Applied Statistics and Data Analysis – Notes –

Andrea Mansi – UniUD

I° Semester – 2020/2021

Contents

Info about the notes

Main contents covered in the notes:

- **Exploratory data analysis:** Data summaries and graphical visualizations to perform a preliminary data exploration. It is the first and (often) the most important step in data analysis.
- **A review of inference concepts:** A quick review of the basic material that is fundamental to statistical inference, ranging from the notions of random sample, statistical model and sampling distribution to the basic concepts of point estimation, confdence interval and test of hypotheses, with the associated measures of statistical accuracy.
- **Linear regression with a single predictor:** This simple regression model may be appropriate for data that can be displayed as a scatterplot. Although the focus is on the straight line model, it is possible to accommodate some specific non-linearities by means of suitable transformations. Many issues are fundamental for any study based on regression methods.
- **Towards multiple linear regression and logistic regression:** Multiple linear regression generalizes the strait line model with a single predictor to allow multiple explanatory or predictor variables. Regression models may be extended to account for non-Gaussian response variables giving, in particular, Bernoulli, Binomial or Poisson distributed outcomes.
- **Predictive and classification methods:** Predictive methods encompasses a variety of statistical techniques aiming at analyzing the available data to make predictions about future or otherwise unknown events. Particular attention is dedicated to predictive methods based on regression models, with quantitative or categorical response variables. The process for predicting qualitative responses is called classifcation.
- **Unsupervised methods (principal component analysis, cluster analysis):** In unsupervised statistical methods there is no outcome (response) measure and the goal is to describe the associations and patterns among a set of input measures. Principal component analysis is a method which projects the data on to a low-dimensional space, commonly two dimensions, suggesting views of the data that may be insightful. Cluster analysis refers to a wide set of techniques for finding subgroups, or clusters, in a data set.

Bibliografy: - UniUD - Prof. Paolo Vidoni: Slides & Lectures

Introduction

Let's start whit the definition of "Statistics":

> Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty.
>
> *Davidian, M. and Louis, T. A.*

Statistical methods are developed in a synergy with the relevant supporting mathematical theory, and more recently with computing, and they are applied in a wide variety of scientifc, social, and business frameworks.

Statistical methods provide crucial guidance in determining what information is reliable and which predictions can be trusted.

Statistical methods often help search for clues to the solution of a scientifc mystery and sometimes keep investigators from being misled by false impressions.

- What do statisticians do?

The world is becoming quantitative and more and more professions depend on data and numerical reasoning.

Statisticians are experts in producing trustworthy data, analyzing data to make their meaning clear and drawing practical conclusions from data.

Statisticians work with people from other professions to solve practical problems and they must know more than statistics.

- New computing tools

The recent advances in statistical computing methodology have made possible the development of new powerful tools for data analysis and prediction.

New types of data and data sets of unprecedented size (e.g. textual data, image data), combined with new data analysis demands, boost the development of new hybrid data analysis approaches, such as machine learning, data mining and analytics.

However, the traditional concerns of professional data analysts remain as important as ever. The size and the complexity of data set are not itself a guarantee of quality and of relevance to issues that are under investigation.

No amount of statistical or computing technology can be a substitute for skill in the use of statistical analysis methodology.

Statistical software systems are one of several components of effective data analysis.

- The main steps of a statistical analysis

- Formulation of the problem: understand the background, specify the objectives of the analysis, put the problem into statistical terms.
- Collection and organization of data: observational or experimental data, missing values, units of measurements, codifcation of data, organization of data.

- Initial (exploratory) data analysis: numerical and graphical summaries to get into data.
- Data analysis.
- Presentation of the results.

> The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill
>
> *Albert Einstein*

Statistics as a way of life

Statistics and Machine Learning fill different places in the large ecosystem of Data Science. Statistics and machine learning share several techniques, but there are some crucial differences between the two disciplines. In our view, some notable points of the Statistics way are as follows:

- **The importance of context:** Statistics is always concerned with the data context, so that data analyses are never fully automatic. A preliminary understanding of the way data are produced is essential for any further step. Solutions can be adapted to a different setting with the necessary changes, but only after studying the context, and never automatically. This approach is less fruitful or even misplaced for some tasks which can be solved by a purely algorithm approach, such as some pattern recognition tasks.
- **The role of principles:** Statistical principles are of paramount importance. They range from optimality in estimation/prediction to principles adhering to likelihood theory and Bayesian theory. Sometimes they are inspired by highly-stylized settings, but they provide guidelines which are useful also in intricate real-life scenarios.
- **Statistical models, aka making sense of data:** Statistical models are of central importance. They are mathematical description of the data generating process, and they include both deterministic and random components. Even methods which are strongly algorithmic in nature (e.g. regression trees or LDA in text mining) are often interpreted as based on statistical models, with the possibility of applying general statistical principles, such as model selection criteria or Bayesian inferential techniques. Statisticians usually endorse the Occam's razor principle, avoiding overly complex models unless the data available are large/rich enough. (That's why we are not so fond of neural networks, which are sometimes misused nowadays).
- **A lingua franca:** Statisticians, with few exceptions, have adopted the open source R software. R is both a statistical software as well as a programming language, with interfaces to many data mining/ML software, such as Weka and H2O. (Despite what many people outside statistics believe, R is far more powerful and versatile than Matlab, it's not just a free version of it). Having a common language has simplifed things a lot within the Stat community.

1 Exploratory Data Analysis

To begin investigation of a new set of data the following points have to be considered.

- Graphs are one of the most important tools.
- Numerical summaries are also important, but they don't go very far without an appropriate graph.
- Statistical models can clarify the information content of data and make prediction possible. But models require assumptions, which must be checked, often via graphical methods.
- One should not over-analyze the data, as "under torture the data may yield false confessions".
- The way data were collected should always kept into account.
- The use of graphs (and of numerical summaries) to display and get insight into data has a long tradition.
- A key point is that data should reveal their information content prior to (or as a part of) a formal analysis.
- Modern computing tools have greatly improved on traditional techniques.

Exploratory Data Analysis (EDA) has at least four important roles:

- It may suggest ideas and understandings not previously considered.
- It may challenge the theoretical knowledge that guided the initial collection of the data.
- It allow the data to cast doubt on an intended analysis and to facilitate checks on assumptions.
- It may reveal additional information and further lines of research.

1.1 Statistical units and variables

Data analysis requires the data to be organised into an ordered database. It is important to identify the statistical units, i.e. the elements in the population (or in the sample) that are considered in the analysis, and the variables, i.e. the characteristics measured for each unit. Two diferent types of variables may defined: **categorical (qualitative)** variables and **numerical (quantitative)** variables.

Categorical variables can be classifed into **nominal**, if the categories do not follow any particular order (i.e. gender, religion professed) or **ordinal**, if the different categories are ordered (i.e. computing skills of a person, credit rate of a company).

Numerical variables are classified into **discrete**, if they have finite or countable potential values (i.e. family size, number of car accidents) or **continuous**, if they can take any value between two real numbers (i.e. height, time spent waiting in a queue).

1.2 Data structures

A simple way to express a database is in terms of a data matrix, where the rows represent the statistical units and the columns represent the variables.

An analysis focusing on a single variable is called **univariate**, whereas it is called **multivariate** whenever two or more variables are jointly considered.

Data structures, more complex than a matrix, arise in a number of frameworks as, for example, when temporal and/or spatial information are relevant (longitudinal and/or spatial data). Further relevant examples are related to text data (a database of text documents, usually related to each other), web data (e.g. log files on the behavior during a web session) and multimedia data (e.g. texts and audio-visual information).

Another challenging situation concerns data structures obtained from the integration of different databases.

1.3 Views of an univariate data set: categorical data

Simple characterizations of the data, and in particular of the frequency distribution, in terms of graphs. Data related to categorical variables are usually described in the form of tables and thus summarized by calculating frequencies. For presentation purposes, table of counts or percentages may be described using the following graphical summaries:

- **barplots:** the height of each bar is proportional to the (relative) frequency of the associated category;
- **piecharts:** the arc length of each slice (and consequently its area) is proportional to the frequency of the associated category.

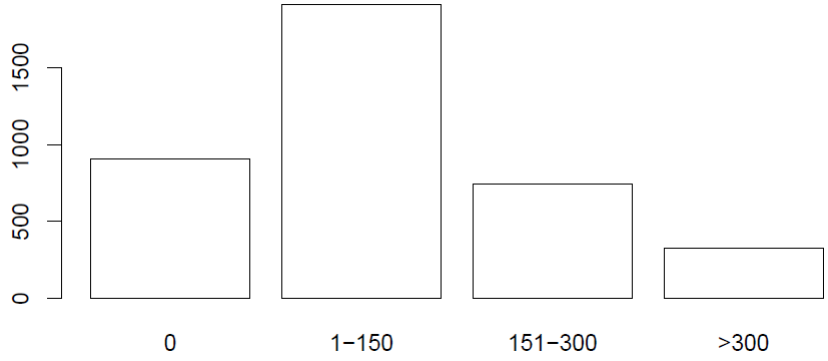
1.3.1 Example: Caffeine and marital status

A two-way table containing data on caffeine consumption (mg/day) by marital status (married, previously married, single) among 3888 pregnant women: observed frequencies and row percentages.

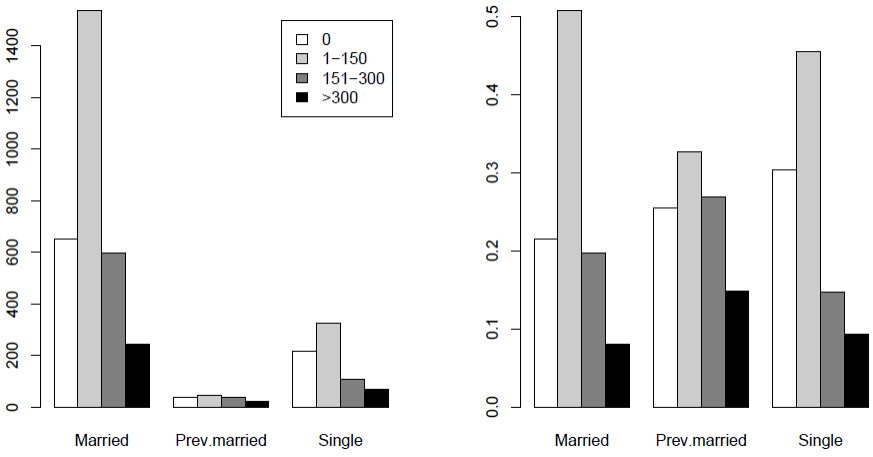
	caffeine	consumption	(mg/day)		
marital status	0	1-50	151-300	> 300	Total
married	652	1537	598	242	3029
prev. married	36	46	38	21	141
single	218	327	106	67	718
Total	906	1910	742	330	3888

	caffeine	consumption	(mg/day)		
marital status	0	1-50	151-300	> 300	Total
married	0.22	0.51	0.20	0.08	1
prev. married	0.26	0.33	0.27	0.15	1
single	0.30	0.46	0.15	0.09	1
Total	0.23	0.49	0.19	0.08	1

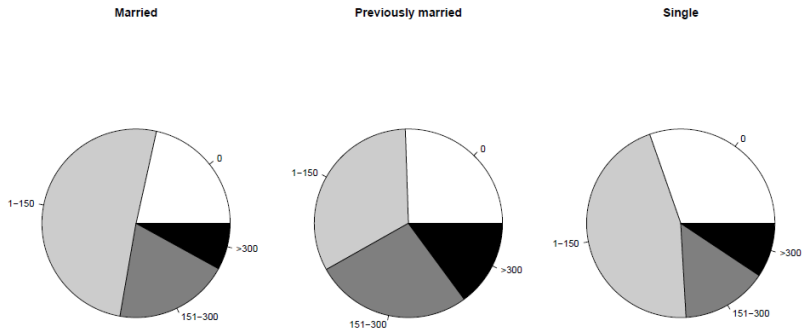
Simple barplot of total caffeine consumption (observed frequencies):



Multiple barplots: observed frequencies (left), row percentages (right). Row percentages enable the comparison of caffeine consumption among the three groups.



Piecharts of caffeine consumption according to marital status:



1.4 Views of an univariate data set: numerical data

Simple graphical representations of numerical variables (both continuous and discrete), and in particular of the frequency distribution of the associated data set, include:

- **histograms:** the area of each rectangle is proportional to the frequency of the observations that lie within the base of the rectangle;
- **density estimate:** for continuous data, an histogram is rough form of density estimate; a better, smooth alternative is a (kernel) density estimate;
- **boxplot:** a notable graphical summary of the data set emphasizing median, quartiles and potential outliers.

Histograms are basic EDA tools for displaying the frequency distribution of a data set. A symmetric and regular histogram often hints to a normal distribution for the data (the "bell curve").

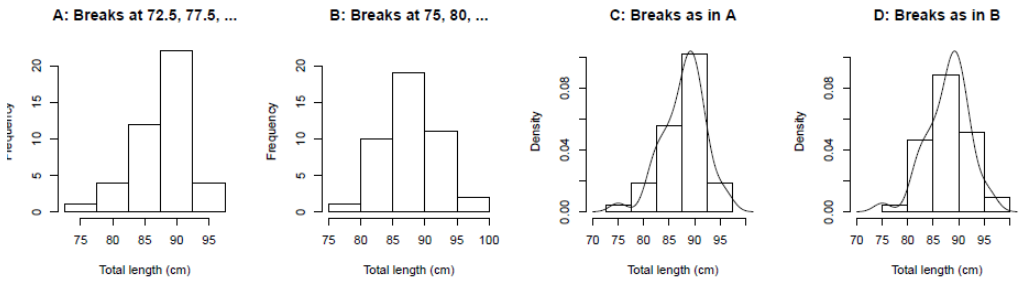
In small samples, the shape can be highly irregular, and the appearance may depend on the choice of breakpoints. A better alternative is, often, a smooth density estimate. Like width of histogram bars, that have to be chosen subjectively, density estimates require the choice of a bandwidth parameter that tunes the amount of smoothing. Software default choices often work well.

Density curves are preferable to histograms for drawing attention to particular forms of non-normality.

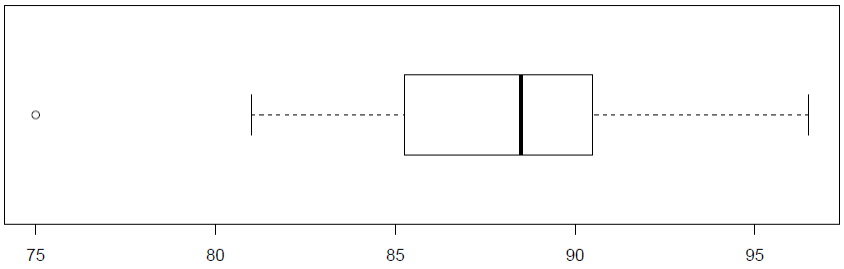
1.4.1 Example: possum dataset

The complete data set has nine morphometric measurements on 104 mountain brushtail possums. Here, attention will be limited to the length (cm) measurements for the 43 females.

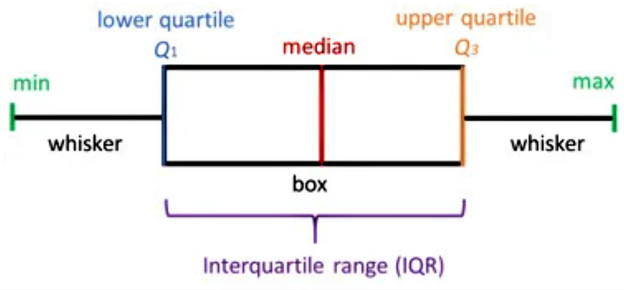
Alternative breakpoints for the histograms suggest different conclusions for the frequency distribution.



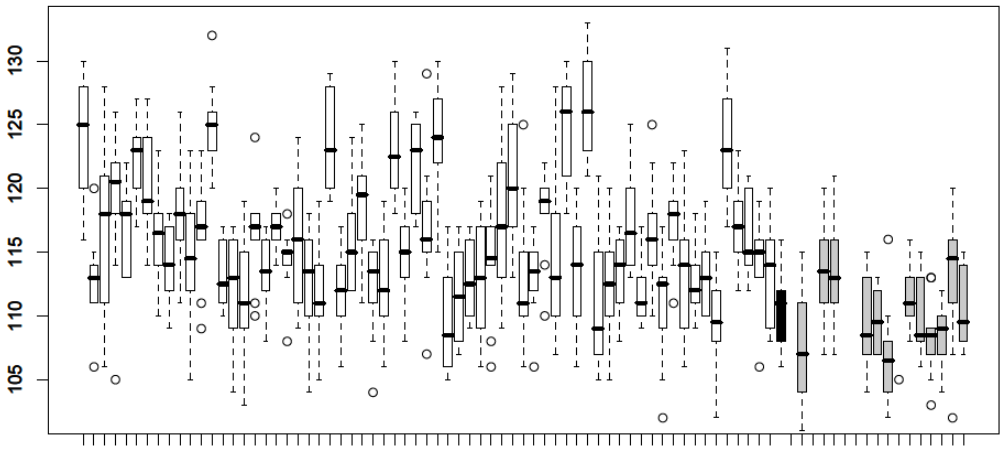
Boxplots allow a comprehension of specific important features of the data at a glance. Indeed, they are useful to identify outliers. With regard to the possum data set and the length measurement:



The whiskers range from the smallest to the largest value (outliers excepted), while the box is defined by the lower quartile and the upper quartile, with indication of the median.

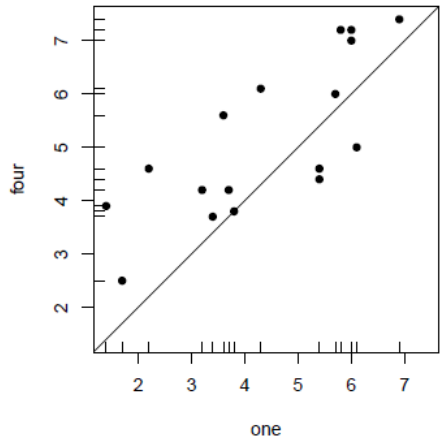


Boxplots are very useful to compare different data sets.



1.4.2 Example: milk sweetness dataset – patterns in bivariate numerical data

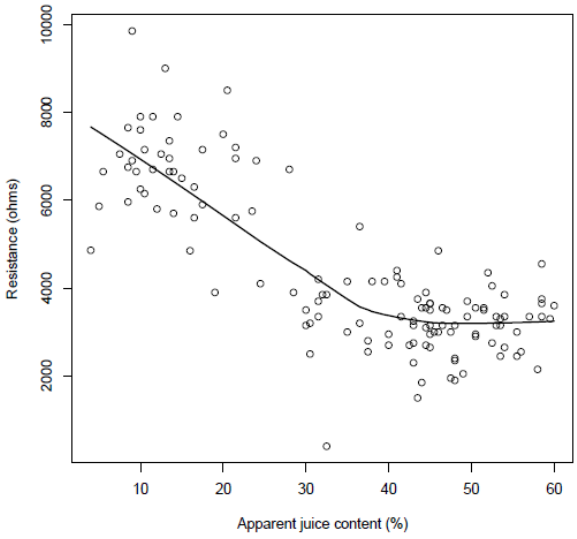
Data from a tasting session where each of 17 panelists assessed the sweetness of two milk samples, one with four units of additive, the other with one unit of additive.



The line $y = x$ assists in comparing the two samples: most panelists rated the sample with four units as sweeter than the sample with one unit.

1.4.3 Example: kiwi electrical resistence dataset – Adding a smooth trend curve

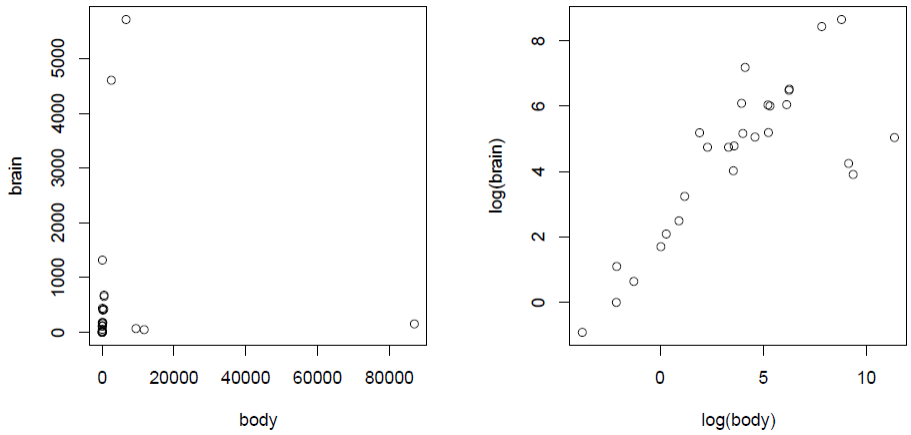
Data from a study that measured both electrical resistance (ohms) and apparent juice content (%) for a number of slabs of kiwifruit.



The curve estimates the relationship between electrical resistance and apparent juice content, which is nonlinear.

1.4.4 Example: brain weight against body weight – what is the appropriate scale?

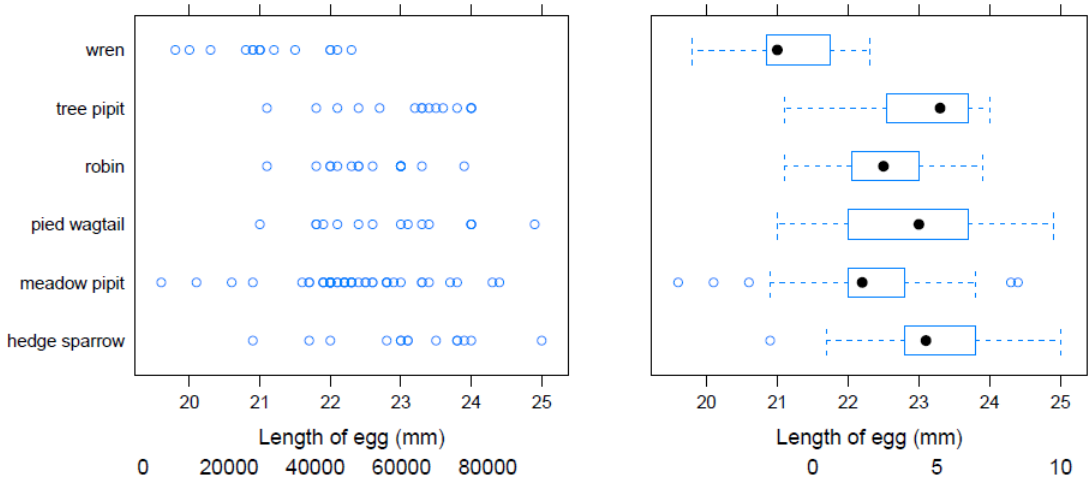
Data of brain weight (g) against body weight (kg), for a number of different animals: untransformed scale vs logarithmic scale:



Logarithmic scale is appropriate for quantities that change multiplicatively, like cells in growing organisms.

1.5 Patterns in grouped data

Cuckoos lay eggs in the nest of other birds, which adopt and hatch the eggs; the egg lengths (mm) are grouped by the species of the host bird and strip plots (left) and boxplots (right) are useful for side-by-side comparisons:

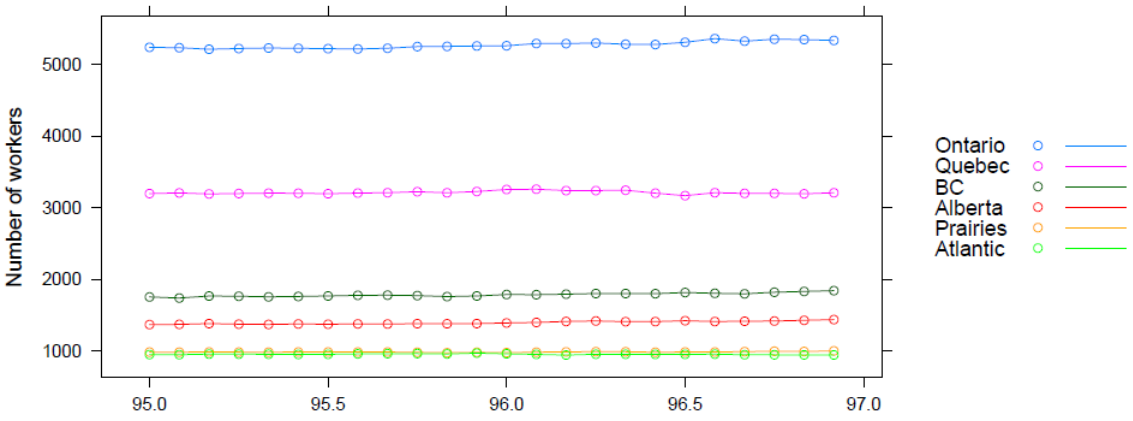


Eggs in wrens's nests appear smaller than eggs planted in other birds's nests; there are several outlying egg lengths in the meadow pipit nests.

