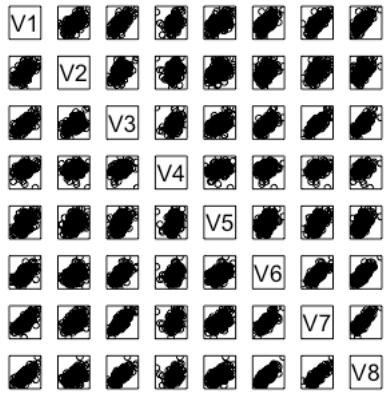


We detect:
-concordance between var1 and var2;
-no association between var1 and var3;
- no association between var1 and var4;
-discordance between var2 and var3;
-no association between var2 and var4;
-no association between var3 and var4.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Scatterplot matrix (*p*=8)



MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Summary measures for categorical variables

- The most important summary measure of a categorical variable is the **mode**.
- The **mode** is the most frequent category.
- To find the mode, we first have to organize the data in a table.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Data visualization

- For a categorical variable, we consider

1. Bar plot
2. Pie chart

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Graphical representation for categorical variables

- The **bar plot** is the graphical representation used for categorical variable.
- In the dataset *Diamonds* diamonds are classified by the categorical variables *Cut* with categories «Fair», «Good», «Very Good», «Premium», «Ideal» and *Color*, with categories «*D*» (best), «*E*», «*F*», «*G*», «*H*», «*I*», «*J*» (worst).

43

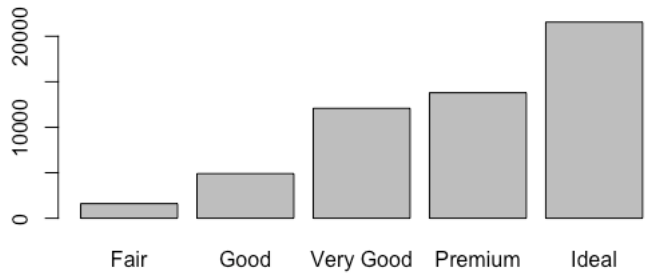
43

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Graphical representation for categorical variables

- **Bar plot** for the categorical variable *Cut*.



44

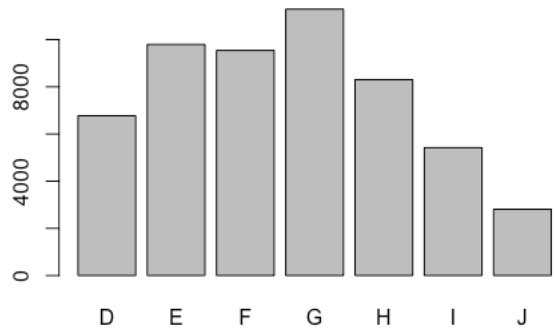
44

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Graphical representation for categorical variables

- **Bar plot** for the categorical variable *Color*.



45

45

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Graphical representation for categorical variables

- The **pie chart** is a circle having a slice of pie for each category. The size of the slice corresponds to the percentage of observation in the category.

46

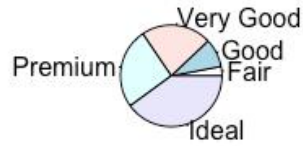
46

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Graphical representation for categorical variables

- **Pie chart** for the categorical variable *Cut*.



47

47

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Graphical representation for categorical variables

- **Pie chart** for the categorical variable *Color*.

48

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Data Handling

- Data analyst has to devote considerable time to "prepare" the data before carrying out the statistical analysis.
- Data Handling processes aim to transform raw data into data that can be effectively analyzed.
- Data Handling processes on a large amount of data are a mix of ad-hoc interventions and automatic (therefore reproducible) procedures.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Data Handling

- We deal with three cases:

1. Missing data
2. Extreme values (outliers)
3. Inaccuracies

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

1. Missing values

- Possible reasons for the lack of data:
  - misfunction of the equipment
  - error in the data-entry step
  - …
- There are two possible approaches:

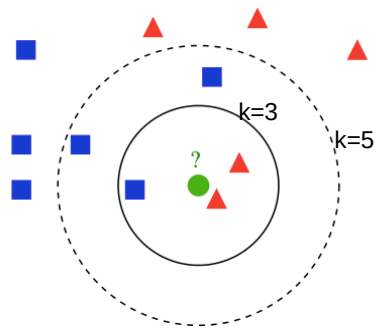
1. Leave out records with missing data (small percentage of missing data)
2. Leave out the variable (high percentage of missing data)
3. Imputation: process of estimation of missing data.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Imputation process: k-NN

- The k-nearest neighbours (k-NN) technique follows the "Learning by analogy" approach ("*Tell me who your friends are, and I'll tell you who you are*")



Imputation process: k-NN

- The procedure is sensitive to the value of *k*.
- k too small ➔we make a decisions based on very few cases
- k too high ➔ many points of other classes are included.
- A possible solution is

$$k = \sqrt{n}$$

2. Extreme values (outliers)

- There is a large literature on the detection of outliers and many definitions of outliers exist.
- Barnett and Lewis (1994) define outlier "an observation (or a set of observations) not consistent with the set of data".
- Hawkins (1980) defines an outlier as "the observation (or set of observations) that is so different from the others as to allow us to hypothesize a different generating mechanism".

Extreme values (outliers)

- Outliers can have several causes. The most common causes are:

1. Human error in data collection (can sometimes be corrected).
2. Voluntary alteration by survey participants (often linked to sensitive issues).
3. Sampling error (some sample units come from a different population than the target population).

- Be careful: some values are far from others because of the high variability in the data.

Identification of outliers

- Failure to detect an outlier can lead to an incorrect specification of the model, distorted parameter estimates and therefore incorrect results.
- There are graphical and analytical techniques for detecting outliers in numerical variables.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Outliers: box plot

- The **box and whiskers plot** (or **box plot**) is a graphical representation to describe the main features of a numerical variable and can be used to detect the presence of outliers.
- The central box of the graph indicates:
  - 1° quartile of X (Q_1)
  - Median of X, Me(X)
  - 3° quartile of X (Q_3)
- Two whiskers are drawn.
- The points outside the whiskers are possible outliers.
- I remove them if they are «very far» from the others.

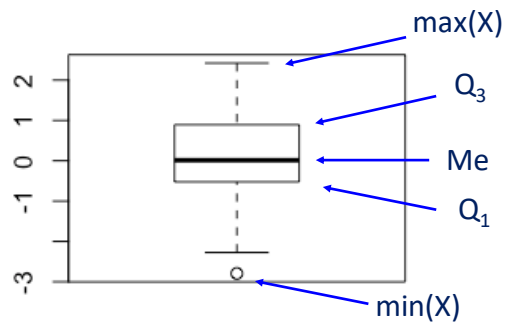
57

57

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Box plot (example 1)



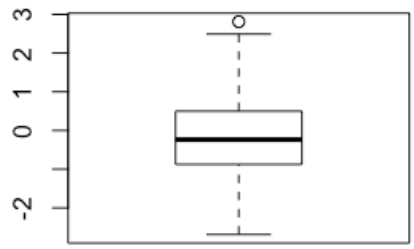
58

58

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Box plot (example 2)



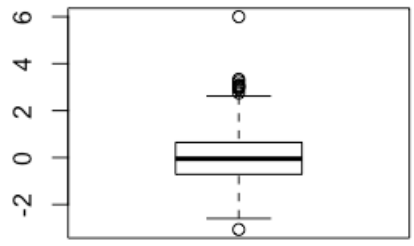
59

59

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Box plot (example 3)



60

60

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

3. Inconsistencies

- The inconsistency does not have a precise definition. Anything that is not consistent or logical falls into this case.
- For example, the categorical variable *Gender at birth* has two categories: male and female. If categories are more than two, it is necessary to do a merge (for example, the categories are male, female, M, F).
- It is possible to find *Pepsi Cola* and *Pepsi*. We need to merge.
- A negative height is an inconsistency and has to be removed (or to be converted to a positive number).