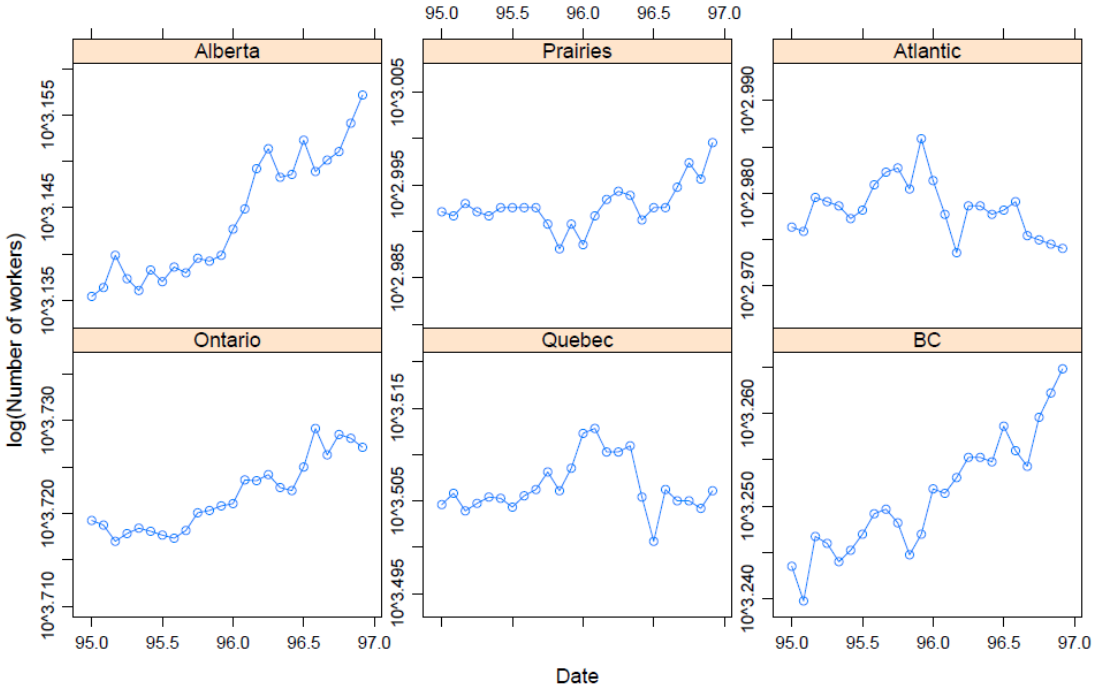
1.6 Temporal data: comparing server time series

Overlaying plots of the time series might seem appropriate for making direct comparisons, provided that the scales are similar for the different series.

Multiple time series of number of workers (in thousands) in Canadian labor force, from 6 regions, from January 1995 to December 1996



The labor forces in the various regions do not have similar sizes so that it is impossible to discern any differences among the regions. In order to account for different sizes and to consider relative changes: six different panels using the logarithmic scale:



1.7 Scatterplots, broken down by multiple factors

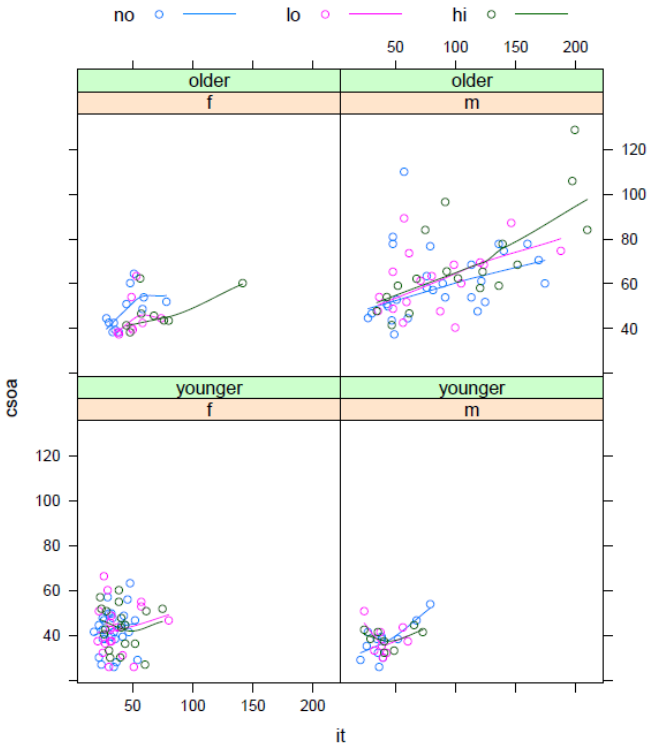
Data from an experiment on the effect of car window tinting on visual performance.

Numerical variables: csoa (time in msec to recognize a target), it (time in msec for a simple discrimination task) and age (to the nearest year).

Categorical variables (factors): tint (3 ordered levels, no, lo, hi), target (2 levels, locon, hicon), sex (2 levels, f,m) and agegp (2 levels, younger, older).

Each of 28 subjects was tested at each level of tint for each of the two levels of target, and all in all there are four factors that may influence two response variables csoa and it, and also the relationship existing between them.

Plot of csoa against it for each combination of sex and agegp, with different colors depending on whether the tint is absent, low or high.



The longest times are usually for the high level of tinting, the relationship between csoa and it seems much the same for different levels of tinting.

1.8 What plots may reveal

- **Outliers**: points that look isolated from the main body of the data; they may convey quite useful information, but are not easy to detect in high-dimensional data sets; they may depend on the scale.
- **Asymmetry of the distribution:** positive or negative skewness is often encountered with socio-economic data; symmetry is preferable, and it is typically obtained by transforming the data; kurtosis, i.e. heaviness of the tails of the distribution, is a further key feature.
- **Changes in variability:** usually not diffcult to detect by graphical summaries; when variability increases as the data values increase, the log transformation (or the square root one) is usually a good idea.
- **Clustering:** clusters of separate points may be quite informative, and often correspond to the values of some relevant variable; it may also happen that such variable is not recorded.
- **Nonlinearity:** linear relationships are often a good approximation to some more complex forms; choosing the right scale may be crucial, as multiplicative relationships become linear on the log scale; sometimes the relationships are inherently non-linear.

1.9 The importance of data summaries

There are at least three reasons to value (numerical) data summaries.

- They might be important per se.
- They may give insight into aspects of data, that can be relevant for subsequent analysis.
- They may be used as data for further analysis, though this requires some caution to avoid information loss.

Appropriate data summaries depend on the nature of the data at hand and the focus here is on:

- summary statistics used to describe the distribution of univariate (numerical) variables with regard to location, variability, symmetry and kurtosis;
- summaries for multivariate and, in particular, for bivariate numerical data sets;
- summaries for counts data (categorical variables).

1.9.1 Some basic summaries for univariate data: measures of location

The most common measure of location is the (sample) **mean** $\bar{y}$, which can be computed only for numerical variables.

Let $y = \{y_1, \ldots, y_n\}$ represent the (sample) values of a variable Y

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

It can be influenced by outlying observations.

Another measure of position is the (sample) **median** $y_{0.5}$, which can be computed for both numerical and ordered categorical data:

$$y_{0.5} = \begin{cases} \text{the middle obs.} & \text{if } n \text{ is odd} \\ \text{the average of the two middle obs.} & \text{if } n \text{ is even} \end{cases}$$

where the sample is sorted in ascending order. It is less influenced by outliers, then it is more robust than $\bar{y}$.

A further simple measure of location is the **mode**, which can be computed for all kind of variables. It is the value associated with the greatest frequency.

The notion of (sample) **quantile** generalizes the notion of median.

The α-quantile y_α, with $\alpha \in (0, 1)$, is the value which splits the frequency distribution into two parts, corresponding (approximately) to $\alpha 100\%$ observations (on the left) and $(1 - \alpha)100\%$ observations (on the right).

The three quartiles, $y_{0.25}, y_{0.5}, y_{0.75}$, divide the distribution into four equal parts; similarly, the percentiles divide the distribution into one hundred equal parts.

1.9.2 Some basic summaries for univariate data: measures of variability

The **range** $R = \max(y) - \min(y)$ and the **interquartile range** $IQR = y_{0.75} - y_{0.25}$ are simple measures for numerical data. The most commonly used measure of variation for numerical data is the (sample) **variance**:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where the denominator $n-1$ is the number of degrees of freedom remaining after estimating the mean with $\bar{y}$.

The positive square root s is called (sample) **standard deviation**.

The median of the transformed data set $|y - y_{0.5}|$ is called **median absolute deviation (MAD)** and it is the robust counterpart of s.

The **coefficient of variation** $CV = s/|\bar{y}|$ is a standardized measure of variability, useful for data comparison.

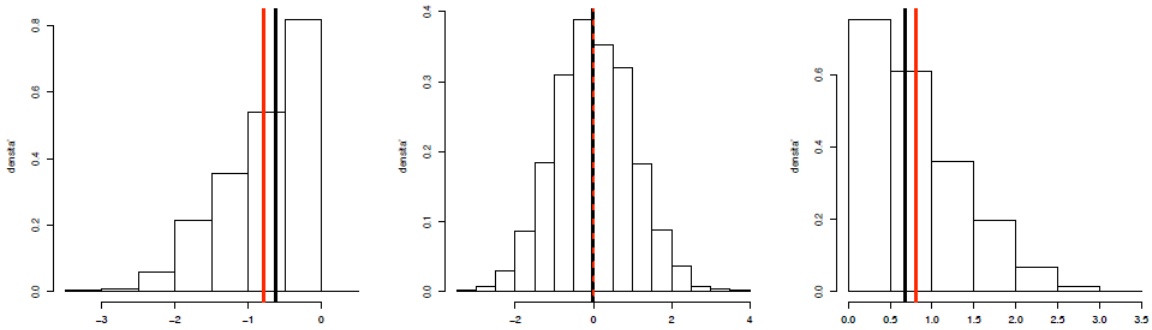
For categorical data: indices of heterogeneity and of concentration.

1.9.3 Some basic summaries for univariate data: measures of asymmetry and kurtosis

Graphs (barplots, histograms, density estimates and boxplots) are useful for investigating the shape of the frequency distribution. A preliminary indication of the **asymmetry (skewness)** of a distribution may be obtained by comparing the mean and the median:

- negative skew: $\bar{y} < y_{0.5}$
- symmetric distribution: $\bar{y} = y_{0.5}$
- positive skew: $\bar{y} > y_{0.5}$

Shown in order:



The most common **index of skewness** is:

$$\gamma = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^3}{s^3}$$

If the distribution is symmetric, $\gamma \approx 0$; if it is skewed to the left, $\gamma < 0$; if it is skewed to the right, $\gamma > 0$.

The analysis of the **kurtosis** concerns the shape of a frequency distribution, focusing on tail weight and peakedness. The kurtosis is evaluated with respect to the normal distribution (the "bell curve"), considered as the reference standard. The most common **index of kurtosis** (it measures tails heaviness) is:

$$\beta = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^4}{s^4}$$

If the distribution is normal shaped, $\beta \approx 3$; if it is hyponormal (thinner tails), $\beta < 3$; if it hypernormal (fatter tails), $\beta > 3$.

1.9.4 Some basic summaries for univariate data: correlation

The relationship between two numerical variables can be graphically represented by a scatterplot; in case of more variables: scatterplot matrix. A relevant measure of the linear relationship between two numerical variables is the (sample) **covariance** s_{xy}.

Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ represent the (sample) values of the variables X and Y, observed on the same units,

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The sign of the covariance shows the tendency, positive or negative, in the **linear relationship** between the variables; $s_{xy} \approx 0$ indicates absence of linear relationship.

The **Pearson correlation coefficient** is a standardized summary measure of linear relationship:

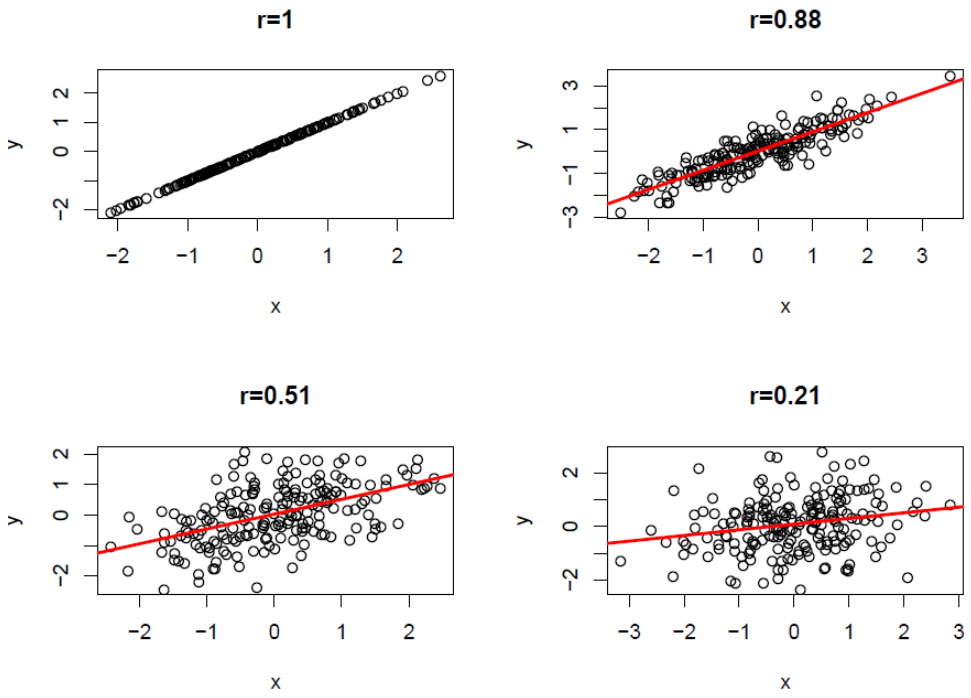
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

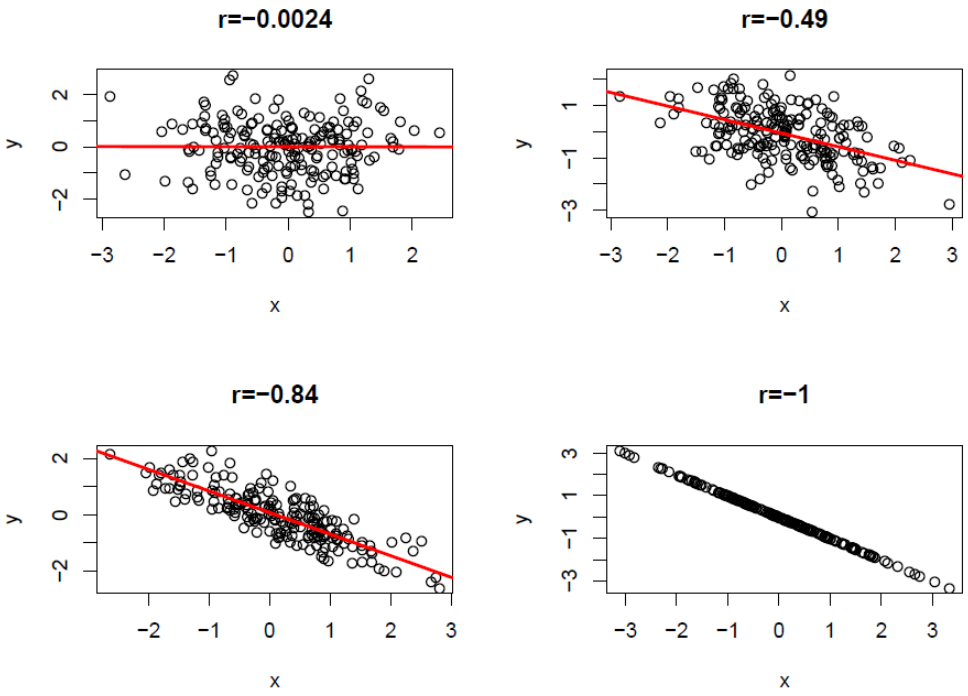
$r_{xy} \in [-1, 1]$ and when $r_{xy} \approx 1$ ($r_{xy} \approx -1$) the data points tend to lay exactly on a line with positive (negative) slope.

The calculation should go together with a scatterplot for checking linearity; a smooth trend is usually helpful. Another fact to check is whether the marginal distributions of the two variables are roughly normal, or at least not highly skewed.

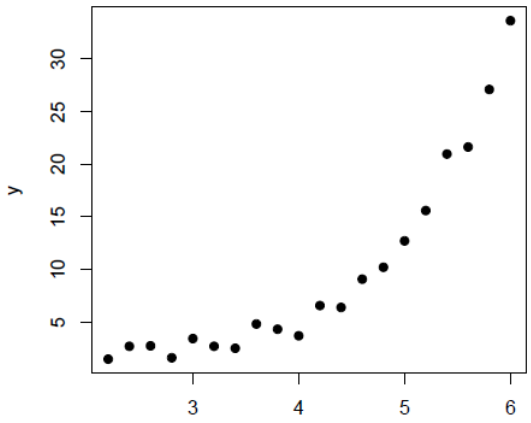
For monotonic nonlinear relationships or asymmetric marginal distributions the **Spearman rank correlation coefficient** should be used, namely $r_{x,y}$ computed using the ranks of x and y.

A further alternative is the **Kendall's tau correlation coefficient**, based on the number of concordant and discordant pairs.





The following scatterpolot shows a strong nonlinear pattern in which, with the above data set, the Pearson correlation coeffcient is 0:887, while the Spearman correlation coeffcient, which better captures the strength of the relationship, is 0:958.



The magnitude of r_{xy} does not of itself indicate whether the fit is adequate; here the linear fit is clearly inappropriate and the graphical representation may guide a suitable numerical analysis.

1.9.5 Summaries for count data

Count data are usually given in the form of contingency tables, collecting the observed frequencies of each combination of variable categories.

Summaries for count data require some care since information may be lost or obscured, for example when summarizing counts across the margins of multi-way tables.

A famous data frame contains the data from a study on unemployed individuals, focusing in particular on differences between those that followed a training program and those that did not.

The two groups, treated and untreated, are not genuinely comparable; by considering who had completed high school and who had not, it is clear that training group has a much higher proportion of dropouts.

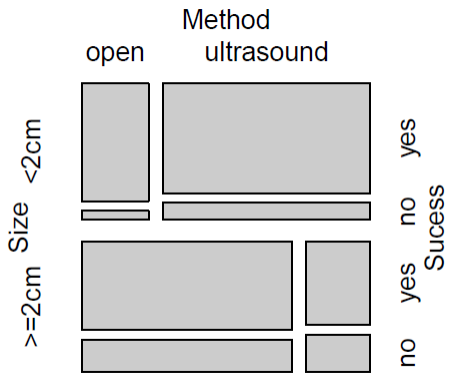
	high school	education	
treatment	completed	dropout	% completed
none	1730	760	68.5
training	80	217	26.9

1.9.6 Example: Kidney stones

Data are from a study that compares outcomes for two different methods of surgery for kidney stones: open, which used open surgery, and ultrasound, which used a small incision and ultrasound. Additional information on the size of the stones: < 2cm, ≥ 2cm.

		success		
method	size	yes	no	% yes
open	<2cm	81	6	93.1
	≥ 2cm	192	71	73.0
ultrasound	<2cm	234	36	86.7
	≥ 2cm	55	25	68.8

The success rate for each size of stone separately favors always open surgery. The multi-way table is summarized using the following mosaic plot.



Summarizing the counts, by summing with respect to the size, produces a loss of information leading to the apparent conclusion that the success rates favor ultrasound.

	success		
method	yes	no	% yes
open	273	77	78.0
ultrasound	289	61	82.6

1.10 Aims and strategies of statistical analysis

The data available are not suitable for any possible question; ideally, they should be collected (or, even, generated) after the aims of the analysis have been planned.

Usually, the aims include scientific understanding of some critical points (is a training program effective in reducing unemployment?) or prediction of some key variables (which is the price that house purchasers may be willing to pay for a certain area and house size?).

Statistical data analysis can greatly help in answering scientific questions, but does not stand alone, and it must be interpreted against a background of subject area knowledge.

A critical distinction is between what can be reached by an experiment and what can be reached by an observational study.

Well designed experiments give highly reliable results, whereas observational studies require a lot of care, and can even be misleading.

1.10.1 Observational versus experimental data: the job training program example

Two different strategies for evaluating a job training program:

- **1:** some enrolled subjects are randomly assigned to training and non-training groups, and after some time from the end of the course (for those under training) their job status is assessed;
- **2:** some subjects freely decide whether to attend a training course or not, and after some time from the end of the course (for those under training) their job status is assessed.

The problem with 2 is that trained subjects may differ systematically from the untrained ones for reasons totally unrelated to the training courses, such as motivation; this does not occur, due to randomization, in 1.

Yet, in practice, 2 is more common than 1 in economic studies. There are methods to analyse those kind of data, but they are more difficult than those for 1.

1.10.2 Statistical analysis strategies

A careful initial analysis of the data is essential in any statistical analysis, and should never been neglected.

EDA techniques are also important to assess the results of more formal analysis, when statistical models are employed.

For planning a formal analysis on experimental or observational data, all the available information should be considered. At times, pilot studies, limited in size, could be performed before designing a more extensive experiment.

EDA of the pilot study, or of data from previous studies, is then a crucial step.

2 A review of inference concepts: Statistical models

2.1 Summary and Introduction

Inferential statistics is about extracting information from data: specifically, information about the "system" that generated the data or about the population from which the sample data are obtained.

Most data contain a component of random variability: replications of the data gathering process several times would give somewhat different data on each occasion.

In many physical sciences, deterministic models are often adequate, as data variability may be small. Outside such cases, and nearly always in the social sciences, variability is a serious issue, and models have to incorporate it, leading to statistical models.

Statistical models involve families of probability distributions, with the aim of providing an adequate description of the data generating system or of the interest phenomenon.

If the model elements (for example, models parameters) were known then an adequate model could generate data that resembled the observed data, including reproducing its variability under replication.

The purpose of statistical inference is to use the statistical model to go in the reverse direction: to infer the values of the model unknowns that are consistent with observed data.

Statistical models for some data are chosen based on previous experience with similar data, subject area knowledge and careful use of EDA findings.

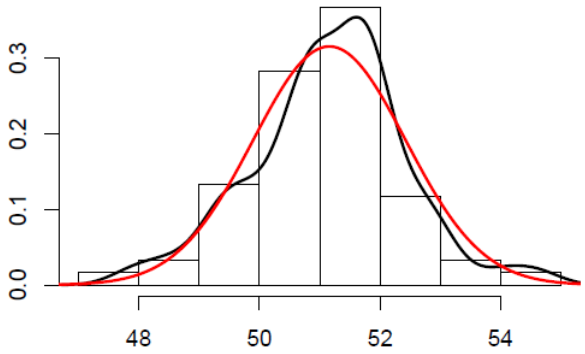
Statistical models often combine a deterministic component and a random component, that is an inherently unpredictable component.

The random component is often called noise or error (but there's typically nothing wrong with it), and sometimes the deterministic part is called signal.

2.1.1 Example: temperatures

Data set y with a 60 year record of mean annual temperatures (°F) in New Haven, Connecticut, from 1912 to 1971.

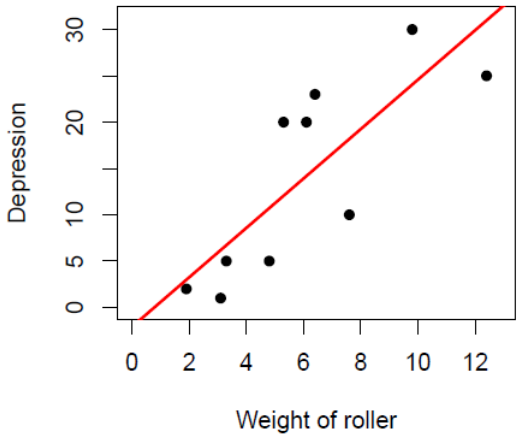
Numerical summaries: $\bar{y} = 51.16,\ y_{0.5} = 51.20,\ s^2 = 1.60,\ \gamma = -0.07\ \beta = 3.38$
(in order: sample mean, median, variance, skewness index, kurtosis index).



Graphical and numerical summaries suggest a simple model which considers y as independent observations from a **normal distribution**, although the tails seem heavier than those of the "bell curve".

2.1.2 Example: roller data

Data from an experiment where different weights (t) of roller were rolled over different part of a lawn, and the depression (mm) measured. Graphical summary: scatterplot of the data, with the **least squares line**.



The assumed model has a linear form for the deterministic part and an additive error term:

$$\text{depression} = \alpha + \beta \cdot \text{weight} + \varepsilon$$

It is called **linear regression model**. Here α and β are model parameters, namely constants which must be estimated using the data. Subscripts allow identification of the individual points: given observations $(x_1, y_1), \ldots, (x_n, y_n)$

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Using the least squares method, parameter estimates are $\widehat{\alpha} = -2.087$, $\widehat{\beta} = 2.667$ and the fitted values are defined by:

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i, \quad i = 1, \ldots, n$$

whereas the observed residuals are given by:

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i, \quad i = 1, \ldots, n$$

The focus of interest can be cast in terms of **interpretation of model parameters or prediction**. A crucial parameter is β, namely the rate of increase of depression with increasing roller weight. Predictions are given by fitted values $\widehat{y}_i$.

One may also predict the depression corresponding to out-of-sample roller weights, though some care would be required.

The model treats the pattern of change of depression with roller weight as a deterministic or fixed effect term. The measured values of depression incorporate also a random term that reflects:

- variation from a part of the lawn to another;
- differences in handling the roller;
- measurement error.

It is assumed that its elements are uncorrelated: size and sign of one element does not provide any information on the other elements. Data from multiple lawns are essential if one wants to generalize results to other lawns.

2.2 Least squares line

It is useful to remember that the **least squares line** has coefficients $\widehat{\alpha}, \widehat{\beta}$ that minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right)^2$$

It takes some simple linear algebra to show that the two coefficients solve a simple linear system giving:

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}, \quad \widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

At times, weighted least squares are useful. Some fixed weights w_i are introduced, and the quantity to be minimized is then:

$$\sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right)^2 w_i$$

2.3 Random variables

The concepts of randomness and probability are central to Statistics and the view of data as coming from a probability distribution is fundamental to understanding statistical methods.

Random variables are building blocks for statistical models, and in particular of their random component.

A random variable takes a different (numerical) value, at random, each time it is observed. It is possible to make probability statements about the values likely to occur, that is to specify its probability distribution. The distribution function of a random variable X is the function $F(x)$ such that:

$$F(x) = P(X \leq x)$$

It gives the probability that the value of X will be less than or equal to $x \in \mathbb{R}$.

From $F(x)$ it is possible to define the potential values for X, which belong to the support $\mathcal{S}$ of X, and the probability of events related to X, such as $X = a, X > a, a < X \leq b$, with $a < b \in \mathbb{R}$.

2.3.1 Discrete and continuous random variables

Discrete random variables take a discrete set of values (finite or countable) and they are suitable for finite or count data. They are described by the **probability (mass) function**:

$$f(x) = P(X = x)$$

Clearly, $f(x) \in [0, 1]$ and, for the potential values of X, that is x_i, $i \in I \subseteq \mathbf{N}$, $f(x_i) > 0$ and $\sum_{i \in I} f(x_i) = 1$.

Continuous random variables take values in a continuous set and the probability of taking any particular value is zero. They are described by the **(probability) density function** $f(x)$ such that:

$$P(a \leq X \leq b) = \int_a^b f(x)dx, a < b \in \mathbf{R}$$

Clearly, $f(x) \geq 0, \int_{-\infty}^{+\infty} f(x)dx = 1, \int_{-\infty}^{b} f(x)dx = F(b)$, so that: $F'(x) = f(x)$, when the first derivative $F'(x)$ exists.

2.4 Mean, variance and quantiles

Instead of considering the distribution of a random variable X completely, for many purposes its first two moments suffice.

In particular, the **expected value (mean)** $\mu = E(X)$ of a discrete or continuous random variable X, given respectively by:

$$E(X) = \sum_{i \in I} x_i f(x_i), \quad E(X) = \int_{-\infty}^{+\infty} x f(x)dx$$

and the **variance** $\sigma^2 = V(X) = E[(X - \mu)^2]$, with σ the **associated standard error**; index of skewness and kurtosis may be defined similarly.

The α**-quantile** x_α of X, with $\alpha \in (0, 1)$, is a value that X will be less than or equal to, with probability α.

The **median** of X corresponds to $x_{0.5}$ whereas the quartiles and the percentiles are obtained with the corresponding choices for α.

The transformed random variable $Z = (X - \mu)/\sigma$ is called **standardized random variable** since $E(Z) = 0$ and $V(Z) = 1$.

2.5 Random vectors

Little can usually be learned, on the interest phenomenon, from single observations. Useful statistical analysis requires multiple observations, viewed as a realization of a **random vector (multivariate random variable)**.

A random vector $(X_1, \ldots, X_n)$ takes values $(x_1, \ldots, x_n) \in \mathbb{R}^n$ namely numerical n-dimensional vectors, according to a **joint probability distribution**.

The probability distribution is defined by the joint distribution function:

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

or, equivalently, by the **multivariate (joint) version of the density (probability) function** $f(x_1, \ldots, x_n)$.

Each marginal component $X_i, i = 1, \ldots, n$, corresponds to a random variable with marginal density (probability) function $f_i(x_i)$.

The two following situations greatly simplify statistical analysis:

- The component random variables $X_i, i = 1, \ldots, n$, are **independent** (the realization of one does not affect the probability distribution of the others) and then:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

- The component random variables $X_i, i = 1, \ldots, n$, are **independent and identically distributed (i.i.d.)**, so that each component follow the same distribution with density (probability) function $g(x)$ and:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i)$$

2.6 Bivariate random variables

The two-dimensional case suffices to illustrate most of the concepts required for higher dimensions. Let us consider a **continuous bivariate random variable** (X, Y) with density function $f(x, y)$; the results for the discrete case are simply obtained by substituting summation for integration.

The **marginal density** of X is:

$$f(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

and similarly for $f(y)$.

The **conditional density** of X given $Y = y$ is:

$$f(x \mid y) = \frac{f(x, y)}{f(y)}$$

assuming $f(y) > 0$, and similarly for $f(y \mid x)$.

Bayes theorem (which leads to a whole school of statistical methods) assuming $f(y) > 0$:

$$f(x \mid y) = \frac{f(x) f(y \mid x)}{f(y)}$$

The component random variables X and Y are **independent** if and only if $f(x, y) = f(x)f(y)$.

The **conditional expectation (mean)** of X given $Y = y$ is:

$$E(X \mid Y = y) = \int_{-\infty}^{+\infty} x f(x \mid y) dx$$

and similarly for $E(Y \mid X = x)$; analogous definition for the conditional variance of X given $Y = y$.

The **covariance** of (X, Y) is:

$$\mathrm{Cov}(X, Y) = \sigma_{XY} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \{x - E(X)\}\{y - E(Y)\} f(x, y) dx dy$$

If X and Y are are independent, then $\mathrm{Cov}(X, Y) = 0$; the reverse does not hold (a relevant exception concerns the multivariate normal distribution).

The **Pearson correlation coefficient** of (X, Y), useful for describing linear dependencies, is:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The first and second order moments of (X, Y) are summarized by the **mean vector** $\mu = (\mu_X, \mu_Y) = (E(X), E(Y))$ and the **variance-covariance matrix**:

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} V(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & V(Y) \end{pmatrix}$$

The Σ matrix is symmetric, since $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$, and positive semidefinite.

2.7 Statistics

A **(sample) statistic** is a function (summary) of a set of random variables and it is itself a random variable.

The probability distribution of a sample statistic is called **sampling distribution** and its form depends on the joint distribution of the initial random vector.

Given a random vector $(X_1, \ldots, X_n)$, well-known examples of statistics are the **sample mean** $\bar{X}$ and the (corrected) **sample variance** S^2 defined, respectively, as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

The uncorrected sample variance is obtained by substituting the degrees of freedom $n-1$ with n.

Further examples of statistics are the sample median, the sample quantiles, the sample MAD, the sample covariance and the sample correlation coefficient.

Some useful results are listed below:

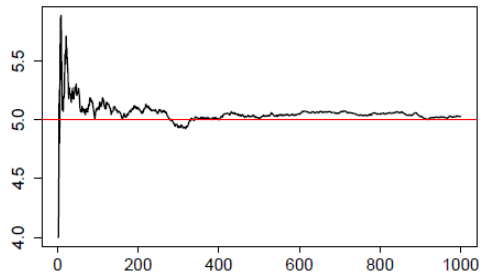
Whenever $X_1, \ldots, X_n$ are uncorrelated (independent) random variables, having the same marginal mean μ and variance σ^2 (e.g. this happens for identically distributed random variables), then:

- $E\left(\sum_{i=1}^{n} X_i\right) = n\mu, V\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2$
- $E(\bar{X}) = \mu, V(\bar{X}) = \sigma^2/n$
- $E\left(S^2\right) = \sigma^2$, while for the uncorrected version the expectation is $\sigma^2(n-1)/n < \sigma^2$

Weak law of large numbers: if $X_1, \ldots, X_n$ are i.i.d. random variables, $\bar{X}$ (sample mean) converges in probability to μ, as $n \to +\infty$ (in symbols, $\bar{X} \xrightarrow{p} \mu$) that is, as n increases, the distribution of $\bar{X}$ is more and more concentrated around the marginal mean μ.

A similar result holds for the (corrected and the uncorrected) sample variance: $S^2 \xrightarrow{p} \sigma^2$, as $n \to +\infty$.

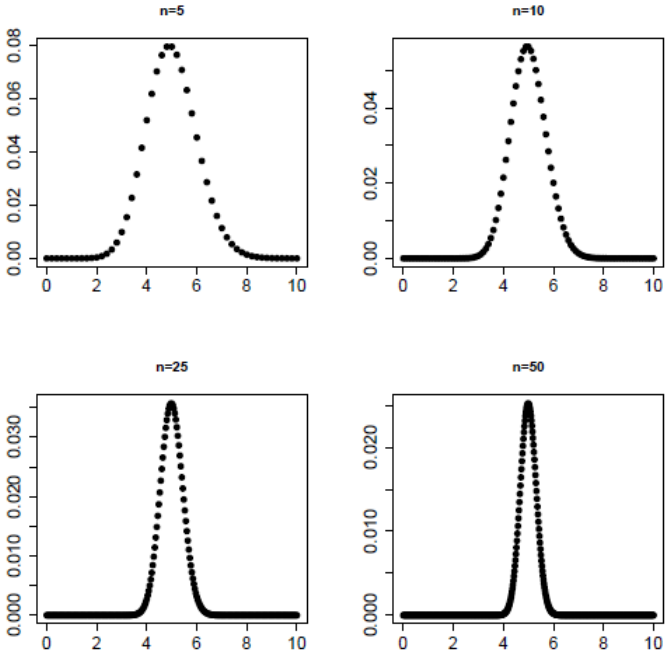
As a simple application of the weak law of large numbers, let $X_1, \ldots, X_n$ be i.i.d. Po(λ) distributed random variables with $\lambda = 5$. A sequence of observed values for the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$, with $n = 1, \ldots, 1000$, is given below:



The sample path shows that, as n increases, the observed values of $\bar{X}$ tend to be more concentrated around $\mu = \lambda = 5$.

Indeed, since the sample sum $\sum_{i=1}^{n} X_i$ follows a $Po(n\lambda)$ distribution, it is easy to specify the distribution of the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$.

The following figure describes the probability function of $\bar{X}$ for $n = 5, 10, 25, 50$:



The mean value is always $\mu = 5$, while the variability lessens as n increases.

2.8 Basic statistical models

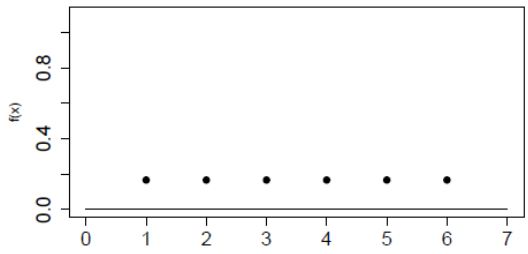
2.8.1 Discrete uniform distribution

A discrete uniform distribution describes an experiment where a finite number of values are equally likely to be observed.

A discrete random variable X follows a **discrete uniform distribution** with values $x_1, \ldots, x_n \in \mathbb{R}, n \in \mathbb{N}^+$, abbreviated as $X \sim Ud(x_1, \ldots, x_n)$, if $\mathcal{S} = \{x_1, \ldots, x_n\}$ and:

$$f(x_1) = \cdots = f(x_n) = \frac{1}{n}$$

Indeed, $E(X) = \sum_{i=1}^{n} x_i/n, V(X) = \sum_{i=1}^{n} \{x_i - E(X)\}^2/n$. The probability function $f(x)$, for $n = 6$ and $x_i = i, i = 1, \ldots, 6$, is:

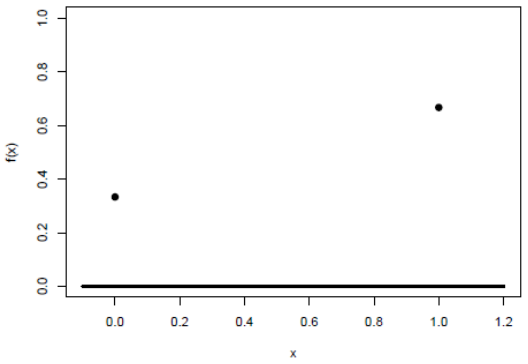


2.8.2 Bernoulli distribution

A discrete random variable X follows a **Bernoulli distribution** with parameter $p \in (0, 1)$, abbreviated as $X \sim \text{Ber}(p)$, if it describes an experiment where the possible outcomes are "success" (or 1) and "failure" (or 0); success may occur with probability p.

$\mathcal{S} = \{0, 1\}$ and $f(1) = P(X = 1) = p, f(0) = P(X = 0) = 1 - p$ indeed, $E(X) = p$ and $V(X) = p(1 - p)$.

The probability function $f(x)$ for $p = 1/3$ is described below:



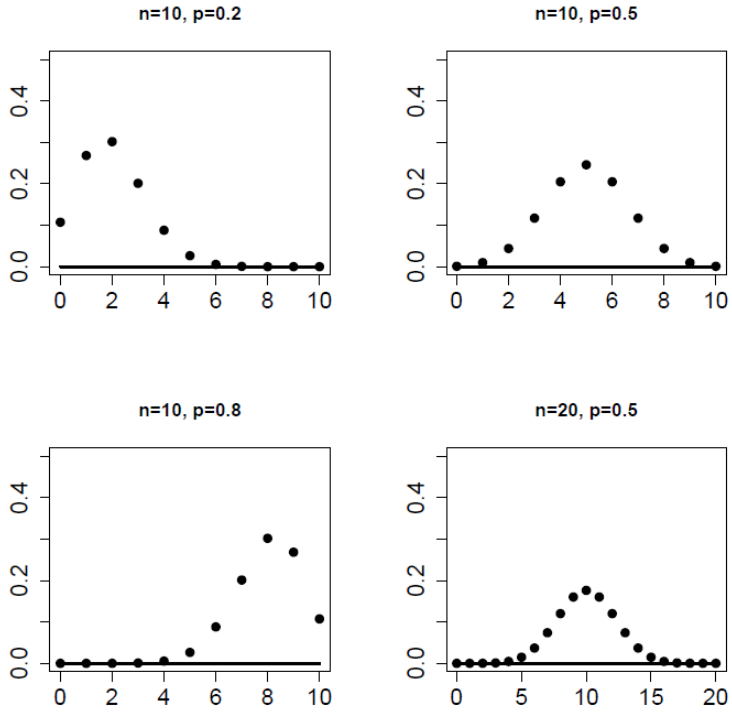
2.8.3 Binomial distribution

A discrete random variable X follows a **binomial distribution** with parameters $n \in \mathbb{N}, p \in (0,1)$, abbreviated as $X \sim Bi(n,p)$, if it describes the number of successes in n independent Bernoulli experiments with the same success probability p. $\mathcal{S} = \{0, \ldots, n\}$ and:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

X may be viewed as the sum of n of independent Ber(p) random variables; note that $Bi(1,p)$ corresponds to Ber(p). Indeed, $E(X) = np$ and $V(X) = np(1-p)$. It is easy to see that the proportion of successes X/n is such that $E(X/n) = p$ and $V(X/n) = p(1-p)/n$.

Probability function $f(x)$ for different n and p values:



2.8.4 Poisson distribution

The Poisson distribution is often used to model the number of events that occur in a certain time interval or in a prescribed spatial region, e.g. numbers of defects observed in manufactured products, number of visits to a website by an individual user.

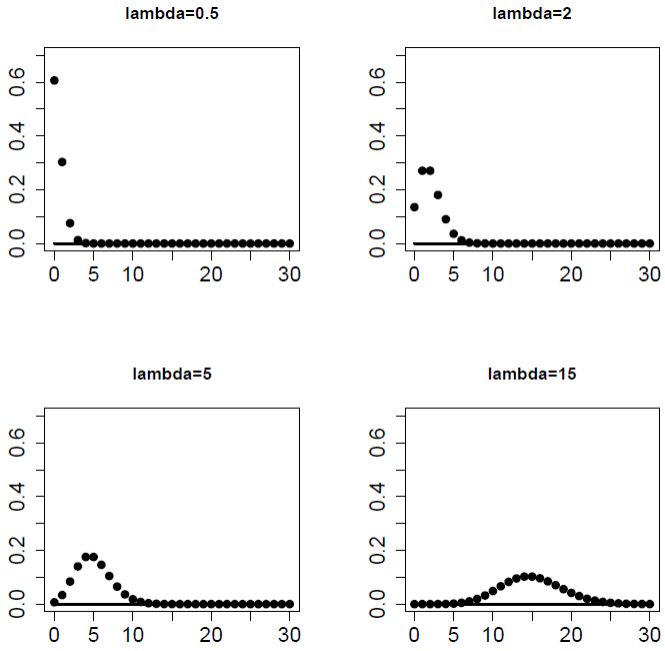
A discrete random variable X follows a **Poisson distribution** with parameter $\lambda > 0$, abbreviated as $X \sim Po(\lambda)$, if $\mathcal{S} = \mathbb{N}$ and:

$$f(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

Indeed, $E(X) = \lambda$ and $V(X) = \lambda$, thus the mean and the variance are both equal to the parameter λ.

Moreover, the probability distribution of a $Bi(n, p)$ random variable, with $n \geq 50$ and $p \leq 1/25$, is close to that of a Poisson distributed random variable with $\lambda = np$.

Probability function $f(x)$ for different λ values:



2.8.5 Geometric distribution

The geometric distribution is similar to the binomial distribution but it records the number of trials for the first success in a sequence of independent Bernoulli experiments with the same success probability p.

A discrete random variable X follows a **geometric distribution** with parameter $p \in (0, 1)$, abbreviated as $X \sim Ge(p)$, if $\mathcal{S} = \mathbb{N}^+$ and:

$$f(x) = \begin{cases} (1-p)^{x-1}p & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

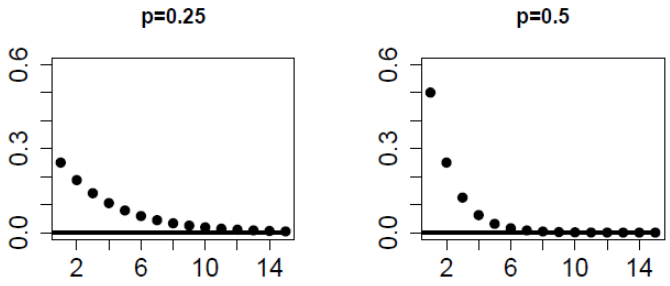
Indeed, $E(X) = 1/p$ and $V(X) = (1-p)/p^2$.

The geometric distribution is **memoryless**; this means that, given that the first success has not yet occurred, the conditional probability distribution of the number of additional trials does not depend on how many failures have been observed:

$$P(X > s + t \mid X > s) = P(X > t) \quad s, t \in S$$

The **negative binomial distribution generalizes the geometric one** by considering the number of trials until the r-th success, with $r \geq 1$, in a sequence of independent Bernoulli experiments with the same success probability p.

The case with $r = 1$ define the geometric distribution. The geometric probability functions $f(x)$ for $p = 0.25$ and $p = 0.5$ are described below:

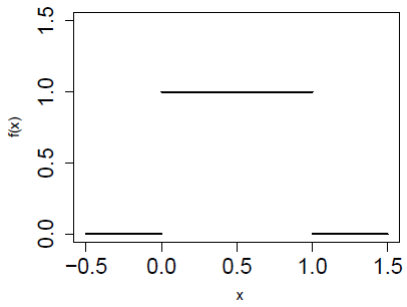


2.8.6 Continuous uniform distribution

The continuous uniform distribution describes equiprobability for continuous experiments, that is all intervals of the same length on the distribution's support are equally probable. A random variable X follows a **continuous uniform (rectangular) distribution** with parameter $a, b \in \mathbb{R}, a < b$, abbreviated as $X \sim U(a, b)$, if $\mathcal{S} = [a, b]$ and:

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Indeed, $E(X) = (a+b)/2, V\,(X = (b-a)^2/12)$ and, for $a = 0$ and $b = 1$, the density function is:



2.8.7 Exponential distribution

The exponential distribution is often used to describe durations, failure times or waiting times, assuming a constant hazard rate λ.

A continuous random variable X follows an **exponential distribution** with parameter $\lambda > 0$, abbreviated as $X \sim Esp(\lambda)$, if $\mathcal{S} = [0, +\infty)$ and:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

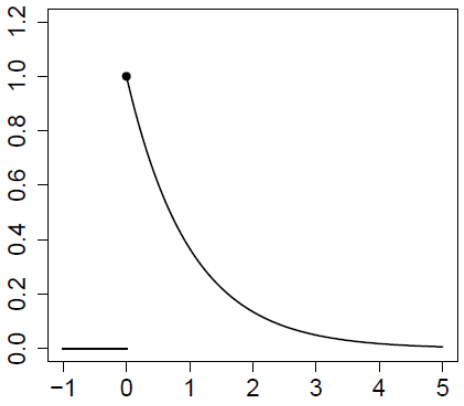
Indeed, $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$.

It can be viewed as a particular case of both the **gamma distribution** and the **Weibull distribution**. It also describes the (independent) times between two subsequent events in a **Poisson process**, which is a particular count process in which the interest events occur continuously and independently at a constant average rate λ.

The exponential distribution is the continuous analogue of the geometric distribution, having the property of being memoryless; namely,

$$P(X > s + t \mid X > s) = P(X > t), s, t \in S$$

The probability of failure in a given time interval is independent of the previous history. The density function $f(x)$ for $\lambda = 1$ is:



2.8.8 Normal distribution

The **normal or Gaussian distribution** has a central place in Probability and Statistics, largely as a result of the central limit theorem.

It has a "bell-shaped" density curve and it is often used as a model for a number of continuous measurement data (sometimes after a suitable transformation).

A continuous random variable X follows a **normal (Gaussian) distribution** with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, abbreviated as $X \sim N(\mu, \sigma^2)$, if $\mathcal{S} = \mathbb{R}$ and:

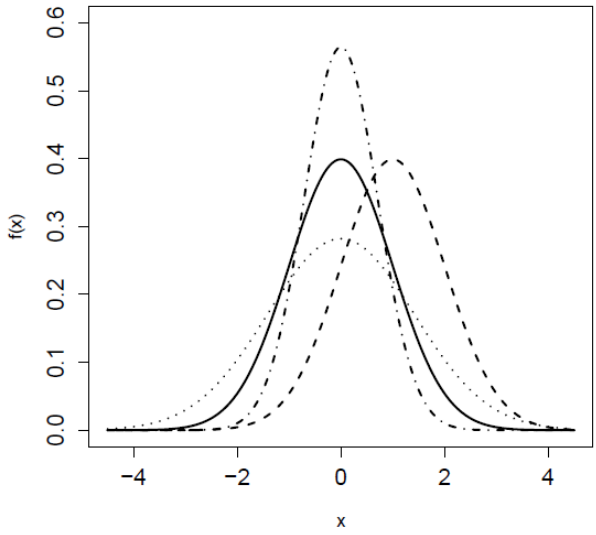
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, x \in \mathbb{R}$$

Indeed, $E(X) = x_{0.5} = \mu$ and $V(X) = \sigma^2$; the normal distribution is closed with respect to linear transformations, namely, if $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$.