61

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

62

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Supervised learning

- Many problems of statistical learning require a supervised learning technique.
- *Supervised learning*: for each statistical unit we know the response variable, Y and p predictors $X_1, X_2, \dots, X_p$.
- The interest lies in the analysis of the relationship between the predictors and the response variable with the aim of predicting the latter for new observations.

Supervised learning techniques

- Classical linear regression, polynomial regression, spline regression, logistic regression, decision tree.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Unsupervised learning

- Unsupervised learning: for each statistical unit we have *p* variables, but no response variable is observed.
- In this case, the goal is the study of the relationship between the variables or between the observations, or grouping the observations into distinct groups.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Linear regression

- Linear regression is a useful and widely used statistical learning method.
- Linear regression analyzes the response of a numerical variable (response variable or dependent variable) to *p* numerical variables (predictors or explanatory variables).
- Simple linear regression: 1 predictor.
- Multiple linear regression: *p* predictors.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Simple linear regression

- In simple linear regression, we denote

  Y = response variable

  X = predictor

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Example of simple linear regression

- Suppose we want to study the sales as a function of Facebook advertising expenses.
- Y = sales
- X = Facebook advertising expenses

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Simple linear regression

- Simple linear regression is given by

$$Y = \beta_0 + \beta_1 X + e$$

- Y = sales (response variable)
- X = Facebook advertising expenses
- β_0 and β_1 are parameters of the model
- e = error
- n statistical units

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Error

- e is the error which prevents from defining a deterministic relationship between Y and the predictor.
- Deterministic relationship between Y and X: only one value of Y is associated to a specific value of X.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

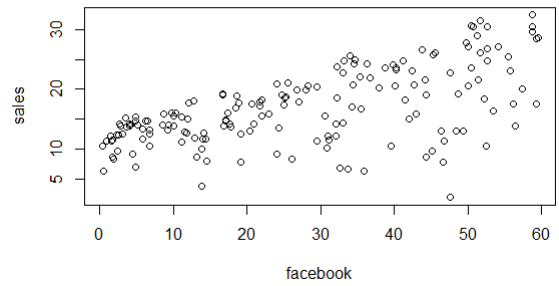
Estimate

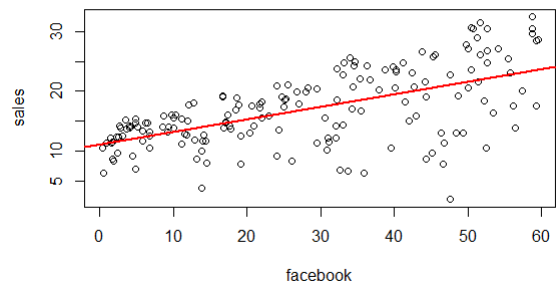
- Parameters β_0 and β_1 are estimated using the Ordinary Least Squares (OLS) method.
- Predicted values are $\hat{Y} = b_0 + b_1 X$
- b_0 and b_1 are the estimates of the parameters β_0 and β_1
- $\hat{Y} = b_0 + b_1 X$ is the regression line.

Parameters

- Parameter b_0 (also known as intercept) is the predicted value of Y when X = 0.
- Parameter b_1 is the predicted change of Y when X increases by one unit.
- If $b_1 > 0$ (< 0), there is a positive (negative) association between $\hat{Y}$ and X.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN
UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE





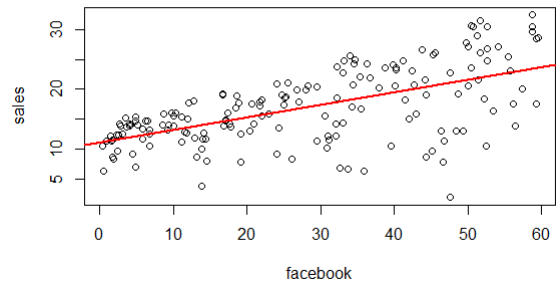
$$\hat{Y} = b_0 + b_1 X$$

$$\widehat{sales} = 11.07 + 0.210\ Facebook$$

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Predicted values



- $Facebook = 55$
- $\widehat{sales}(55) = 11.07 + 0.210 \cdot 55 = 22.609$

75

75

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

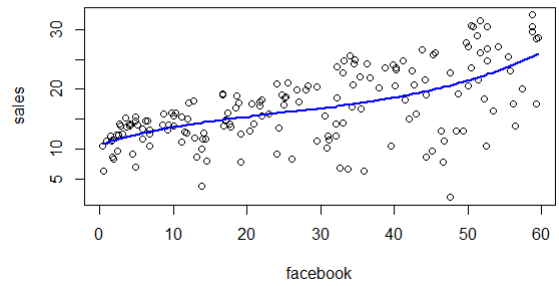
UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Polynomial regression

- Linear regression involves a linear relationship between the response variable and the predictors.
- Sometimes, the relationship is not linear.
- Polynomial regression is a simple strategy to extend a linear model to capture a non-linear relationship.
- In practice, polynomial regression adds further terms given by some power of the original predictors.
- Non-linear relationship: relationship not adequately represented by a straight line.

76

76



77

Improvement of the predictions

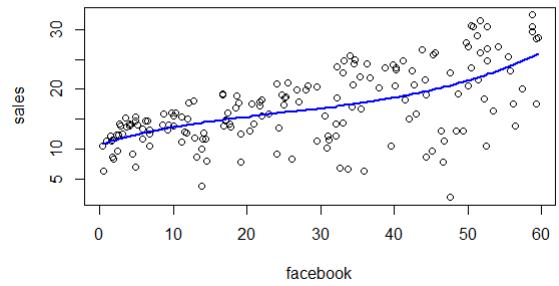
- The choice of a better model involves an improvement of the predictions.

78

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Predicted values



- $Facebook = 55$
- $\widehat{sales}(55) = 23.542$

79

79

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Multiple linear regression (two predictors)

- Multiple linear regression with two predictors is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

- Y = sales (response variable)
- X_1= Facebook advertising expenses (1st predictor)
- X_2 = Newspaper advertising expenses (2nd predictor)
- β_0, β_1 and β_2 are parameters of the model
- e = error
- n statistical units

80

80

Predicted values

- Predicted values are given by

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

- This is not a regression line, but a regression plane!

Parameters

- Parameter b_0 (also known as intercept) is the predicted value of Y when $X_1 = X_2 = 0$.
- Parameter b_1 is the predicted change of Y when X_1 increases by one unit, if X_2 remains constant.
- If $b_1 > 0$ (< 0), there is a positive (negative) association between Y and X_1.
- Parameter b_2 is the predicted change of Y when X_2 increases by one unit, if X_1 remains constant
- If $b_2 > 0$ (< 0), there is a positive (negative) association between Y and X_2.

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Adjusted R-squared

- The goodness of fit is measured using $\bar{R}^2$ which is a measure of the fit but also takes into account the number of the parameters of the model (model complexity).

MEIM MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Example

- Let's try to predict the variable *sales* using *Facebook advertising expenses* and *Newspaper advertising expenses*.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

$$\widehat{sales} = b_0 + b_1 Facebook + b_2\ Newspaper$$

$$\widehat{sales} = 10.68 + 0.203\ Facebook + 0.016\ Newspaper$$

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

Multiple regression
Selection of the predictors

- In presence of many (potential) predictors, we can use a function (*step*) that automatically selects the useful variables, defining the «best» model.
- First, we estimate the complete (with all the possible predictors) model, then we apply the function *step* and identify the best model.

MASTER IN ENTREPRENEURSHIP INNOVATION MANAGEMENT IN COLLABORATION WITH MIT SLOAN

IN COLLABORATION WITH MIT MANAGEMENT SLOAN SCHOOL

UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

MASTER MEIM 2023

Thank you for your attention