A normal distributed random variable Z having mean 0 and variance (standard deviation) 1 is referred to as the **standard normal random variable**; note that $Z = (X - \mu)/\sigma$

The distribution function of a normal distributed random variable is not explicitly known, however numerical approximation are readily available, giving also the associated :α-quantiles.

Density function $f(x)$ of $X \sim N(\mu, \sigma^2)$ with $\mu = 0, \sigma^2 = 1(-), \mu = 1,\ \sigma^2 = 1(--), \mu = 0, \sigma^2 = 2(\cdots)$ and $\mu = 0, \sigma^2 = 1/2(- \cdot -)$:

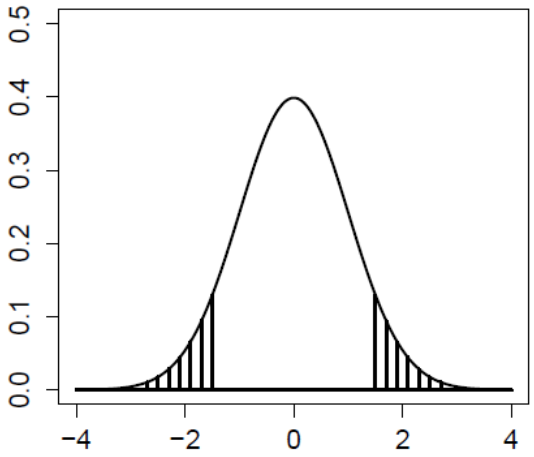


The standard normal density and distribution functions are usually indicated as $\phi(z)$ and $\Phi(z)$, respectively; indeed,

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

As a consequence of the symmetry of the normal density function, $\Phi(-z) = 1 - \Phi(z), z \geq 0$. Moreover, as described in the following figure, for $z \geq 0$:

$$P(|Z| < z) = \Phi(z) - \Phi(-z), \quad P(|Z| > z) = 2\{1 - \Phi(z)\}$$



With regard to statistical applications, the notion of critical value of a standard normal distribution may be useful.

The α-critical value of Z, with $\alpha \in (0, 0.5)$, is the value z_α such that:

$$P(Z > z_\alpha) = P(Z < -z_\alpha) = \alpha$$

z_α identifies the right α-level tail of $\phi(z)$, while $-z_\alpha$ defines the symmetric α-level tail on the left-hand side. In particular,

α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
z_α	1.28	1.65	1.96	2.33	2.58	3.09	3.29

As a straight forward application of the above mentioned results,

$$P(\mu - \sigma < X < \mu + \sigma) \doteq 0.68$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) \doteq 0.95$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) \doteq 0.997$$

2.8.9 The central limit theorem

Consider i.i.d. random variables $X_1, \ldots, X_n$, with mean μ and variance σ^2, then in the limit $n \to +\infty$:

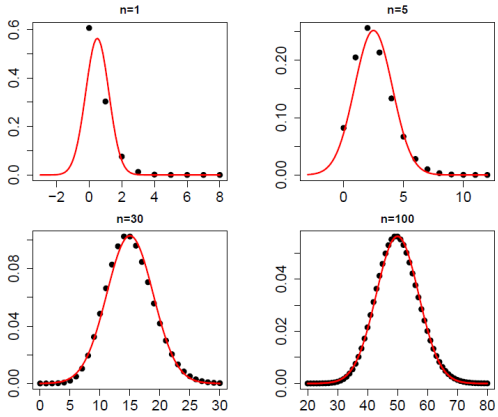
$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0, 1)$$

Thus, for a large n, the following approximations hold:

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right), \quad \sum_{i=1}^n X_i \sim N\left(n\mu, n\sigma^2\right)$$

As a simple application, let $X_1, \ldots, X_n$ be i.i.d. $\text{Po}(\lambda)$ distributed random variables; the sample sum $\sum_{i=1}^n X_i$ follows a $\text{Po}(n\lambda)$ distribution.

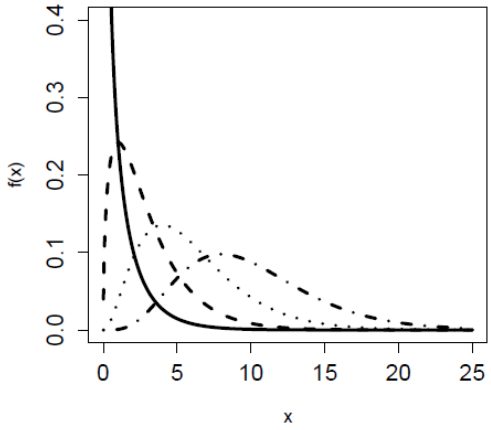
The following figure compares the true probability function of $\sum_{i=1}^n X_i$ with the density function of the **approximating normal distribution** $N(n\lambda, n\lambda)$, for $\lambda = 0.5$ and $n = 1, 5, 30, 100$.



2.8.10 Chi-squared distribution

Let $Z_1, \ldots, Z_k$ be i.i.d. standard normal distributed random variables; the random variable $Y = \sum_{i=1}^{k} Z_i^2$ follows a **chi-squared distribution** with $k \geq 1$ degrees of freedom, abbreviated as $\chi^2(k)$; it is a special case of the gamma random variable.

It is a continuous random variable with $\mathcal{S} = [0, +\infty), E(Y) = k, V(Y) = 2k$; the density function for $k = 1(-), k = 3(--), k = 6(\cdots), k = 10(- \cdot -)$ is given below:

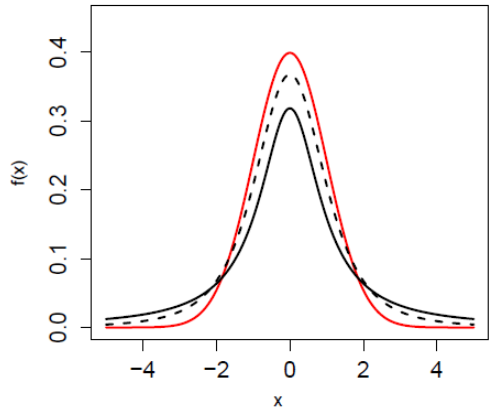


2.8.11 Student's t distribution

Let $Z \sim N(0,1)$ and $Y \sim \chi^2(k)$ be independent random variables; the random variable $T = Z/\sqrt{Y/k}$ follows a **Student's** t **distribution** with $k \geq 1$ degrees of freedom, abbreviated as $T \sim t(k)$.

It is a continuous random variable with $\mathcal{S} = \mathbb{R}, E(T) = 0$, for $k > 1$ and $V(T) = k/(k-2)$, for $k > 2$; if $k \to +\infty$, it converges to a $N(0,1)$ random variable.

The density is symmetric and "bell-shaped", like the normal density (in red), but it has heavier tails; some examples below for $k = 1(-), k = 3(--)$:

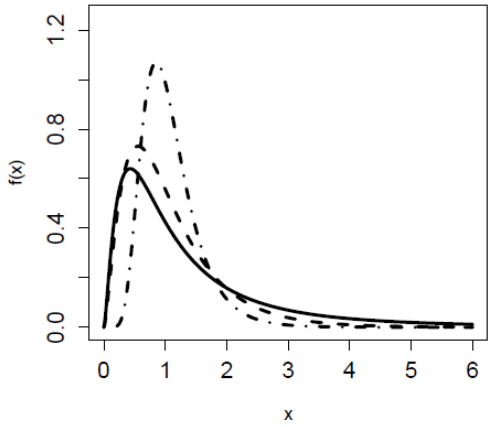


2.8.12 Fisher (F) distribution

Let $X \sim \chi^2(k)$ and $Y \sim \chi^2(m)$, with $k, m \geq 1$, be independent random variables; the random variable $F = (X/k)/(Y/m)$ follows an F-distribution with k and m degrees of freedom, abbreviated as $F \sim F(k, m)$.

It is also known as the **Fisher or the Snedecor distribution**; it is a continuous random variable with $\mathcal{S} = [0, +\infty)$ and $E(F) = m/(m-2)$, for $m > 2$.

Density function for $k = 5, m = 5(-), k = 5, m = 25(--), k = 25,\ m = 25(-\cdot-)$:



2.9 Sampling from probability distributions

Modern statistical software have routine to generate repeated random samples from a specified distribution. Such a task is referred to as **simulation**, and it plays a prominent role in modern statistical practice.

Summary statistics, possibly derived from a certain model, can be computed for each generated (simulated) sample, and their properties can be studied without intricate mathematics.

Simulation is widely used to determine the properties of statistical procedures in cases where it has not been possible or convenient to derive analytical results.

It has to be kept in mind, however, that computers generate pseudo-random numbers, typically based on deterministic algorithms, having specified a set of initial values.

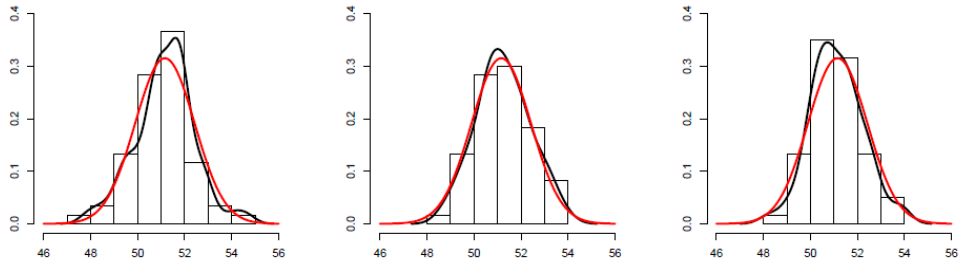
It is possible to specify the initial values by setting the random seed, thus forcing the generator to produce the same numbers.

2.9.1 Simulations from a normal distribution

Consider the data set y with 60 mean annual temperatures (°F) in New Haven, Connecticut, from 1912 to 1971.

Left panel: histogram and density estimate from the original data.

Central and right panels: histogram and density estimate based on simulated samples of dimension $n = 60$ from a normal distribution with $\mu = \bar{y} = 51.16$ and $\sigma^2 = s^2 = 1.60$

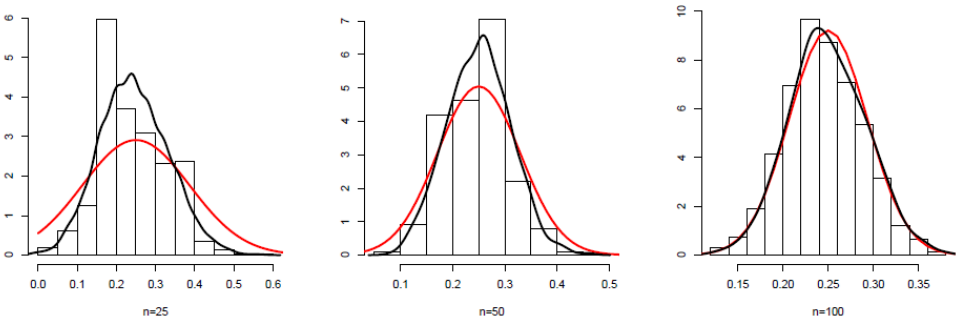


2.9.2 Simulations of the sample mean

The central limit theorem implies that, given an i.i.d. sample $X_1, \ldots, X_n$ for large enough n, the sampling distribution of $\bar{X} = \sum_{i=1}^{n} X_i/n$ will be closely approximated by a $N\left(\mu, \sigma^2/n\right)$ distribution.

1000 random samples of size $n = 25, 50, 100$ are simulated from a Ber(p) distribution with $p = 0.25$.

The resulting plots show the distribution of the sample mean estimated by simulation using the histogram and the density estimate, together with the approximating normal density.



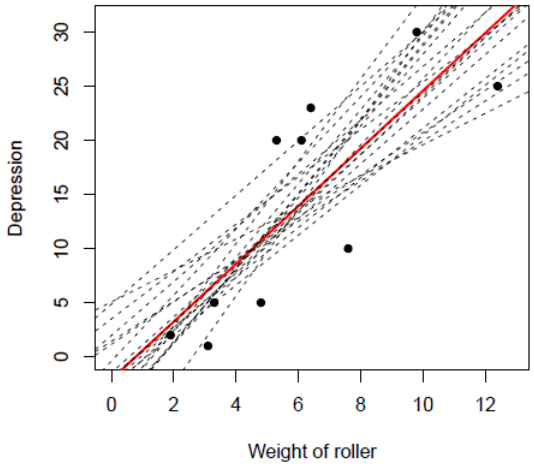
2.9.3 Simulations of regression data

It is possible to simulate a sample of n observations from the linear model

$$y_i = \alpha + \beta x_i + \varepsilon_i, i = 1, \ldots, n$$

with $\varepsilon_i \sim N\left(0, \sigma^2\right)$, independent of one another, and fixed x_i values.

Consider the roller data set. Least squares lines for each of the 20 simulated sample, with $n = 10, \alpha = -2.087, \beta = 2.667$ and $\sigma^2 = 45.367$, together with the original data and the original fitted line.



2.10 Sampling from a finite population

It is possible to generate a sample from a finite population, which amounts to sampling from a finite set of numbers.

Two variants are contemplated: sampling **without-replacement**, where no element is selected more than once, and sampling **with-replacement**, where repeated observations are allowed.

This kind of sampling is useful for practical implementation of **randomization** techniques, that are very important in experimental design and have also a role in statistical inference.

Cluster sampling is one of many different probability-based variants on simple random sampling: the clusters are independent, whereas the elements within the clusters are usually dependent.

2.11 Common model assumptions

Common assumptions for statistical models, having a deterministic and a random component, are **independence and normality** of the elements of the random terms and **homogeneity of variance** (that is, standard deviations of all measurements are the same).

When some assumptions do not hold, a statistical model may be invalid, failing to provide an adequate representation of the data. Some of the assumptions may be less important, and a certain method may be robust against them.

An important part of applied statistics is about which assumptions are important and need to be carefully checked.

Non-parametric methods have been developed to handle situations where normality or other assumptions are in question (without sensible alternatives); these methods assume little structure into a model and they are only sometimes useful.

Particular attention is dedicated to the independence and the normality assumptions.

2.11.1 Randomness

Typically, the data at hand are used as a window onto a much wider population, and they

should be collected in such a way that the randomness assumption is guaranteed.

For this reason, randomization in design experiments and random sampling in surveys are very important.

Samples chosen haphazardly or in a careless fashion (e.g. a survey interview involving individuals found in a shopping center) and self-selected samples can totally invalidate a statistical analysis.

Failure of the randomness assumption is a common reason for wrong statistical inferences, therefore it is crucial to identify the nature of any possible lack of randomness.

As a matter of fact, random sampling assumption is made even when data selection mechanism does not guarantee randomness; in such case, it is crucial to consider carefully how this lack of randomness will affect the data.

2.11.2 Indipendence

It is quite common to assume that the elements of a random sample are independent (and following the same distribution). However, suitable modifications of this simple independent random sampling scheme may be considered and therefore basic methods have to be modified or extended to handle such deviations from the basic experimental framework.

It may happen that the lack of independence is due to the fact that sampled units are close in time or space or belong to the same cluster, such as in the case of subjects from the same street or from the same household.

Whenever the randomness is guaranteed, if the data presents anyway temporal, spatial or cluster dependence, specific statistical models and methods have to be considered.

2.11.3 Checks for normality

Many data analysis methods are based on the assumption (at times implicit) that the data are normally distributed.

Real data are never exactly normally distributed, being the normal assumption at best approximate (usually after a suitable data transformation).

Large departures from normality is worrisome, whereas small departures are usually of no consequence.

Things to check are skewness, heavy or thin tails, outliers and undue discreteness (as that caused by excessive rounding).

With modest-sized samples, only gross departures will be detectable, and not even them for very small samples (with size 10 or less).

Graphical tools for checking for normality: histograms are usually not effective and a better tool for assessing normality is the normal probability plot (quantile-quantile plot). Formal statistical tests for normality may also be considered.

2.11.4 Quantile-quantile plot

For a normal probability plot, the data are sorted and then plotted against the ordered values to be expected if the data were from a normal distribution; namely, the observed quantiles are plotted against the theoretical normal quantiles.

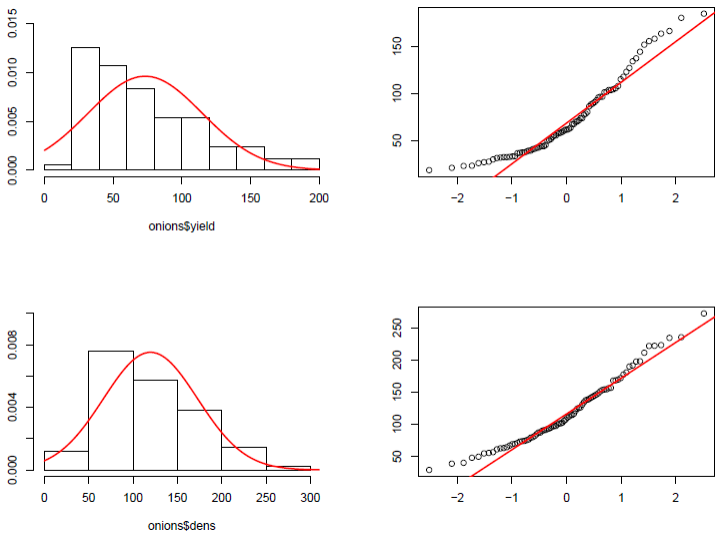
In case the data actually are from a normal distribution, with any mean and standard deviation, the plot should be approximately a straight line.

It is actually useful to add a line that passes through two given quantiles (such as the lst and 3 rd quartiles) to help the eye to assess the linearity.

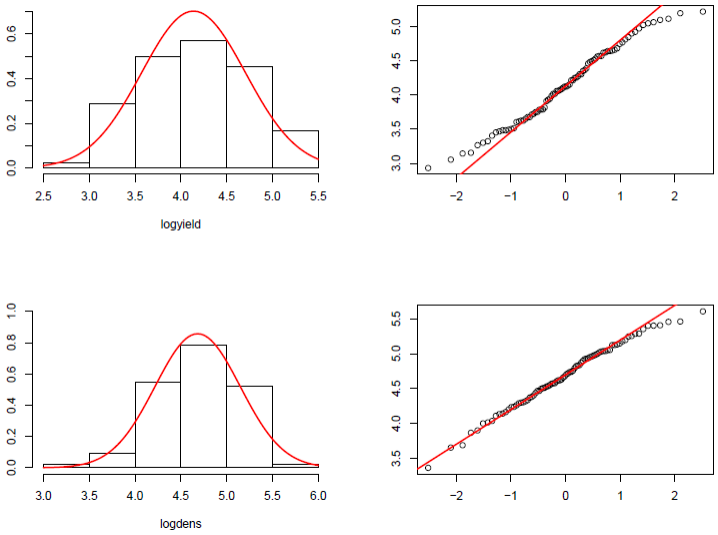
The same idea can actually be employed for any interest distribution, other than the normal one, by plotting the sorted data against the ordered values that might be expected from the relevant distribution.

2.11.5 Example: onions

Data from an experiment on the production of white spanish onions in two South Australian locations: $n = 84$ observations for dens, areal density of plants (plants per m^2) and yield, onion yield (gr per plant):



The original observations for dens and yield show departures from normality. The normal distribution assumption seems more plausible for the log-transformed data logdens and logyield.



2.11.6 Why models matter: the Simpson's paradox

Statistical models are important because they can consider all the relevant information simultaneously, in a way that simple summaries do not allow for.

Let us consider data on the admission frequencies by gender, for the six largest departments at the University of California at Berkeley in 1973: frequencies classified by admission status, gender and department.

The focus concerns evidence, across the University as a whole, of sex-based discrimination. Marginal admission rates for males and females:

	admission	status	
gender	admitted	rejected	% admitted
male	1198	1493	44.5
female	557	1278	30.4

Apparently, female were discriminated, and this went under the name of Berkeley gender bias case. A look at the results in each department show, however, that no single department was biased against women, as confirmed by the marginal admission rates for males and females for the six departments:

	dept					
gender	A	B	C	D	E	F
male	62.1	63.0	36.9	33.1	27.7	5.9
female	82.4	68.0	34.1	34.9	23.9	7.0

As a fraction of those who applied, females were strongly favored in department A, and males somewhat favored in departments C and E. The explanation of this paradox is in the different proportions of department applications for males and females, as described in the following table.

	dept					
gender	A	B	C	D	E	F
male	30.7	20.8	12.1	15.5	7.1	13.9
female	5.9	1.4	32.3	20.4	21.4	18.6

The overall bias arose because males favored departments where there were a relatively larger number of places, such as departments A and B.

This is just an instance of the **Simpson's paradox**, which refers to the fact that a relationship between two variables may change when the data are partitioned in subgroups, namely when another variable is taken into account.

Statistical models aim at considering all the relevant variables simultaneously and then they could be are the right tool to avoid such pitfalls, due to unsatisfactory and potentially misleading data summary.

3 A review of inference concepts: Statistical inference

3.1 Summary and Introduction

Making **inferences** about a population or about an interest phenomenon, based on a **random sample**, is a major task in statistical inference.

Methods for analyzing data characterized by an inherently random variability so that the conclusions drawn are generally valid, even though obtained from a single set of data.

For the most part this involves the use of **parametric statistical models**, which are suitable families of probability distributions, specified by one or more unknown parameters, describing hopefully how the data might have been generated.

The aim is to infer the values of the unknown model parameters that are consistent with observed data, and to **provide a measure of the accuracy of the inferential conclusion**.

The focus of interest may be the interpretation of the model parameters, and then of the interest phenomenon, or the prediction of future observations or outcomes using the estimated model.

The available data $y = (y_1, \ldots, y_n)$ are analyzed as **observations of a random vector** $Y = (Y_1, \ldots, Y_n)$, following an unknown joint probability distribution.

A **parametric statistical model** is a family of joint density (probability) functions $f(y_1, \ldots, y_n; \theta), \theta \in \Theta$, which hopefully contains the unknown generating probability distribution or a at least a suitable approximation of it.

The quantity θ, which specifies the density (probability) functions of the family, is a vector of **unknown parameters**; some of these parameters would answer the questions of interest about the system generating y.

Statistical model may also depend on some further data x that are usually treated as known and called **covariates or predictor variables**.

If the value of θ were known, a correct statistical model would allow the simulation of random data vectors which resemble the observed data y.

3.2 Random sampling

A random sample is a set of units selected from a larger population, and in a (uniform) random sample all the elements of the population have the same chance to be included in the sample.

A **random sample** describes also repeated observations of a random experiment or of a random phenomenon. Random sampling is the backbone of statistical inference, the collection of methods that allow to draw conclusions from a sample that are valid for the entire population or for the interest random phenomenon.

Since the observed sample data y can be always thought as generated by random vector Y, the analyst may assume some specific properties about the distribution of the marginal component random variables like **independence and identical distribution** (i.i.d.), **independence** but not identical distribution or some specific form of **dependence**.

3.2.1 Example: temperatures

Data set y with a 60 year record of mean annual temperatures (°F) in New Haven, Connecticut, from 1912 to 1971.

A simple model would treat the data as independent observations from normal distribution $N(\mu, \sigma^2)$, with unknown parameters $\theta = (\mu, \sigma^2)$. Then the density function of a single measurement $Y_i, i = 1, \ldots, 60$, is

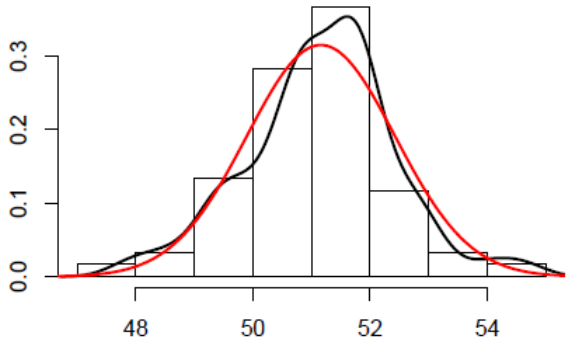
$$f(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\}$$

and the joint density of the vector data Y is

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{60} f(y_i; \mu, \sigma^2)$$

Numerical summaries $\bar{y} = 51.16, y_{0.5} = 51.20$ may provide a "guess" for μ, whereas $s^2 = 1.60$ a "guess" for σ^2 (sample mean, sample variance).

Estimates of the the generating probability distribution: histogram, smooth density estimate, normal density with $\mu = 51.16$ and $\sigma^2 = 1.60$



The tails seem heavier than those of the normal density; a better model might be to consider the data as independent observations from a Student's t distribution; namely, $Y_1, \ldots, Y_{60}$ i.i.d. random variable's such that:

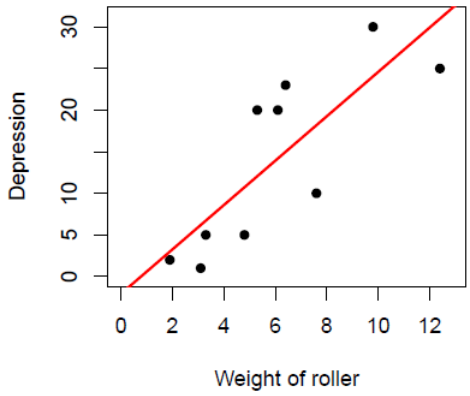
$$\frac{Y_i - \mu}{\sigma} \sim t(k)$$

with $\theta = (\mu, \sigma, k)$ unknown parameters.

3.2.2 Example: roller data

Experiment where different weights (t) of roller were rolled over different part of a lawn, and the depression (mm) measured.

Vector y includes data on the depression measured (response variable) and vector x includes the weights of the roller (covariate or predictor variable) which are taken as fixed.



Graphical summary (scatterplot of the data, with the least squares line) suggests that a useful model might be the linear regression model.

The **simple linear regression model** has a linear deterministic part and an additive error term so that:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

with the assumption that, given the covariate x_i, **the error (or noise) term is normally distributed**, $\varepsilon_i \sim N(0, \sigma^2)$, and errors of different units are independent.

This amounts to say that, given x_i (which is taken as fixed), the observed response y_i is viewed as a realization of the r.v. $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, independent from the other response random variables.

Here $\theta = (\alpha, \beta, \sigma^2)$ and the values $\widehat{\alpha} = -2.087, \widehat{\beta} = 2.667$, obtained with the least squares method, are a plausible "guess" for α and β.

Concerning the **interpretation of the model**, a crucial parameter is β, namely the rate of increase of depression with increasing roller weight.

With regard to **prediction**, using the estimated regression model, it is possible to predict the depression corresponding to out-of-sample roller weights, though some care is required.

3.3 Inferential questions

Given some data, y, and a statistical model with unknown parameters θ, there could be four basic points to consider:

- find values for θ which are most consistent with data y: **point estimation**;
- find ranges of values (usually, intervals) for θ which are consistent with data y : **interval estimation**;
- evaluate if some prespecified restriction on θ (hypothesis) is consistent with data y: **hypothesis testing**;
- evaluate if the model is consistent with the data y for any values of θ: **model selection/checking**.

There is a further point to be considered when the data-gathering process can be controlled; it concerns the organization of this process in order to take on the preceding question as accurately as possible: **experimental/survey design**.

There are two main classes of methods for answering these questions: **the frequentist and the Bayesian approaches**.

3.3.1 The frequentist approach

Basic inferential methods, like those taught in this notes, are usually frequentist. The model parameters θ are interpreted as fixed states of nature, about which the aim is to learn using the available data y. Probability is used to investigate what would happen to the inferential analysis under repeated replication of the data. Frequentist methods are usually based on the concept of **likelihood function**:

$$L(\theta; y) = f\left(y_1, \ldots, y_n; \theta\right), \theta \in \Theta$$

Function in the argument θ which gives the "probability" of observing the sample y (that was observed) by considering different values for θ. Given the observed data y, a suitable "guess" $\widehat{\theta}$ for the unknown parameters θ is the value which maximizes the likelihood function (or its logarithmic transformation called **loglikelihood function**):

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(\theta; y)$$

Likelihood-based procedures provide general solutions, with nice theoretical properties, to the above mentioned inferential problems. It is possible to derive **suitable (sample) statistics** (summaries of the random variables in the sample) for making inference on θ.

For example, if $y = (y_1, \ldots, y_n)$ are observations from i.i.d. $N\left(\mu, \sigma^2\right)$ random variables, the likelihood-based sample statistics for μ and σ^2 are the sample mean and the (uncorrected) sample variance:

$$\widehat{\mu} = \bar{Y}, \quad \widehat{\sigma}^2 = S^2(n-1)/n$$

A further rather general inferential approach is the **least square method**, which is particularly relevant for regression models.

A simple approach relies on the **method of moments** where, broadly speaking, the sample statistics are defined by considering the sample versions of the interest parameters, whenever available.

For example, the sample mean is considered for making inference on the population mean, the sample median for the population median, the (corrected) sample variance for the population variance, etc.

3.4 Point estimation: Estimators and standard errors

Sample statistics, used to estimate a certain parameter θ, are generally called **estimators**, and their value computed using the observed data y is just called the **point estimates** of θ and specified as $\widehat{\theta} = \widehat{\theta}(y)$.

Since an estimator is a summary of the random sample Y, it is itself a random variable (if θ is scalar), which is also denoted as $\widehat{\theta} = \widehat{\theta}(Y)$ (distinction will be clarified by the context).

Every estimator follows a sampling distribution, which describes the values assumed by it across (hypothetical) repeated random samples; if not explicitly known, the sampling distributions can be assessed by simulation.

Two theoretical properties are desirable for an estimator $\widehat{\theta}$:

- **unbiasedness:** $E(\widehat{\theta}) = \theta$ or at least $|E(\widehat{\theta}) - \theta|$ should be small;
- **low variance:** $V(\hat{\theta})$ should be small.

There is a trade-off between the two properties, so it is usual to seek both.

Under this respect, a suitable measure of the estimation error is the **mean square error (MSE)**:

$$\text{MSE} = E\left\{(\widehat{\theta} - \theta)^2\right\} = V(\widehat{\theta}) + |E(\widehat{\theta}) - \theta|^2$$

The square root of MSE yields the **standard error** $\text{SE} = \sqrt{\text{MSE}}$, which is a measure of the estimation accuracy having the same unit of measurement as the quantity being estimated.

A **parameter estimate** should always be accompanied by its **estimated standard error**, obtained by substituting θ with $\widehat{\theta}$ in SE.

If the estimator $\widehat{\theta}$ is unbiased, $\text{MSE} = V(\widehat{\theta})$ and then $\text{SE} = \sqrt{V(\widehat{\theta})}$ namely the standard deviation of the sampling distribution of $\widehat{\theta}$.

It is quite common to look for minimum variance unbiased estimators. A further relevant property is **consistency**: $\widehat{\theta} \xrightarrow{p} \theta$, as $n \to +\infty$.

Under suitable assumptions, **the maximum likelihood estimators (MLE's)** are (asymptotically) unbiased, consistent and achieves, in the large sample limit, a normal distribution with variance equal to the Cramér-Rao lower bound.

3.4.1 Estimation of the mean

The **sample mean** $\bar{Y}$ is a particular important instance of estimator, widely used to estimate the population mean μ.

It has some appealing properties, under i.i.d. assumptions, for any distribution of the variable of interest:

- unbiasedness: $E(\bar{Y}) = \mu$
- consistency: $\bar{Y} \xrightarrow{p} \mu$, as $n \to +\infty$

The central limit theorem ensures that the sampling distribution of $\bar{Y}$ **can be approximated by the normal distribution** $N(\mu, \sigma^2/n)$; in case of a sample from a normal distribution, $\bar{Y}$ is exactly normal.

The standard deviation of the sampling distribution of the sample mean is the **standard error of the mean (SEM)** $\sigma/\sqrt{n}$, whose estimate for random sampling is:

$$\text{SEM} = \frac{S}{\sqrt{n}}$$

where S^2 is a suitable estimator for σ^2, usually the (corrected) sample variance. So $S = \sqrt{S^2}$.

3.4.2 Estimation of the difference of means

With two independent i.i.d. samples of size n_X and n_Y, the comparison is usually in the form of a **sample difference** $\bar{X} - \bar{Y}$, where $\bar{X}$ and $\bar{Y}$ denote the respective sample means.

If the corresponding standard errors are SEM_X and SEM_Y, then the **standard error for the difference (SED)** is:

$$\mathrm{SED} = \sqrt{\mathrm{SEM}_X^2 + \mathrm{SEM}_Y^2}$$

In case it is reasonable to assume a common standard deviation σ for the two samples, it can be estimated by:

$$\mathrm{SED} = S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

where S_p^2 is an estimator for σ^2 based on both samples, usually the **pooled sample variance**:

$$S_p^2 = \frac{\sum \left(X_i - \bar{X}\right)^2 + \sum \left(Y_i - \bar{Y}\right)^2}{n_1 + n_2 - 2}$$

3.4.3 Example: elastic bands

Data from an experiment on the effect of heat on the amount of stretch (mm) of elastic bands: 21 bands were randomly divided into two groups, one of $n_X = 10$ and one of $n_Y = 11$.

Bands in the first group were tested for the amount that they stretched under a weight; the other group was placed in hot water for four minutes and then measured for amount of stretch under the same weight.

Two independent i.i.d. samples X and Y : $\bar{x} = 253.5, \bar{y} = 244.1$ $s_X = 9.92, s_Y = 11.73, \mathrm{SEM}_X = 3.14, \mathrm{SEM}_Y = 3.54$ since the separate standard deviations are similar, the pooled standard deviation estimate $s_p = 10.91$ is an acceptable summary of the variation in the data.

The mean difference is $\bar{x} - \bar{y} = 9.41$, with a SED = 4.77; therefore, the mean change is positive and it corresponds to $9.41/4.77 = 1.97$ times the estimated standard error. Saying that the difference is about 2 times the estimated standard error, means that the difference is strange (because 2 in a standard normal distribution is a rare case) so maybe there is a difference in elongation with or without the hot water process.

3.4.4 Estimation of a proportion

The aim is to estimate the probability of occurrence p of a "success" event in a sequence of n i.i.d. Ber(p) random variables $Y_1, \ldots, Y_n$; an alternative interpretation is in terms of a realization of a $Bi(n, p)$ random variable.

An unbiased, consistent estimator for p is the observed proportion of the event of interest in the n trials, which corresponds to the sample mean:

$$\widehat{p} = \bar{Y}$$

The associated standard error is:

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}$$

which is estimated by substituting p with $\widehat{p}$. For example, a random sample of $n = 132$ freshmen is selected in order to evaluate the proportion of freshmen that are displaced from their home. since 37 out of 132 freshmen are displaced: $\widehat{p} = 0.28, \mathrm{SE} = 0.039$.

3.4.5 Sampling distribution of z- and t- statistics

Consider an i.i.d. sample $Y = (Y_1, \ldots, Y_n)$ from a $N(\mu, \sigma^2)$ distribution, the **z-statistic (standardized sample mean)** is such that:

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

but this is not directly useful unless σ is known.

When the component random variables do not follows a normal distribution, the above result holds approximately for large n.

The **t-statistic (studentized sample mean)** is obtained by substituting σ with S and it is such that:

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Given the observed data y, formula $t = (\bar{y} - \mu)/$ SEM provides a standardized measure (in number of SEMs) of the distance between the sample mean and the true value μ.

3.4.6 Estimation of the variance

The **(corrected) sample variance** S^2 is widely used as estimator for the population variance σ^2. Under i.i.d. assumptions it is **unbiased**, $E(S^2) = \sigma^2$, and **consistent**:

$$S^2 \xrightarrow{p} \sigma^2, \text{ as } n \to +\infty$$

In the case of i.i.d. observations from a $N(\mu, \sigma^2)$ distribution, S^2 follows a scaled chi-squared distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

With two independent i.i.d. samples of size n_X and n_Y, the comparison concerning the distribution variances is usually in the form of a **sample variance ratio** S_X^2/S_Y^2, where S_X^2 and S_Y^2 are the respective corrected sample variances. In case of normal samples from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$:

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n_X - 1, n_Y - 1)$$

3.5 Confidence Intervals

Confidence intervals (interval estimates) provide more satisfactory estimation results than point estimates alone, giving an entire set of values (usually an interval) to estimate the population parameter.

Interval estimation gives also an implicit idea of the accuracy of the estimation procedure. A $(1 - \alpha)100\%$ confidence interval for a scalar parameter θ is an observation of a random

interval, based on a suitable sample statistic and designed to have a prescribed probability $1-\alpha$ **(confidence level)** of including the true value of θ.

The inferential procedure is expected to specify, over repeated samples, intervals that include the true parameter value with a certain proportion of the times, corresponding to the confidence level.

The confidence level specify a probability referred to the random interval and not to the observed confidence interval.

3.5.1 Confidence interval for the mean

Consider an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution with σ^2 known; by some algebra, using the z-statistics, it follows that:

$$P\left(\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

with $z_{\alpha/2}$ the $\alpha/2$-critical value of a $N(0,1)$ distribution. The random interval:

$$\left[\bar{Y} \pm z_{\alpha/2}\sigma/\sqrt{n}\right]$$

is the $(1-\alpha)100\%$-level **confidence interval** for μ; levels commonly used are $90\%, 95\%$ and 99%. The $(1-\alpha)100\%$-level **observed confidence interval**:

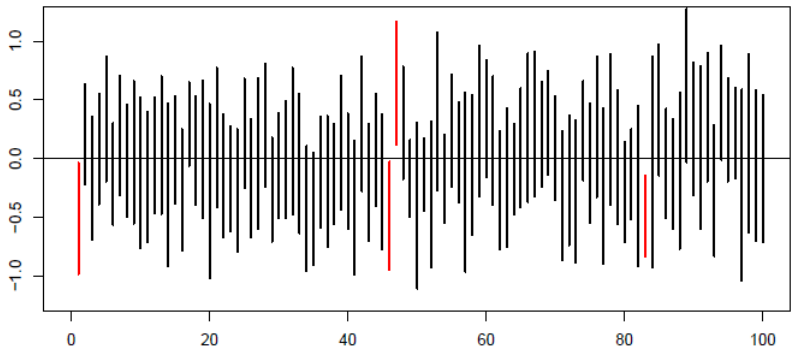
$$\left[\bar{y} \pm z_{\alpha/2}\sigma/\sqrt{n}\right]$$

summarizes the information provided by the observed data on the unknown μ.

When σ^2 is unknown, the $(1-\alpha)100\%$-level **confidence interval** for μ is based on the t-statistics and it corresponds to:

$$\left[\bar{Y} \pm t_{n-1;\alpha/2}S/\sqrt{n}\right]$$

with $t_{n-1;\alpha/2}$ the $\alpha/2$-critical value of a $t(n-1)$ distribution. 95%-level observed confidence intervals based on 100 simulated normal samples of size $n = 15$, with $\mu = 0$ and $\sigma^2 = 1$; four intervals fail to contain the true value:



The estimated (by simulations) confidence level is 0.96; increasing the number of simulated samples would get the result closer to 0.95.

3.5.2 Example: cork stoppers

Data set y with the total perimeter of the defects (in pixels) measured in $n = 50$ high quality cork stoppers. Numerical summaries: $\bar{y} = 365, y_{0.5} = 363, S^2 = 12167, S = 110$, $\text{SEM} = S/\sqrt{50} = 15.6$. Observations interpreted as i.i.d. realizations of a normal distribution. The 95% confidence interval for μ is:

$$[\bar{y} \pm t_{49;0.025}\text{SEM}] = [334, 396]$$

with $t_{49;0.025}$ the 0.025 -critical value of a $t(49)$ distribution. The observed confidence interval [334,396] is obtained using a statistical procedure which is characterized by a risk of 5% of giving wrong results (that is, intervals not containing the true value of μ).

3.5.3 Confidence intervals in general

In broad generality, $(1-\alpha)100\%$-level confidence intervals for a generic parameter θ have the form:

$$\left[\widehat{\theta} \pm z_{\alpha/2}\,\text{SE}(\widehat{\theta})\right]$$

where $\hat{\theta}$ is an estimator of $\theta, \text{SE}(\widehat{\theta})$ is its (estimated) standard error.

The above formula is valid if $\widehat{\theta}$ is normally distributed; whenever this holds only approximately for large n, the $(1-\alpha)100\%$ confidence level is approximate.

A relevant example is the confidence interval for a proportion p; indeed, the t-statistic-based confidence level for normal samples is also of the kind above, as $t_{n-1;\alpha/2} \approx z_{\alpha/2}$, for large n.

For specific models and parameters (variance, ratio of variances, correlation coefficient, etc), there may exist exact confidence intervals (exactness refer to the coverage probability) and whenever they exist they represent a better option than approximate intervals.

3.6 Hypotesis testing

Statistical procedures for hypothesis testing play a fundamental role in statistical inference and they are an essential item in many scientific studies.

The focus here is mainly on **parametric tests**, which rely on the specification of a parametric statistical model and aim at stating whether a prespecified restriction on the model parameter θ (i.e. a parametric hypothesis) is consistent with data y.

Nonparametric tests are sometimes called distribution-free tests because they are based on fewer assumptions and make no strict assumptions about the probability distributions of the random sample. In spite of these differences, the fundamental notions concerning hypothesis testing apply to parametric and to nonparametric tests alike.

3.6.1 Significance level and critical region

Hypothesis testing is a procedure for validating a **null hypothesis** H_0 made on possible values of θ; the null hypothesis is evaluated against an alternative hypothesis H_1 using the sample data y. A fundamental instance corresponds to:

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

where the alternative hypothesis is **two-sided**; H_1 could be also defined as **one-sided**: lower, $H_1 : \theta < \theta_0$, or greater $H_1 : \theta > \theta_0$.

A **test statistic** is a sample statistic, usually (a suitable transformation of) an estimator for the interest parameter θ, such that observed values far from θ_0 (greater or lower) may lead to rejection of the null hypothesis.

It is essential to derive, at last approximately, the **sampling distribution** of the test statistic **under the null hypothesis**; in standard cases this will be a well-known result, as for example, for the z-statistics or the t-statistics.

To perform a statistical test it is crucial to select a **significance level** α, common values are 5% and 1%, which is the accepted (by the analyst) probability of the **type I error** (reject H_0 when it is true).

The knowledge of the distribution of the test statistic under the null hypothesis enables a partition of its possible values into those for which the null hypothesis is rejected **(critical region)** and those for which it is not.

The critical region is derived so that its probability under H_0 is α, at least approximately.

Given the observations y, if the observed value of the test statistic is in the critical region, **the null hypothesis is rejected**, otherwise it is accepted or "not rejected".

Besides the probability α of the type I error, it is important to evaluate the probability β of the the **type II error** (accept H_0 when it is false); the value of β characterizes the power of statistical tests with a fixed significance level α.

3.6.2 P-value of a test

The ***p*-value** is the probability, under the null hypothesis (if it is true), of obtaining a test statistic value, across (hypothetical) repeated random samples, that is at least as extreme (against H_0) as that which was observed.

The computation of the p-value requires the knowledge of the distribution of the test statistics under H_0, at least approximately, and it depends on the form of the alternative hypothesis (two-sided or one-sided).

It is a measure of closeness of the data to the null hypothesis, and a small p-value (e.g. < 0.05 or < 0.01) is an evidence against H_0.

It is common to compare the p-value with the chosen significance level α and to reject H_0, in favor of H_1, if and only if its value is less than α.

The two testing approaches (critical region and p-value) are equivalent from the decision making perspective (given α, both lead to the same decision), however the second one is usually preferred since it provides a measure of the evidence against or in favor of H_0.

3.6.3 The ASA's statement on p-values

The American Statistical Association (ASA) published a clear guidance on the proper use and interpretation of the p-value and, more generally, on good statistical and scientific practice. Four principles (out of six) made by the ASA: p-values can indicate how incompatible the data are with a specified statistical model. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. Scientific conclusions and business or policy decisions should not be based only

on whether a p-value passes a specific threshold. Proper inference requires full reporting and transparency.

3.6.4 Testing the means

The purpose is to asses whether the mean of a population (from which the i.i.d. sample was collected) has a certain value or not:

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Here the alternative hypothesis H_1 is two-sided; H_1 could be also one-sided: lower, $H_1 : \mu < \mu_0$, or greater $H_1 : \mu > \mu_0$.

In **case of normal observations**, the test is called (one-sample) t test since it is based on the t-statistic, so that, under H_0,

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

Given a significance level α, the critical region for the t test is

$$R_\alpha = \left\{ y : |t| \geq t_{n-1;\alpha/2} \right\}$$

that is, H_0 is rejected if the difference, in absolute value, between $\bar{y}$ and μ_0 is at least equal to $t_{n-1;\alpha/2}$ times the SEM.

If H_1 is one-sided, the critical region is $R_\alpha = \{y : t \leq t_{n-1;\alpha}\}$, lower alternative, or $R_\alpha = \{y : t \geq t_{n-1;\alpha}\}$, greater alternative. Given the observed value t of the test statistic T, the p-value is

$$p = 2 \min \{P_{H_0}(T \leq t), P_{H_0}(T \geq t)\} = P_{H_0}\{|T| \geq |t|\}$$

whereas, for a one-sided lower alternative, $p = P_{H_0}(T \leq t)$ and, for a one-sided greater alternative, $p = P_{H_0}(T \geq t)$. When the population variance σ^2 is known (or the sample size n is sufficiently large), the z-statistic is considered and the test is called (one-sample) z test; then, under H_0,

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

In the **case of non-normal observations** and a large sample size: z test using the standard error evaluated under H_0. In the case of non-normal observations and a small sample size: ad hoc exact tests.

3.6.5 Example: maximum temperature

Data set y with maximum temperature (°C) registered in 1981 at $n = 25$ weather stations in Portugal. Numerical summaries: $\bar{y} = 39.8$, $y_{0.5} = 40$, $S = 2.739$, SEM $= 0.548$. Observations interpreted as i.i.d. realizations of a normal distribution. A "typical' year has an average maximum temperature of 37.5°C and then the aim here is to perform a t test with $\alpha = 0.05$ on:

$$H_0 : \mu = 37.5 \quad H_1 : \mu \neq 37.5$$

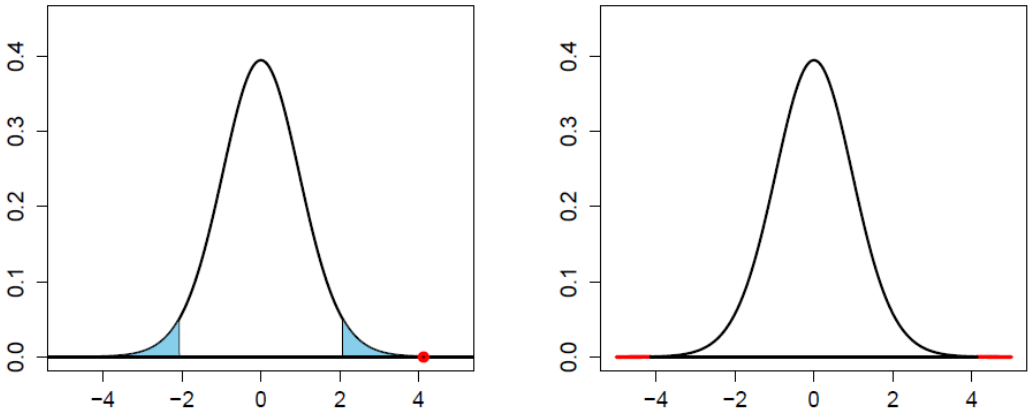
The reference sampling distribution is $t(24)$ so that $t_{24;0.025} = 2.064$ and $R_{0.05} = \{y : |t| \geq 2.064\}$ since the observed value of the test statistic is $t = 4.199$, the null hypothesis is rejected at the level $\alpha = 0.05$ of significance.

The p-value leads once again to the rejection of H_0:

$$p = P_{H_0}\{|T| \geq |4.199|\} = P_{H_0}\{T \leq -4.199 \text{ or } T \geq 4.199\} = 0.0003$$

The 95% confidence interval for μ is $[\bar{y} \pm t_{24;0.025}\text{SEM}] = [38.67, 40.93]$, which does not contain the null value $\mu_0 = 37.5$.

Left panel: observed value of the test statistic and critical region, giving an overall area equal to the level $\alpha = 0.05$ of significance. Right panel: p-value corresponding to the tails area.



3.6.6 Different ways to report results

Confidence intervals and, to a lesser extent, hypothesis testing provide a way to report summary results in a more interpretative way. For instance, for the maximum temperature data, the following statements may be considered.

- The mean maximum temperature is 39.8, with SEM = 0.548 and $n = 25$
- The observed value of the t-statistic is $t = 4.199$, on 24 d.f., namely the difference $\bar{y} - \mu_0$ is 4.199 times the standard error.
- A 95% confidence interval for the mean is [38.67,40.93].
- The null hypothesis, that the true mean maximum temperature is 37.5, is rejected ($p = 0.0003$)

Alternatives $3rd$ and $4th$ state differently and interpret the information in $1st$ and $2nd$.; alternative $3rd$ is probably the most informative.

3.6.7 Example: physical activity

The percent of adults (≥ 18 years old) in US who met the guidelines for aerobic physical activity in 2014 is 49.2% (National Health Interview Survey, US, 2014). A city's council wants to know if the proportion in their city is different from 49.2%: random sample bf $n = 200$ adults, 108 meet the guidelines. Observations interpreted as i.i.d. realizations of a Ber(p) distribution; the aim here is to perform a test with $\alpha = 0.05$ on:

$$H_0 : p = 0.492 \quad H_1 : p \neq 0.492$$

since p is the mean of a Bernoulli random variable and the sample size is large: z test statistic, which is approximately normal distributed.

$\widehat{p} = 0.54$, estimated SE $= \sqrt{\hat{p}(1-\widehat{p})/200} = 0.035$ and, considering the SE under H_0, $z = (\widehat{p} - 0.492)/\sqrt{0.492(1-0.492)/200} = 1.358$

The (approximate) p-value is $p = P_{H_0}\{|Z| \geq |1.358|\} = 0.175$; the null hypothesis is not rejected: there is not sufficient evidence to state that the proportion of citizens meeting the guidelines is different from 0.492.

3.6.8 Testing the means

For testing the equality of the mean of **two independent i.i.d. normal samples**, in case of (unknown) **equal variances**, a well-known test is the (two-sample) t test based on

$$T = \frac{\bar{X} - \bar{Y}}{\text{SED}}$$

where $\text{SED} = S_p\sqrt{n_X^{-1} + n_Y^{-1}}$, with S_p^2 the pooled estimate of σ^2.

The computation of the critical regions and of the p-values consider that, under H_0, the test statistic T follows a $t\,(n_X + n_Y - 2)$ distribution; there are general formulas for large samples, employing the normal distribution.

If variances are heterogeneous (i.e. **unequal variances**), the t-statistic based on the pooled variance estimate is inappropriate.

In this case, the **Welch test**, based on $T = (\bar{X} - \bar{Y})/\text{SED}$, with $\text{SED} = \sqrt{\text{SEM}_X^2 + \text{SEM}_Y^2}$, gives an adequate approximate solution; under H_0, T has a t distribution with suitable degrees of freedom.

Paired data arise when the same units are measured under two different conditions, generating two different observations x_i and y_i for each unit, $i = 1, \ldots, n$.

In this case, the methods for two independent samples cannot be used; a simple solution consists in using one-sample tests applied to the individual differences $d_i = x_i - y_i$ (e.g. $\boldsymbol{t}$ test for paired data).

In the case of two **non-normal independent samples** and large sample sizes: z test based on $\bar{X} - \bar{Y}$, using the SED evaluated under H_0.

In the case of two non-normal independent samples and a small sample size: ad hoc exact tests.

Specific tests may be considered for two **non-normal dependent samples**; a well-known large sample test for Bernoulli observations is the **McNemar's test**.

3.6.9 Example: white and red wines

Data x and y, with $n_X = 30$ and $n_Y = 37$, correspond to the aspartame content (mg/l) in two independent samples of white and red wines. Summaries: $\bar{x} - \bar{y} = 27.06 - 20.86 = 6.203, s_X = 10.51, s_Y = 10.97$ Observations interpreted as independent i.i.d. realizations of two normal distributions; the variances are considered as equal, although a formal statistical test would be required.

The point is whether the mean aspartame content can distinguish white wines from red wines: two-sample t test with $\alpha = 0.05$ on:

$$H_0 : \mu_X - \mu_Y = 0 \quad H_1 : \mu_X - \mu_Y \neq 0$$

since SED $= 2.645$, the observed value of the test statistic is $t = 2.345$ and the p-value is 0.022.

Note that for $\alpha = 0.05$ the null hypothesis is rejected, but not for lower values of the significance level, such as $\alpha = 0.01$.

3.6.10 Example: temperatures

Data x and y correspond to the maximum temperature (°C) registered in 1980 and in 1981 at $n = 25$ weather stations in Portugal. The objective is to compare the maximum temperatures in year 1980 with those of year 1981; since the measurements are performed in the same stations, the data sets x and y define pair data. Numerical summaries: $\bar{x} = 37.44, \bar{y} = 39.80, s_X = 2.20, s_Y = 2.739$; with regard to the differences $d = x - y, \bar{d} = -2.36$, SEM $_D = 0.412$. Observations interpreted as two dependent normal i.i.d. samples: t test for paired data on:

$$H_0 : \mu_X - \mu_Y = 0 \quad H_1 : \mu_X - \mu_Y \neq 0$$

The observed value of the test statistic is $t = -5.731$ and the p-value is $6.632 \cdot 10^{-6}$: the null hypothesis is rejected; strong evidence on the fact that the mean maximum temperature in 1981 exceeds that one in 1980.

3.6.11 Example: labor training program

Two groups of individuals are randomly selected: $n_X = 297$ who had participated in labor training programs and $n_Y = 128$ who had not. Evaluate if the proportion of high school dropouts is the same in the two reference populations: 217 and 65 dropouts observed, respectively.

Observations viewed as independent i.i.d. samples from a Ber (p_X) and a Ber (p_Y) distribution, respectively; the objective is to perform a test on

$$H_0 : p_X - p_Y = 0 \quad H_1 : p_X - p_Y \neq 0$$

thus, to assess whether the probability distribution of the dichotomous random variable school dropout is the same in the two populations. As the sample sizes are large: z test based on $Z = \left(\widehat{p}_X - \widehat{p}_Y\right) / \text{SED}$, where $\text{SED} = \sqrt{\hat{p}(1-\widehat{p})\left(n_X^{-1} + n_Y^{-1}\right)}$, with $\hat{p}$ is the estimated proportion based on the pooled sample. Since $\hat{p}_X = 0.731, \widehat{p}_Y = 0.508$ and $\hat{p} = 0.664$, then $z = 4.46$ and the (approximate) p-value is $8.189 \cdot 10^{-6}$, leading to H_0 rejection.

3.6.12 Testing the medians

Although the (one-sample) t tests are fairly robust against departures from the normal distribution, especially in larger samples, a nonparametric alternative is the **Wilcoxon signed-rank test**.

It assumes only that the distribution is continuous and symmetric and the hypotheses concern the median instead of the mean.

The Wilcoxon signed-rank can be considered as a nonparametric test for comparing the median of two paired samples, when the population cannot be assumed to be normally distributed.

It is the same as the previous test, applied on the differences of the paired observations. A nonparametric alternative to the (two-sample) t test for independent i.i.d. samples is the **Wilcoxon rank-sum test**.

It assumes only that the distributions are continuous and the hypotheses usually concern the difference in medians; it is equivalent to the **Mann-Whitney U test**.

3.6.13 Testing the variances

One normal i.i.d. sample: the aim is to asses whether the variance of a population has a certain value or not:

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 \neq \sigma_0^2$$

The test statistic is $(n-1)S^2/\sigma_0^2$, with null distribution $\chi^2(n-1)$.

Two independent normal i.i.d. samples: the purpose is to decide whether the two reference populations have the same variance:

$$H_0 : \sigma_X^2/\sigma_Y^2 = 1 \quad H_1 : \sigma_X^2/\sigma_Y^2 \neq 1$$

A well-known test is called F **test** for the homogeneity (equality) of variance (homoscedasticity) and it is based the ratio of the two (corrected) sample variances S_X^2/S_Y^2, which under H_0 follows an $F\left(n_X - 1, n_Y - 1\right)$ distribution, since $\sigma_X^2 = \sigma_Y^2$.

The F test is extremely sensitive to non-normality; the **Levene's test or the Bartlett's test** are more robust alternatives.

Example: red and white wines

In order to decide whether the mean aspartame content can distinguish white wines from red wines, a two-sample t test was considered, assuming equality of the population variances.

However, to evaluate the homogeneity of variance, an F test with $\alpha = 0.05$ can be performed by considering the hypotheses:

$$H_0 : \sigma_X^2/\sigma_Y^2 = 1 \quad H_1 : \sigma_X^2/\sigma_Y^2 \neq 1$$

since $s_X^2 = 110.43, s_Y^2 = 120.34$ and $n_X = 30, n_Y = 37$, the observed value of the test statistic is $s_X^2/s_Y^2 = 0.9178$ and, using the $F(29, 36)$ null distribution, the p-value corresponds to:

$$p = 2\min\left\{P_{H_0}(F \leq 0.9178), P_{H_0}(F \geq 0.9178)\right\} = 0.819$$

The hypothesis of homoscedasticity is accepted, having a strong support from the observed data.

3.6.14 Correlation Test

Given a random sample from the bivariate random variable (X, Y), it might be of some interest to test whether X and Y are correlated; namely:

$$H_0 : \rho_{XY} = 0 \quad H_1 : \rho_{XY} \neq 0$$

It should be emphasized that a correlation between two variables does not necessarily imply that one causes the other (causation).

A **correlation test** can be obtained under the assumption that the random sample derives from a bivariate normal distribution (it may be enough to check that both X and Y are normally distributed).

A well-known test is based on the sample version of the **Pearson correlation coefficient**:

$$r_{XY} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}$$

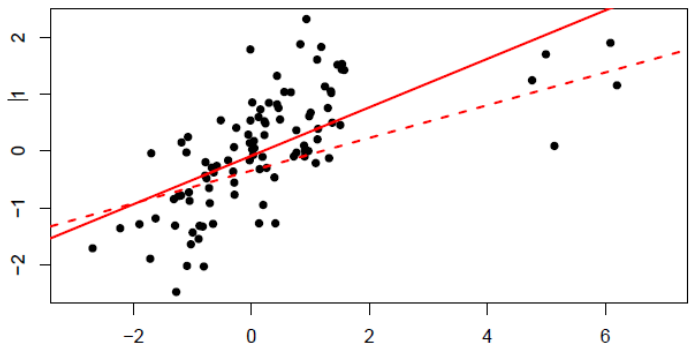
The test statistics is obtained by transforming r_{XY} so that, under H_0, it follows a suitable t distribution.

There are some nonparametric variants of the Pearson correlation test, having the advantage of not depending on the normal distribution; however, their interpretation may be not quite clear. A popular test is based on the **Spearman's rank correlation coefficient**, which is obtained by replacing the observations by their rank and computing the Pearson correlation coefficient.

A further nonparametric test involves the **Kendall's τ coefficient**, which is based on counting the number of concordant and discordant pairs (two pairs are concordant if the difference in the x-coordinate is of the same sign as the difference in the y-coordinate).

The null distribution of the above mentioned test statistics may be calculated exactly for small samples; for larger samples, it is common to use suitable approximations. The null distributions and the p-values, and even the confidence intervals for ρ_{XY}, can be computed using suitable simulation-based approximate methods.

The figure below shows that the five outlies in the upper right corner change the least square line:



The correlation tests based on the Pearson and on the Spearman correlation coefficient both suggest rejection of the null hypothesis; the rank correlation is more robust, though the p-values are always very small.

Full sample: sample correlation 0.64 (p-value $< 10^{-12}$), 95% confidence interval [0.51,0.75] rank correlation 0.75 (p-value $< 10^{-16}$).

Without outliers: sample correlation 0.73 (p-value $< 10^{-16}$), 95% confidence interval [0.61,0.81]; rank correlation 0.73 (p-value $< 10^{-16}$).

3.7 Basic concepts of model selection

3.7.1 Model Checking

The aim of model checking is to verify if the assumed statistical model is seriously wrong, taking into account that "all models are wrong but some are useful." (George Box)

"The type of statistical inference used may be less important to the conclusions than choosing a suitable model or models in the first place." (The GAMLSS team)

Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably." (OxfordDictionaries.com)

Although an **underfitted** model is clearly not adequate, an overfitted model contains too many parameters, since it tends to improperly model also the residual random variation observed in a data set. "Entities should not be multiplied beyond necessity". (Occam's Razor)

Although "all models are wrong, but some are useful", if a model is wrong statistically, then the statistical conclusions drawn from it are usually unreliable and questionable.

Graphical checks; when the model is wrong, they frequently indicate how it is wrong: quantile-quantile plot, plots of the standardized residuals (useful for regression-type models).

Parametric tests may be viewed as a procedure for comparing two alternative nested models: the null hypothesis defines a simplified (restricted) version of the statistical model taken into account.

The **goodness of fit tests** are statistical tests which considered as null hypothesis the compatibility of the observed sample with a parametric statistical model: **e.g. Kolmogorov-Smirnov test, chi-square goodness of fit test, Shapiro-Wilk test for normality**.

3.7.2 Akaike's information criterion

The Akaike's information criterion (AIC) enables model comparison, seeking the "best" model from a set of (two or more) models, which need not necessarily be nested.

A (theoretical) measure of goodness of fit for discriminating among alternative models is the **Kullback-Leibler divergence**, which, in the continuous case, is

$$K\left(\widehat{f}, f_T\right) = \int \left\{\log f_T(y) - \log f(y;\widehat{\theta})\right\} f_T(y)dy$$

where f_T is the true density of Y and $\widehat{f} = f(y;\widehat{\theta})$ is the estimated density under the assumed model.

According to the AIC, the selected model has the lowest value of:

$$\text{AIC} = -2\ell(\widehat{\theta}; y) + 2\dim(\theta)$$

with $\ell(\widehat{\theta}; y) = \log f(y;\widehat{\theta})$ the maximized loglikelihood based on data

The AIC is an estimate, based on the observed sample y, of the expected value of $K\left(\widehat{f}, f_T\right)$, with respect to the distribution of the estimator $\widehat{\theta}$; it requires the assumption that the model is the true one.

It specifies a trade-off between the goodness of fit of the model (measured by the maximized loglikelihood) and the complexity of the model (described by the dimension of θ).

An alternative criterion recognizes that $K\left(\widehat{f}, f_T\right)$ depends on the model only via $-\int \log f(y;\widehat{\theta})\Big\} f_T(y)d$ a cross-validation estimate for this quantity is

$$CV = -\sum_{i=1}^{n} \log f\left(y_i; \widehat{\theta}_{-i}\right)$$

where $\widehat{\theta}_{-i}$ is the MLE based on the data y with y_i omitted.

The **cross validation score criterion** points to the model which minimizes CV and measures the average ability to predict data to which it was not fitted.

3.7.3 Bayesian information criterion

An obvious Bayesian approach is to consider all possible models, and then to compute the marginal posterior probability for each model; sensitivity to the priors put on the model parameters.

In the Bayesian framework the goal of summarizing the evidence for or against two alternative models can be achieved by the **Bayes factor**, which unavoidably depends on the choice of the prior.

A Bayesian criterion, similar to the AIC, is the **Bayesian information criterion (BIC)** which selects the model minimizing:

$$\text{BIC} = -2\ell(\widehat{\theta}; y) + \log n \dim(\theta)$$

The BIC does not consider the prior and it is obtained by using a suitable approximation for the marginal likelihood $\int f(\theta) f(y \mid \theta) d\theta$ and by dropping the prior, in a somewhat artificial way.

The term penalizing model complexity is larger in BIC than in AIC; beyond the Bayesian interpretation, the BIC (as the AIC) belongs to the class of information (or predictive) criteria for model selection.

Example: black cherry trees

Data set with measurements of the girth (diameter of the tree, in inches, measured at a fixed distance above the ground), height (ft) and volume of timber (cubic ft) in $n = 31$ felled black cherry trees:

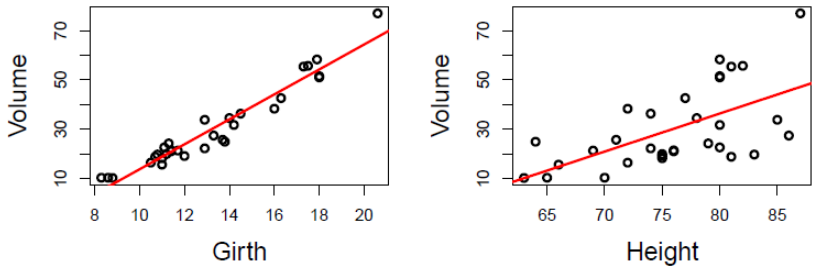
Two linear regression models (the residuals ε are i.i.d. $N(0, \sigma^2)$ random variables):

Model 1 : volume $= \alpha + \beta \cdot$ girth $+\varepsilon$

Model 1 : volume $= \alpha + \beta_1 \cdot$ girth $+\beta_1 \cdot$ height $+\varepsilon$

Model 1 : $\log \text{Lik} = -87.82, \text{AIC} = 181.64, \text{BIC} = 185.95, \text{CV} = 92.36$

Model 2 : $\log \text{Lik} = -84.45, \text{AIC} = 176.91, \text{BIC} = 182.65, \text{CV} = 90.62$



3.8 Contigency tables

3.8.1 Bivariate tables

Two-way tables, called contingency tables, display the observed frequencies associated to bivariate sample data.

Given an i.i.d. n-dimensional sample from a bivariate random variable (X, Y), where X has $r > 1$ categories and Y has $c > 1$ categories, the observed counts $n_{ij}, i = 1, \ldots, r, j = 1, \ldots, c$, related to the combination of categories, can be summarized in:

	y_1	y_2	$\cdots$	y_c	
x_1	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{1+}
x_2	n_{21}	n_{22}	$\ldots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r+}
	n_{+1}	n_{+2}	$\ldots$	n_{+c}	n

where n_{i+} and n_{+j} are, respectively, the row and the column sums.

The marginal random variables X and Y are categorical with a finite number of possible values; also discrete and continuous random variables can be suitably categorized.

3.8.2 Chi-squared test of independence

The underlying statistical model is a **multinomial distribution** with $r \times c$ cells with probabilities $p_{ij} = P(X = x_i, Y = y_j), i = 1, \ldots, r \; j = 1, \ldots, c$, and the number of independent trials is n.

The multinomial distribution is a generalization of the binomial distribution for experiments with more than two possible outcomes.

There are several tests for contingency tables, but the most commonly used one is the **chi-squared test of independence**.

The test is suitable to assess the null hypothesis that paired observations on two variables are independent of each other, that is:

$$\begin{aligned} H_0: &\quad p_{ij} = p_{i+}p_{+j}, \text{ for each } (i,j) \\ H_1: &\quad p_{ij} \neq p_{i+}p_{+j}, \text{ for at least one } (i,j) \end{aligned}$$

with $p_{i+} = \sum_{j=1}^{c} p_{ij} = P(X = x_i), p_{+j} = \sum_{i=1}^{r} p_{ij} = P(Y = y_j)$ An equivalent expression is $H_0 : p_{ij}/p_{+j} = p_{i+}$ for each (i,j) (so that $p_{ij}/p_{i+} = p_{+j}$) : namely, $P(X = x_i \mid Y = y_j) = P(X = x_i)$ (and $P(Y = y_j \mid X = x_i) = P(Y = y_j)$).

Under H_1, the cell probabilities p_{ij} are estimated by the **observed proportions:**

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}$$

whereas, under H_0, they are estimated by the **expected proportions under independence:**

$$\hat{p}^0_{ij} = \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{e_{ij}}{n}$$

The chi-squared statistic compares the observed n_{ij} with the expected (under independence) $e_{ij} = n_{i+}n_{+j}/n$, and computes the score:

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Under H_0, for **large tables**, the test statistic X^2 follows, approximately, the chi-squared distribution $\chi^2((r-1)\cdot(c-1))$. Large values of X^2 point against H_0, so that, given the observed value x^2 of X^2, the (approximate) p-value is $p = P_{H_0}(X^2 \geq x^2)$.

The chi-squared distribution is a good approximation for the true distribution of X^2, provided that $e_{ij} \geq 5$ for all the cells.

For **sparse tables**, the **exact Fisher test** can be considered; alternatively, the p-value can be obtained by simulation, generating many samples under H_0 and computing the distribution of X^2.

The requisite of random sampling is really important, and lack of thereof may invalidate the procedure; nonrandom sampling may occur in case of **repeated observation** of the same units over time, **clustering and self-selection of individuals**.

In case of small p-value, it might be useful to investigate the reasons that lead to it by analyzing the **(Pearson) standardized residuals**:

$$\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

In large tables, under H_0, such quantities are roughly standard normally distributed, so residuals with absolute value larger than 2 point to departure from the independence hypothesis.

Example: steel rods

Four machines produce steel rods, whose diameter can be not defective (ok), too short (short) or too long (long).

A sample of $n = 500$ steel rods is randomly selected and two categorical variables, type of machine and diameter, are observed; the available sample counts n_{ij} form the two-way table:

type of machine		diameter short	ok	long	Total
machine	1	10(15.84)	102(96.48)	8(7.68)	120
machine	2	34(26.40)	161(160.80)	5(12.80)	200
machine	3	12(13.20)	79(80.40)	9(6.40)	100
machine	4	10(10.56)	60(64.32)	10(5.12)	80
Total		66	402	32	500

where in parenthesis are the expected observations e_{ij} under independence. The presence of dependence between type of machine and diameter implies that the proportions of rod types produced by different machines are different.

The observed value of X^2 is 15.58, with p-value 0.016, giving a moderate evidence against the independence hypothesis. The standardized residuals are reported below:

type of machine		diameter short	ok	long
machine	1	-1.4673552	0.56197944	0.1154701
machine	2	1.4791480	0.01577201	-2.1801663
machine	3	-0.3302891	-0.15613491	1.0277402
machine	4	-0.1723281	-0.53865504	2.1566757

There are two large residuals in the cells with coordinates (2,3) and (4,3). The evidence is that the proportion of rods with diameter too long is too low for machine 2, whereas the proportion of rods with diameter too long is too high for machine 4.

3.8.3 Comparing multinomial populations

The $r \times c$ contingency table and the chi-squared test is a convenient formalism also whenever there are r i.i.d. independent samples obtained from r distinct multinomial populations. In this case the rows represent different populations and report the observed frequencies of the c categories of the interest r.v.Y.

The row totals $n_{i+}, i = 1, \ldots, r$, are assumed to be fixed and correspond to the dimension of the r samples, so that $n = \sum_{i=1}^{r} n_{i+}$.

The **chi-squared test** X^2 may be considered as-well for assessing whether the r populations follows the same multinomial distribution:

$$H_0: \quad p_{1j} = p_{2j} = \cdots = p_{rj}, \text{ for each } j = 1, \ldots, c$$

H_1 : $\exists i, j$ such that $p_{ij} \neq p_{kj}$ for a category j. For a 2×2 contingency table, the chi-squared test compares two independent Bernoulli populations and it is equivalent to the test assessing whether the "success" probabilities are reasonably the same.

Example: labor training program

Data on the high school dropouts: two samples of, respectively, 128 individuals who had participated in labor training programs and 297 individuals who had not are summarized in a 2×2 contingency table:

	high school graduate		
program	ye s	no	Total
yes	63(43.07)	65(84.93)	128
no	80(99.93)	217(197.07)	297
Total	143	282	425

In parenthesis are the expectled observations e_{ij} under the null hypothesis that the proportion of dropouts is the same in the two populations.

The observed value of the test statistic X^2 is 19.893 and, since under H_0 it is approximately $\chi(1)$ distributed, the p-value is lower than 0.00001 , giving substantial evidence against the null hypothesis.

The conclusion is in accordance with that of the z test based on the difference between the observed proportions (the value of X^2 is the square of that of Z), thought it does not reflect the sign of the difference.

3.8.4 Concerning statistical reasoning

Data on the high school dropouts: two samples of, respectively, 128 individuals who had participated in labor training programs and 297 individuals who had not are summarized in a 2×2 contingency table:

	high school graduate		
program	yes	no	Total
yes	63(43.07)	65(84.93)	128
no	80(99.93)	217(197.07)	297
Total	143	282	425

In parenthesis are the expectled observations e_{ij} under the null hypothesis that the proportion of dropouts is the same in the two populations.

The observed value of the test statistic X^2 is 19.893 and, since under H_0 it is approximately $\chi(1)$ distributed, the p-value is lower than 0.00001 , giving substantial evidence against the null hypothesis.

The conclusion is in accordance with that of the z test based on the difference between the observed proportions (the value of X^2 is the square of that of Z), thought it does not reflect the sign of the difference.

4 Linear regression with a single predictor

Let's start with an introduction: An important objective in scientific research concerns the study of the relation (useful for both interpretation and prediction purposes) among a **response variable** and some **explanatory variables** (regressors, predictors or covariates).

The focus here is on the straight line model, namely the **simple linear regression model** which is based on the linear relation of one response variable and a single predictor variable.

Data for which these models may be appropriate can be displayed as a scatterplot. By convention, the x-variable, plotted on the horizontal axis, has the role of explanatory variable, whereas y-variable, plotted on the vertical axis, has the role of response or outcome variable.

Although the interest is on the linear model, the use of transformations makes it possible to accommodate specific forms of non-linear relationship within this framework.

Many of the issues that arise for these simple regression models are fundamental to any study of regression methods.

4.1 Data about two variables

If data about the response Y and the regressor X are available, it is appropriate to start from a data visualization by means of a scatterplot, perhaps supplemented by the analysis of the correlation.

If there are many observations, it is often useful to compare the fitted line with a fitted smooth curve. If they differ substantially, then straight line regression could be inappropriate.

Fitting a straight line is often quite natural, and this corresponds to assume a **simple linear regression model** defined as:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where, given the x_i, the error term is **normally distributed**, namely $\varepsilon_i \sim N(0, \sigma^2)$, and errors of different units are **independent**.

This amounts to say that, given x_i (which is taken as fixed in regression models), the i-th response is $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, independent from the other responses.

The normality assumption and the independence assumption are important for obtaining confidence intervals and perform hypothesis testing on the model parameters, and on β in particular (β is the key parameter, connecting the two variables together).

Two types of predictor variable may be considered:

- **metric predictor variables**, that is measurements of some quantity that may help to predict the value of the response (i.e. if the response is the blood pressure of patients, then age or fat mass are potential metric predictors);
- **factor predictor variables**, that is labels that serve to categorize the response measurements into groups, as in the ANOVA model (i.e. if the response is the blood pressure of patients, then a factor may be the drug treatment).

Here, metric predictor variables are mainly considered and the case of factor predictor variables is briefly discussed, with the aim of introducing ANOVA models and comparing regression models and ANOVA models when the levels of the factor are quantitative.

4.2 Estimation and testing

A popular procedure for estimating the regression parameters α, β is the **least squares method**, giving:

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}, \quad \widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

These estimators are unbiased and consistent; as a consequence of the normality assumption, they correspond to the MLE's (maximum likelihood estimations).

Given the standard errors $\mathrm{SE}(\widehat{\alpha}), \mathrm{SE}(\widehat{\beta})$, the test statistics for the nullity of the coefficients (that is, $H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$ and $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$) are, under H_0

$$\frac{\widehat{\alpha}}{\mathrm{SE}(\widehat{\alpha})} \sim t(n-2), \quad \frac{\widehat{\beta}}{\mathrm{SE}(\widehat{\beta})} \sim t(n-2)$$

Testing the null hypothesis that $\beta = 0$ is particularly relevant; a small p-value leads to the rejection of H_0 and it is consistent with an evident linear trend.

Using the parameter estimates $\widehat{\alpha}, \widehat{\beta}$, it easy to estimate the means $\mu_i = \alpha + \beta x_i$ of the response random variables Y_i, which correspond to the fitted values (prediction of points on the fitted line):

$$\widehat{y}_i = \widehat{\mu}_i = \widehat{\alpha} + \widehat{\beta} x_i, i = 1, \ldots, n$$

The observed residuals are given by:

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i, i = 1, \ldots, n$$

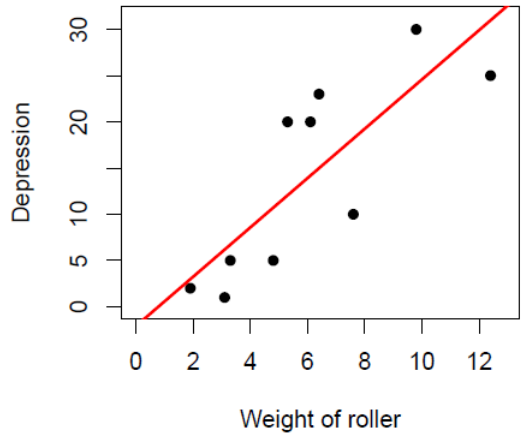
An estimate for σ is the **residual standard error**:

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i\right)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n-2}}$$

with $n - 2$ being the **degrees of freedom**.

4.3 Example: roller data

Experiment where different weights (t) of roller were rolled over different part of a lawn, and the depression (mm) measured. The scatterplot of the data, with the fitted regression line is given below:

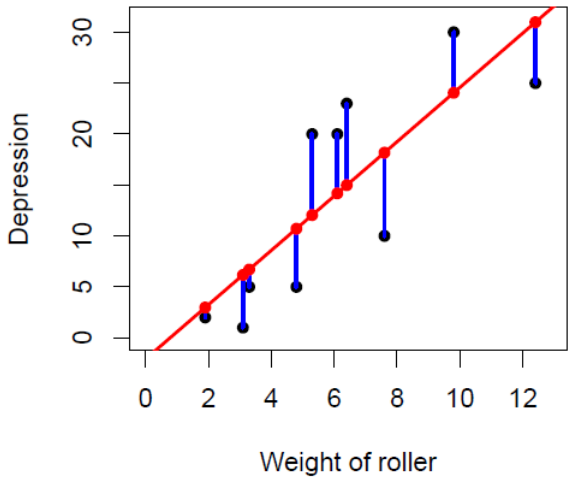


The intercept of the fitted line is $\widehat{\alpha} = -2.09$ ($\mathrm{SE}(\widehat{\alpha}) = 4.75$), the estimated slope is $\widehat{\beta} = 2.67$ ($\mathrm{SE}(\widehat{\beta}) = 0.70$). The standard deviation σ of the noise term is estimated by the residual standard error 6.735 , with 8 degrees of freedom.

The p-value for the slope (testing $\beta = 0$) is 0.005, consistent with the evident linear trend.

The p-value for the intercept (testing $\alpha = 0$) is 0.67, i.e. the difference from zero may well be random sampling error (it would be reasonable to fit a model that lacks an intercept term).

Fitted values (red points on the fitted line) and observed residuals (blue segments) for the roller data:



4.4 Condifence Intervals

Confidence intervals may be calculated for the model parameters and for the regression line at some given value x_0, namely $\mu_0 = \alpha + \beta x_0$. A 95% confidence interval for β has the form:

$$\left[\widehat{\beta} \pm t_{n-2;0.025}\mathrm{SE}(\widehat{\beta})\right]$$

and analogously for α. A 95% confidence interval for μ_0 has a similar form and it is given by:

$$\left[\widehat{\mu}_0 \pm t_{n-2;0.025}\,\mathrm{SE}\left(\widehat{\mu}_0\right)\right]$$

with $\widehat{\mu}_0 = \widehat{\alpha} + \widehat{\beta} x_0$ and:

$$\mathrm{SE}\left(\widehat{\mu}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

where σ can be estimated by $\widehat{\sigma}$. Note that $\mathrm{SE}\left(\widehat{\mu}_0\right)$ is affected by how far x_0 is from the covariates sample mean $\bar{x}$.

4.5 Prediction Intervals

A **prediction interval** for the response random variable $Y_0 = \mu_0 + \varepsilon_0$ at a new predictor value x_0 can also be obtained. In this case the interval provides a set of values for a random variable and it incorporates also the variability due to the random term ε_0.

The best **point predictor** $\widehat{Y}_0$ for Y_0 is again $\widehat{\mu}_0$, namely

$$\widehat{Y}_0 = \widehat{\mu}_0 + \widehat{\varepsilon}_0 = \widehat{\mu}_0 + 0 = \widehat{\mu}_0$$

since the best prediction of the random term ε_0 is $E\left(\varepsilon_0\right) = 0$.

The **prediction error** is $Y_0 - \widehat{Y}_0 = Y_0 - \widehat{\mu}_0$ and then

$$E\left(Y_0 - \widehat{\mu}_0\right) = 0, \quad V\left(Y_0 - \widehat{\mu}_0\right) = \sigma^2 + \mathrm{SE}\left(\widehat{\mu}_0\right)^2$$

The square root of $\sigma^2 + \mathrm{SE}\left(\widehat{\mu}_0\right)^2$, describing the variability of the point predictor, defines the **standard error of prediction**, whose estimate is denoted as $\mathrm{SE}\left(\widehat{Y}_0\right)$, which is greater than $\mathrm{SE}\left(\widehat{\mu}_0\right)$.

The 95% prediction interval for Y_0 (wider than that for μ_0) is:

$$\left[\hat{Y}_0 \pm t_{n-2;0.025}\,\mathrm{SE}\left(\widehat{Y}_0\right)\right]$$

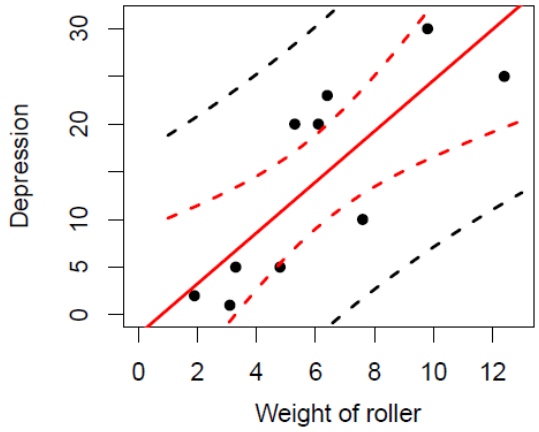
4.6 Example: roller data

For the roller data, the 95% confidence interval for β is:

$$[2.67 \pm 2.31 \times 0.70] = [1.05, 4.28]$$

It does not contain 0; as seen before, the null hypothesis $H_0 : \beta = 0$ is rejected with a 5% significance level.

Figure below shows the 95% pointwise confidence bounds for (points μ_0 on) the fitted line (dashed lines) and the 95% pointwise prediction bounds for new data Y_0 with different values for x_0 (**dashed lines**).



https://www.graphpad.com/support/faqid/1506/ ↦ Link to confidence vs prediction intervals.

4.7 One-Way ANOVA

One-way analysis of variance (ANOVA) is a set of techniques to compare the means of several groups, generalizing two-sample comparisons.

It can be interpreted as the study on how the **mean level** of a continuous response variable depends on the **level of a factor**, which may be viewed as a categorical-type regressor.

It can be extended to more than one factor, obtaining **two-way** or **multi-way ANOVA**.

It is widely used in data from designed experiments, but it has a role also with observational data; indeed, it is an inferential method sometimes employed as an EDA tool.

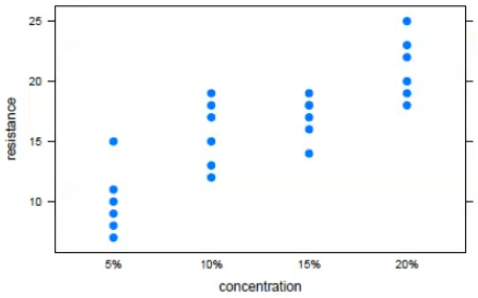
The underlying statistical model is a special case of a linear regression model.

4.7.1 Example: paper resistance

The table below reports the data on an experiment to study the relation between paper resistance and wood fibre concentration in pulp. There are 4 different levels of concentrations, and 6 trials are made at each level (the data are balanced, as appropriate for ANOVA).

	observation							
concentration	1	2	3	4	5	6	Total	Mean
5%	7	8	15	11	9	10	60	10.00
10%	12	17	13	18	19	15	94	15.67
15%	14	18	19	17	16	18	102	17.00
20%	19	25	22	23	18	20	127	21.17

A graphical representation of the data is given below with a strip-plot:



The stripplots display within-group variability and give an indication of differences among the group means (of the response variable resistance). There is a single explanatory variable (regressor), namely the concentration, with one level for each of the different treatments that were applied.

Variances seem similar for the four different treatments (the levels of concentration).

A simple-minded approach is to calculate the means for each of the four treatments, and then examine all pairwise comparisons.

The use of an analysis of variance (really, as noted below, the fitting of a linear model) enables an overall analysis.

4.8 Statistical model for one-way ANOVA

The setting is that there are a levels of a factor of interest (treatment), which identify different groups of observations. The statistical model for one-way ANOVA is:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $i = 1, \ldots, a, j = 1, \ldots, n$, and:

- i identifies the treatment level;
- j identifies the observation;
- μ is the general mean;
- τ_i is the treatment effect, namely the deviation from the general mean for the i-th group;
- ε_{ij} is a random error.

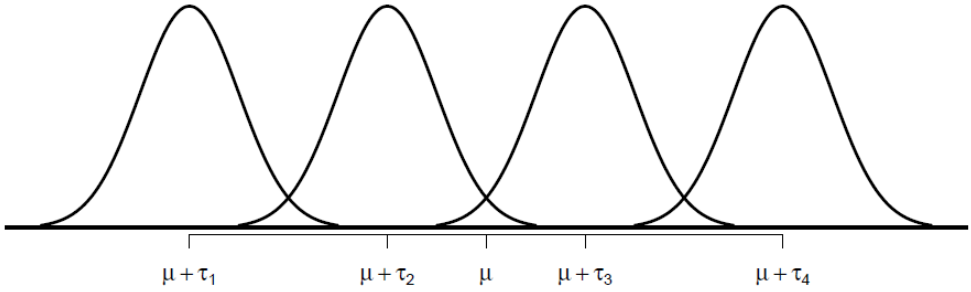
The random errors are assumed i.i.d. normal distributed, that is:

$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right) \qquad \text{i.i.d.}$$

Namely, the Y_{ij} are independent random variables with:

$$Y_{ij} \sim N\left(\mu + \tau_i, \sigma^2\right)$$

For example, if there are $a = 4$ levels, the general mean is $\mu = 0$ and the treatment effects are $\tau_1 < \tau_2 < 0$ and $\tau_4 > \tau_3 > 0$, the probability distributions describing the $a = 4$ groups are such that:



4.9 Hypothesis testing in one-way ANOVA

The null hypothesis states that all the means are equal:

$$H_0: \quad \tau_1 = \tau_2 = \ldots = \tau_a = 0$$

$$H_1: \quad \tau_i \neq 0 \text{ for at least one group}$$

If H_0 is true, the data are a random sample from a $N\left(\mu, \sigma^2\right)$ distribution, with no effect of the factor on the mean response.

The test is performed by comparing two estimates of the variance σ^2:

$$\widehat{\sigma}_0^2 = \frac{\sum_{i=1}^{a} n\left(\bar{y}_i - \bar{y}\right)^2}{a-1} \qquad \widehat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2}{an-a}$$

where $\bar{y}_i$ are the **group means**, and $\bar{y}$ the **grand mean**.

The numerator of $\widehat{\sigma}_0^2$ and $\widehat{\sigma}^2$ are referred to as between-group sum of squares and residuals sum of squares, and their denominators as treatment d.f. and residual d.f.

While $E\left(\widehat{\sigma}^2\right) = \sigma^2$, for $\widehat{\sigma}_0^2$ it holds $E\left(\widehat{\sigma}_0^2\right) = \sigma^2$ only when H_0 is true: from their comparison we can obtain some information on the plausibility of H_0.

Indeed, the F test for one-way ANOVA is just the ratio $F = \widehat{\sigma}_0^2/\widehat{\sigma}^2$ of the two estimators of σ^2 and under H_0:

$$F = \frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} \sim F(a-1, an-a)$$

The results of the analysis can be trusted as long as the statistical model is reasonable; model checking can be made following the guidelines for linear regression models. The nonparametric equivalent of the F test is the Kruskal-Wallis test based on ranks, which extends the Mann-Whitney U test when there are more than two independent groups.

The normal distribution for the residuals is not required and, when the group distributions have the same shape and scale, the comparison concerns the group medians.

4.10 Post-hoc analysis

If the p-value is small, the ANOVA test is significant, so that at least one mean must differ from the others (because the null-hypotesis that all the means are the same can be rejected).

In this case, it is recommendable to investigate the results by means of a post-hoc analysis, such as the **Least Significant Difference (LSD)** or the **Tukey's Honest Significant Difference (HSD)**.

LSD and HSD perform testing on which effects are significantly different at a given level (5% or 1% are typically used); both these two procedures provide a fixed quantity as output, and effects which differ by less than it are deemed as not statistically different.

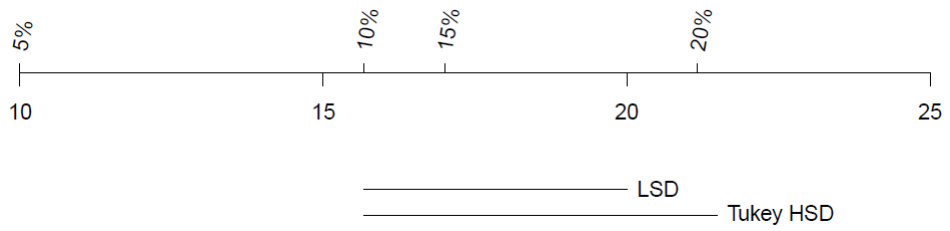
LSD does not take into account that there are $a(a-1)/2$ possible comparisons rather than a single test, hence it is often liberal: the reported output is too short and tends to differentiate too much the various group effects.

On the contrary, HSD focuses on the maximum difference, hence it is conservative: the reported output is too long and blurs the comparisons; considering both methods is therefore a sensible approach.

4.10.1 Example: paper resistance

For the data on paper resistance at $a = 4$ different level of wood fibre concentration, it easy to obtain $\widehat{\sigma}_0^2 = 127.6$, with $a - 1 = 3$ d.f., and $\widehat{\sigma}^2 = 6.51$, with $an - a = 20$ d.f. The observed value of the F statistic is $F = 19.605$, that gives a p-value $p = 3.6 \cdot 10^{-6}$, leading to a substantial evidence against H_0.

Post-hoc analysis here is useful, and it is provided by considering the LDS and HDS results summarized in the following figure:



The effects for 10% and 15% concentration are surely equivalent, while, for example, the effects for 5% and 15% are different.

4.11 Regression vs qualitative ANOVA

The aim is to compare regression models and ANOVA models, when the levels of the factor are quantitative.

In this context, the factor in the one-way ANOVA may be considered as the predictor in a simple linear regression model.

In the ANOVA framework the statistical tests for qualitative differences between treatment effects ignore the fact that the levels are quantitative.

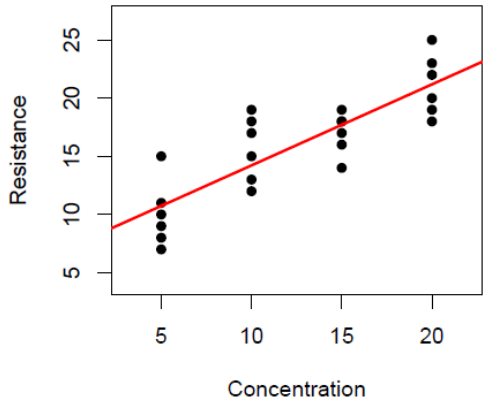
A test for linear trend is more powerful than an analysis of variance test, that treats the levels as qualitatively different levels. The p-values in the first case will on average be smaller.

Fitting a line or a curve, where this is possible, rather than fitting an analysis of variance model that has a separate parameter for each separate level of the explanatory variable, takes proper advantage of structure in the data.

This allows interpolation between successive levels of the explanatory variable and it enables a convenient description for the pattern of the response variable.

4.11.1 Example: paper resistance

Data on paper resistance and wood fibre concentration: 4 levels of concentrations ($5\%, 10\%$, $15\%, 20\%$) and 6 trials at each level. One-way ANOVA uses a different mean for each concentration, but do not use the information about the actual amount. This can be done by fitting a linear model, which passes all the diagnostic checks:



The p-value for the slope β is $2.43 \cdot 10^{-7}$, whereas that one found by the ANOVA analysis is $3.6 \cdot 10^{-6}$. Both values suggest a strong relation, but the p-value for the linear model is about 10 times smaller.

4.12 Checking the residuals

A crucial part of any model fitting concerns judging if the model is appropriate for the data, **by checking residuals and outliers**.

With small data sets, departures from assumptions will be hard to detect. As it is not possible to observe the error term, the observed residuals are used instead. In particular, various **residual plots** are routinely considered:

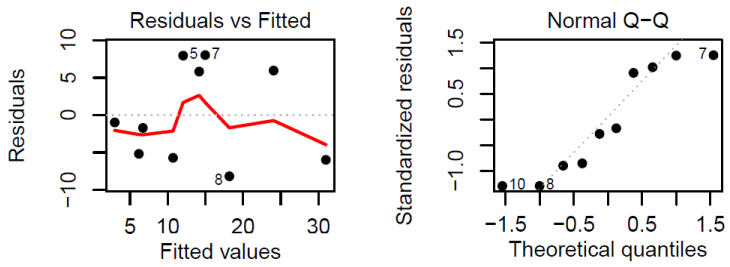
- plot of the observed residuals $\widehat{\varepsilon}_i$ against the fitted values $\hat{y}_i$, for checking lack of systematic patterns (e.g. correlation and clustering);
- plot of the square root of absolute values of the residuals $\sqrt{|\widehat{\varepsilon}_i|}$ against the fitted values $\widehat{y}_i$, for checking if variance is constant;
- plot of $\widehat{\varepsilon}_i$ against x_i, useful for detecting nonlinearity effects of x_i (for a single predictor is equivalent to the plot of $\widehat{\varepsilon}_i$ vs $\hat{y}_i$);
- normal probability plot for checking the normality assumption.

It is possible to show that, no matter whether the model is a good one or not, the residuals invariably have sample mean equal to 0, and variance that depends (to some extent) on x_i.

4.12.1 Example: roller data

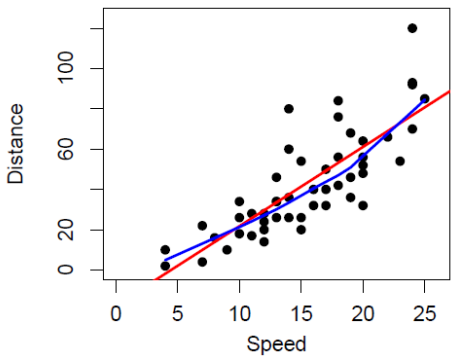
For the roller data, the left panel gives some mild suggestion of clustering, but the sample size is too small to be sure about it.

It is not easy to interpret the normal probability plot (right panel), due to the small data set and the lack of a reference standard. It could be useful to compare this plot against a number of independent plots from normal simulated data with the same number of observations.



4.12.2 Example: cars

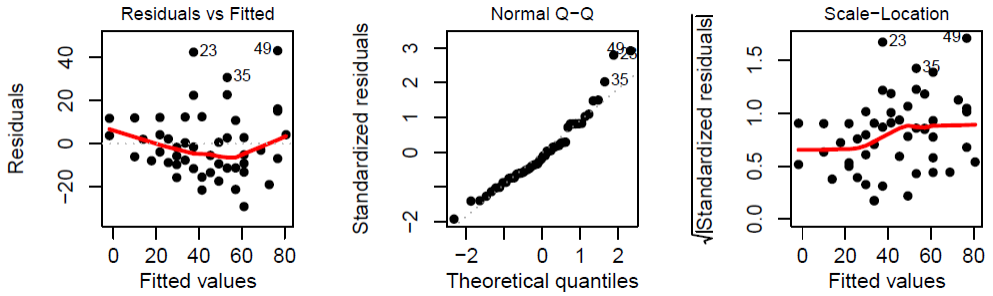
The data set gives the speed of cars (mph) and the distance taken to stop (ft). The observations, recorded in the 1920s regard $n = 50$ cars. The scatterplot of the data, with the fitted regression line and the fitted smooth curve, is given below:



The intercept of the fitted line is $\widehat{\alpha} = -17.58$ (SE($\widehat{\alpha}$) = 6.76), the estimated slope is $\widehat{\beta} = 3.93$ (SE($\widehat{\beta}$) = 0.42). The p-value for the slope is closed to 0, whereas the p-value for the intercept is 0.012 .

The smooth curve gives a better indication of the pattern in the data than the straight line.

The following three diagnostic plots confirms that the simple linear regression model does not gives an adequate description for the cars data:



The curvature in the first plot (correlation for the residuals) and in the third plot (non constant variance) is apparent.

The residuals are from the straight line model. Different conclusions could be obtained from the equivalent plots based on residuals from the smooth curve.

4.13 The ANOVA results and the R^2

Also for linear regression models it is possible to produce the ANOVA results, which decompose the total variability of the response variable into two parts:

- a part accounted for by the linear model;
- a residual part, that for a good model should be small.

In this context, the F test gives a statement on model adequacy which is analogous to that obtained from test statistic for the slope β.

The total variance of the response variable Y can be evaluated by considering the **total sum of squares** (about the mean):

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{SSM} + \text{SSE}$$

which can be expressed as the composition of the **sum of squares accounted for by the linear model** $\text{SSM} = \sum (\widehat{y}_i - \bar{y})^2$ and the **residual sum of squares** $\text{SSE} = \sum (\widehat{y}_i - y_i)^2 = \sum \widehat{\varepsilon}_i^2$. The contribution of X in explaining the variability of Y can be summarized with the quantity R^2 (**coefficient of determination**):

$$R^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

that indicates the proportion of the total variability of Y which is accounted for by the linear function of the predictor X.

The R^2 statistic corresponds to the square of the Pearson correlation coefficient $\hat{\rho}_{XY}$ and its values ranges from 0 in case of no contribution of X (horizontal regression line), to 1 in case of perfect regression (all observations on the regression line).

A less optimistic measure (in general preferable to R^2) is the:

$$\text{adjusted } R^2 = 1 - \frac{\text{SSE/d.f.SSE}}{\text{SST/d.f.SST}}$$

taking into account the degrees of freedom of SSE and SST. Neither statistic gives any direct indication of how well the regression equation will predict when applied to a new data set.

https://blog.minitab.com/blog - Link with some references about R^2;

https://365datascience.com/sum-squares/ - About SST=SSM+SSE.

4.13.1 Example: roller data

For the roller data, the total sum of square is SST $= 657.97 + 362.93 = 1020.90$ and the introduction of weight cuts it down to 362.93.

Then the coefficient of determination is:

$$R^2 = \frac{1020.90 - 362.93}{1020.90} = 1 - \frac{362.93}{1020.90} \doteq 0.64$$

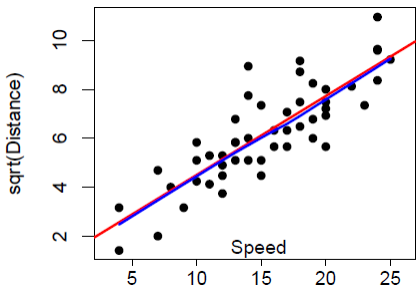
that shows that about 64% of the total variability is accounted for by the linear function of weight. Such result is achieved by passing from the model that just uses only the sample mean to the linear model, which uses two coefficients: the number of degrees of freedom available for estimating the noise variance and the total variance are $10 - 2 = 8$ and $10 - 1 = 9$, respectively. The adjusted R^2 takes this into account and corresponds to:

$$\text{adjusted } R^2 = 1 - \frac{362.93/8}{1020.90/9} \doteq 0.60$$

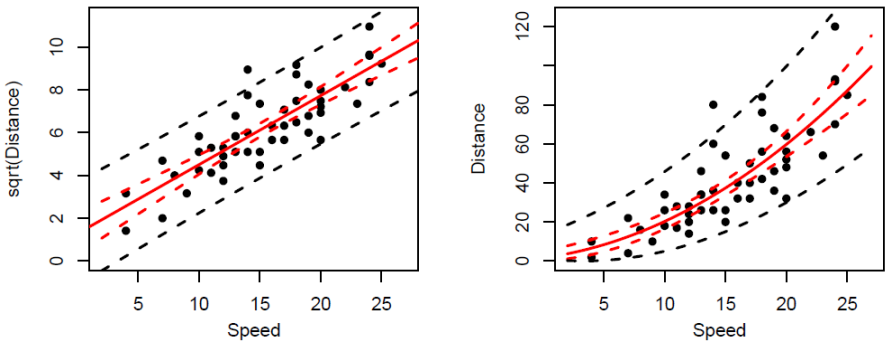
4.13.2 Example: cars

For the cars data, $R^2 = 0.651$ (adjusted $R^2 = 0.644$) shows that about 65% of the total variability of distance is described by the linear function of speed.

In order to account for the non-linearity, that appears in the scatterplot, a linear model with $\sqrt{\text{distance}}$ as response in considered. In this case, the coefficient of determination increases to $R^2 = 0.709$ (adjusted $R^2 = 0.702$) and the improved fit is confirmed by the following scatterplot, with the fitted regression line and the fitted smooth curve:



For the linear model taking $\sqrt{\text{distance}}$ as response, the 95% confidence bounds for the mean response (dashed lines) and the 95% prediction bounds (**dashed lines**) are computed for a range of speed values. The result are reported both on the transformed and on the original scale; the latter choice is safer.



4.14 Outliers, leverage and influence

Outliers are points that lie away from the bulk of the data. Outliers are important because they may:

- carry very useful information, as they may corresponds to the best (or the worst) conditions;
- have an undue influence on the conclusions.

Often, but not always, they correspond to points with a large residual. To this end, it is useful to introduce the concepts of leverage and influence.

The leverage of a data point describes the (potential) impact on the fitted line of moving the point on the y-coordinate.

Points with high leverage are at the extreme end of range of x-values; they may exert a greater pull on the regression line than points towards the center of the range.

Influential points have a strong influence on the model results, and if omitted they would change the fitted line; they are points with a large residual, a large leverage or both.

4.15 Identification and treatment of outliers

Points with high leverage are flagged in the diagnostic plots and they usually correspond to points with large or small x-values.

Influential points can be detected by means of the **Cook's distance**; it measures the change in the fitted line if the point were omitted.

Points with Cook's distances greater than one, or substantially larger than for other points, require investigation.

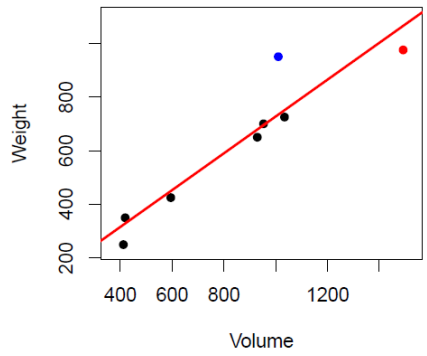
Looking at residuals may not reveal influential points, since an outlier with high leverage tend to attract the fitted line and therefore it may have a small residual.

If some outliers are identified, the first thing to do is to check them carefully, as they may just be a recording error. If an outlier seems a genuine data value, a good practice

is to perform the analysis both with and without the outlier, to assess its impact on the conclusions. Another possibility is to use robust methods, which fit a linear regression model reducing the influence of outliers.

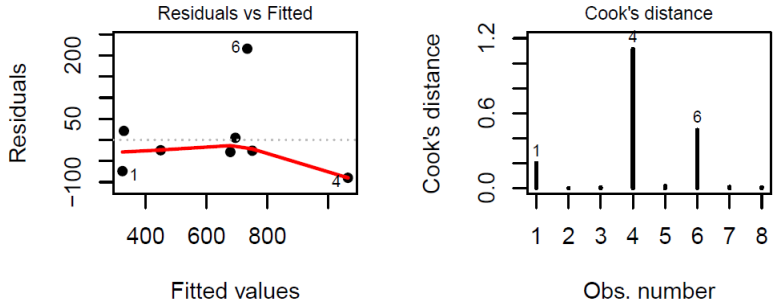
4.15.1 Example: books

The data set gives the volume (cm^3) and the weight (gr) of $n = 8$ paperback books. The scatterplot of the data, with the fitted regression line, is given below:



The intercept of the fitted line is $\widehat{\alpha} = 41.37 (\text{SE}(\widehat{\alpha}) = 97.56)$, the estimated slope is $\widehat{\beta} = 0.686 (\text{SE}(\widehat{\beta}) = 0.11)$. The p-value for the slope is 0.0006, whereas the p-value for the intercept is 0.686.

Indeed, $R^2 = 0.875$ and adjusted $R^2 = 0.854$.



Observations 4 (red point in the scatterplot) and 6 (blue point in the scatterplot) are influential points; observations 4 is also a leverage point.

Although observation 6 has the largest residual, its Cook's distance is relatively small.

Observation 4 has the largest Cook's distance. In part, because this point is a high leverage point, since its y-value is lower than would be predicted by the line, it pulls the line downward.

Points 4 and 6 are both candidates for omission, however with only eight observations, it would not make sense to omit any of them.

4.16 Assessing the predictive accuracy of a model

An important definition in Statistics (and in Machine Learning) is that of **training data**: the data set used to estimate a given model.

If the training data is used also to evaluate the predictive accuracy of a model, this gives an optimistic assessment because the same data are used twice.

An attempt to correct for this fact motivates the use of the adjusted R^2 in place of the R^2 for simple linear regression models.

The ideal approach is to assess the performance of the model on a new data set. Then it is a good practice to evaluate the predictive accuracy of a given model by splitting the data into two separate groups:

- the training set, used to estimate the model;
- the test set, used to assess its predictive performance.

This is a very general idea that can be applied to nearly every statistical method.

4.17 Cross-validation

When this approach is not practical, mainly due to limitations of the available data, **cross-validation** can be used instead.

This consists in splitting the data in k sets (folds), which are used in turn as a test set, with the remaining folds giving the training data.

The predictive assessments for each of the k folds are then combined together into a single measure. Values of k between 3 and 10 are typically used.

Several measures of performances may be defined.

For linear regression models, the estimate of σ^2 (or of the sum of squared residuals or of the mean square error) is a good choice, as it corresponds to the residual variability left unexplained by the regression line.

A particular case of the k-fold procedure is the **leave-one-out cross-validation**, where k is set equal to the sample dimension n, so that each single observation is used in turn as a test set.

4.18 Transformations

Diagnostic plots often point out that something is wrong with the model assumptions.

In many cases, it is just a matter of choosing the right scale for the response variable. There two notable special cases:

- **logarithmic scale**, useful for size measurements of biological organisms, but also for many socio-economic variables that are highly skewed, such as income or satisfaction; also a good option if the ratio between the largest and the smallest response value is large, namely greater than 10 or even 100;
- **square root scale**, useful for count data, it often stabilizes the variance and it is actually a special case of power transformations y^p, with $p = 1/2$; also $p = 1/3$, the cube root scale, is at times useful.

Though the usual course is to transform the y-values, in some instances it may make sense to transform the x-values instead, or both.

4.19 The Box-Cox transformation

Choosing the right scale for the response is often a matter of trial and error, and it might be convenient to be able to do it in a semi-automatic way.

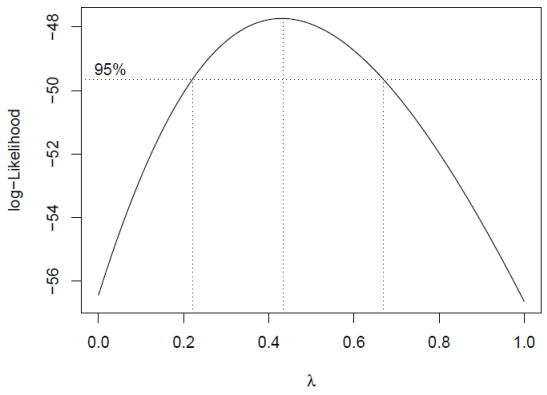
The **Box-Cox transformation** provides exactly that for models with positive responses. It corresponds to a general power transformation, depending on a real parameter λ

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

The specification of a confidence interval for λ may give a quite useful indication. In particular, this would quickly suggest whether the log or the power transformations are likely to work well.

4.19.1 Example: cars

For the cars data, the following plot displays a sort of log-likelihood function for λ and the 95% confidence interval for λ can be read off. The value $\lambda = 0.5$, corresponding to the square root transformation used previously, is among the most supported values.



4.20 The matrix form of simple linear regression

It can be useful to rewrite the simple linear regression model in the following compact matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

- $\mathbf{y} = (y_1, \ldots, y_n)^T$ is column vector collecting all the response values;
- $\mathbf{X}$ is the $n \times 2$ model matrix (or design matrix) defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

- $\boldsymbol{\beta} = (\alpha, \beta)^T$ is the column vector with the model coefficients;

- $\boldsymbol{\varepsilon} =\mid (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the column vector collecting all error terms.

It is not difficult to verify that this matrix form is exactly equivalent to the definition of the model given before.

5 Multiple linear regression and logistic regression

5.1 Introduction to multiple regression

In straight line regression, a response variable Y is regressed on a single explanatory variable.

Multiple linear regression generalizes this methodology to allow **multiple explanatory (predictor) variables, denoted as covariates**.

Multiple linear regression model is one of the most fundamental statistical model.

It is not always the right model, as it is based on some assumptions that are not always reasonable. However, to some extent, a large number of statistical models are an extension of it. Even if it is a rather simple model, and just a rough representation of reality in many cases, it may be extremely useful for both interpretation and prediction purposes.

5.2 Model assumptions

If data about the response Y and the p regressors $X_1, \ldots, X_p$ are available, a **multiple linear regression model** is defined as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where, given the covariates $x_{ij}, j = 1, \ldots, p$, the error term is **normally distributed**, namely $\varepsilon_i \sim N(0, \sigma^2)$, and errors of different units are **independent**.

Thus, given $x_{ij}, j = 1, \ldots, p$, (which are taken as fixed), the i-th response Y_i is normally distributed, independent from the other responses, with constant variance σ^2 and mean defined as a linear combination of the covariates, namely:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

In general, the assumption on $E(Y_i)$ is likely to be false, however, it can be a good approximation or a reasonable starting point for subsequent analysis. The multiple linear regression model can be expressed in matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

- $\mathbf{y} = (y_1, \ldots, y_n)^T$ is column vector collecting all the observed response values;
- X is the $n \times (p+1)$ model matrix, with rank $p+1$ and $n > p+1$, defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the column vector with the model coefficients;
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the column vector collecting the error terms.

Then, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{I}_n$ the identity matrix.

As for simple linear regression, the predictor variables, which form the basic ingredients for X, can be metric (numeric) variables or factor variables.

The focus here is on metric regressors, even though the case with factor regressors will be briefly discussed.

To understand the structure of X, the following example is considered: the response y is described by two numeric regressors, u and v, and by a factor g with labels dividing the observations into three groups.

Factors may be included in the model by means of **dummy variables** (regressors which assume only two values, 1 and 0): in this case, there are three dummy indicators showing whether the corresponding observation belongs to the group, or not.

With regard to numeric variable, they could enter the model also non-linearly: the model is linear in the parameters and error term, but not necessarily in the predictors.

A model matrix, accounting for non-linearities concerning numeric regressors and for a factor regressor with three levels, might be:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & u_1 & v_1 & v_1^2 & u_1 v_1 \\ 1 & 0 & 0 & u_2 & v_2 & v_2^2 & u_2 v_2 \\ 1 & 0 & 0 & u_3 & v_3 & v_3^2 & u_3 v_3 \\ 0 & 1 & 0 & u_4 & v_4 & v_4^2 & u_4 v_4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & u_n & v_n & v_n^2 & u_n v_n \end{pmatrix}$$

Then, the first three observations are in the first group, the fourth is in the second group and the last is in the third group.

When working with factor regressors some care is required to ensure that the model matrix X has full rank, as required in the assumptions. Otherwise, there will be a lack of identifiability: the model parameters can not uniquely be determined from data. Notice that, in this case, the first column in X, which specifies a common intercept parameter, is not present.

5.3 Inference

Point estimates of the linear model parameters $\boldsymbol{\beta}$ can be obtained by the method of least squares, giving

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

The estimator is (minimum variance, linear) unbiased and consistent; for the normality assumption, it corresponds to the MLE (Maximum Likelihood Estimation). Since $\widehat{\beta}$ is just a linear transformation of a normal random vector,

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\beta}))$$

where the variance matrix is $\mathbf{V}(\boldsymbol{\beta}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. Using $\widehat{\boldsymbol{\beta}}$, it easy to estimate the mean vector $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ of the response random vector Y, which correspond to the fitted values

$$\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^T = \widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

with $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ the hat matrix (such that $\text{tr}(\mathbf{H}) = p+1$ and HH = H).

A result, useful for testing hypotheses about individual β_j, as well as for finding confidence intervals for β_j, is:

$$\frac{\widehat{\beta}_j - \beta_j}{\text{SE}\left(\widehat{\beta}_j\right)} \sim t(n-p-1)$$

where the (estimated) standard error $\text{SE}\left(\widehat{\beta}_j\right)$ is given by the square root of the j-th diagonal element of matrix $\mathbf{V}(\boldsymbol{\beta})$.

Interpreting the inferential results on the regression coefficient β_j is not as straightforward as it might appear. As a matter of fact, the p-value is used to test whether the coefficient could be zero, given that the other coefficients remain in the model (i.e. are non-zero).

Since the estimators for the various coefficients are usually not independent, dropping one term (setting it to zero), will change the estimates of the other coefficients and hence their p-values.

For this reason, if the aim is to refine the model by dropping some coefficients, the strategy is to drop only one term at a time (starting from those with the highest p-values) and to refit after each drop.

It is also of interest to obtain distributional results for testing, for example, the simultaneous equality to zero of several model parameters. Such tests correspond to F **tests** with appropriate degrees of freedom. A relevant F test is that one focusing on the global significance of all the regression coefficients, namely on:

$H_0 : \beta_1 = \cdots = \beta_p = 0$ vs $H_1 : \beta_j \neq 0$, for at least one j

Furthermore, the observed **residuals** are given by $\widehat{\varepsilon} = \mathbf{y} - \widehat{\mathbf{y}} = \left(\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n\right)^T$, with $\widehat{\varepsilon}_i = y_i - \widehat{y}_i, i = 1, \ldots, n$

An estimate for σ is the **residual standard error**:

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \left(y_i - \widehat{y}_i\right)^2}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-p-1}}$$

with $n-p-1$ being the **degrees of freedom** and $\sum_{i=1}^n \widehat{\varepsilon}_i^2$ the **residual sum of squares**.

5.4 Confidence and prediction intervals

Confidence intervals may be calculated for the model parameters and for the regression function at some given value for the regressors $\text{x}_0 = \left(1, x_{01}, x_{02}, \ldots, x_{0p}\right)^T$, that is for the expected response at x_0:

$$\mu_0 = \mathbf{x}_0^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \ldots + \beta_p x_{0p}$$

A 95% confidence interval for β_j has the form:

$$\left[\widehat{\beta}_j \pm t_{n-p-1;0.025}\,\text{SE}\left(\widehat{\beta}_j\right)\right]$$

A 95% confidence interval for μ_0 has a similar form and it is given by:

$$\left[\widehat{\mu}_0 \pm t_{n-p-1;0.025}\,\text{SE}\left(\widehat{\mu}_0\right)\right]$$

with $\widehat{\mu}_0 = \mathbf{x}_0\widehat{\boldsymbol{\beta}}$ and:

$$\mathrm{SE}\left(\widehat{\mu}_0\right) = \sigma\sqrt{\mathrm{x}_0^T\left(\mathrm{X}^T\mathbf{X}\right)^{-1}\mathrm{x}_0}$$

where σ can be estimated by $\widehat{\sigma}$.

A **prediction interval** for the response random variable $Y_0 = \mu_0 + \varepsilon_0$, at some new predictor values x_0, can also be obtained. In this case the interval provides a set of values for a random variable and it incorporates also the variability due to the random term ε_0.

The best **point predictor** for Y_0 is again $\widehat{Y}_0 = \widehat{\mu}_0$ and the **prediction error** is $Y_0 - \widehat{Y}_0 = Y_0 - \widehat{\mu}_0$; then:

$$E\left(Y_0 - \widehat{\mu}_0\right) = 0, \quad V\left(Y_0 - \widehat{\mu}_0\right) = \sigma^2 + \mathrm{SE}\left(\widehat{\mu}_0\right)^2$$

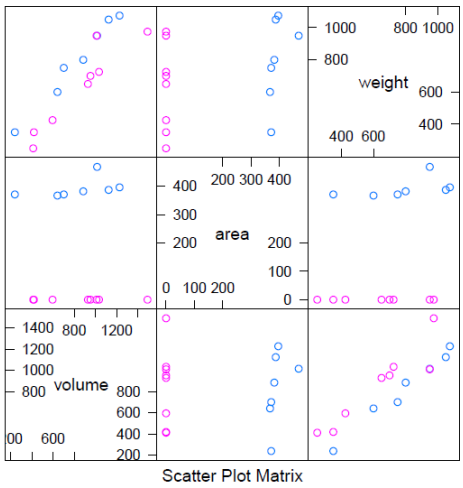
The square root of $\sigma^2 + \mathrm{SE}\left(\widehat{\mu}_0\right)^2$, describing the variability of the point predictor, defines the **standard error of prediction**, whose estimate is denoted as $\mathrm{SE}\left(\widehat{Y}_0\right)$, which is greater than $\mathrm{SE}\left(\widehat{\mu}_0\right)$. The 95% prediction interval for Y_0 (wider than that for μ_0) is

$$\left[\hat{Y}_0 \pm t_{n-p-1;0.025}\mathrm{SE}\left(\widehat{\mathrm{Y}}_0\right)\right]$$

In this framework, also more general confidence (prediction) regions or simultaneous confidence (prediction) intervals, involving more than one parameter (future observation), can be specified.

5.4.1 Example: book weight

Data about a sample of $n = 15$ books; the variables are book volume $\left(\mathrm{cm}^3\right)$, hard board cover area $\left(\mathrm{cm}^2\right)$, book weight (gr) and cover, a factor with levels hardback and paperback. The scatterplot matrix for the numerical variables is given below:



Leaving aside the factor cover (as it provides similar information to cover area), a sensible model for book weight is:

$$\text{weight}_{\ i} = \beta_0 + \beta_1\ \text{volume}_{\ i} + \beta_2\ \text{area}_{\ i} + \varepsilon_i$$

The coefficient estimates are $\widehat{\beta}_0 = 22.41\left(\mathrm{SE}\left(\widehat{\beta}_0\right) = 58.40\right), \widehat{\beta}_1 = 0.708$

$\left(\text{SE}\left(\widehat{\beta}_1\right) = 0.061\right), \widehat{\beta}_2 = 0.468 \left(\text{SE}\left(\widehat{\beta}_2\right) = 0.102\right)$, whereas the estimate for the noise standard deviation (the residual standard error) is $\widehat{\sigma} = 77.66$.

The low p-values for volume ($7.07 \cdot 10^{-8}$) and area (0.0006) highlights that they are both important predictors of book weight.

These results should be used informally, rather than as a basis for formal tests of significance, since the model parameter estimators, and the associated t-tests, are usually not independent.

The information on individual regression coefficients can readily be adapted to obtain a confidence interval for the coefficients and for the regression function and prediction intervals for the book weight.

Indeed, the multiple R^2 and the adjusted multiple R^2 are, respectively, 0.928 and 0.917. An F-statistic can be defined to provide an overall test of significance on the regression, and in particular on the global significance of the regression coefficients, that is on:

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

In the example, both this test and each of the individual t-tests on β_1 and β_2 are strongly significant.

However, a significant F-test does not necessarily imply that all the individual t-tests are significant too. The global test is not concerned with the intercept β_0.

Despite what stated in the course textbook, testing about the significance of the intercept makes little sense.

The intercept is something which is convenient to include in a model, since the simplest possible model (the null model) includes at least the intercept. Furthermore, without the intercept, the multiple R^2 measures are not interpretable.

The ANOVA results may be derived also for multiple linear regression, although the interpretation requires great care.

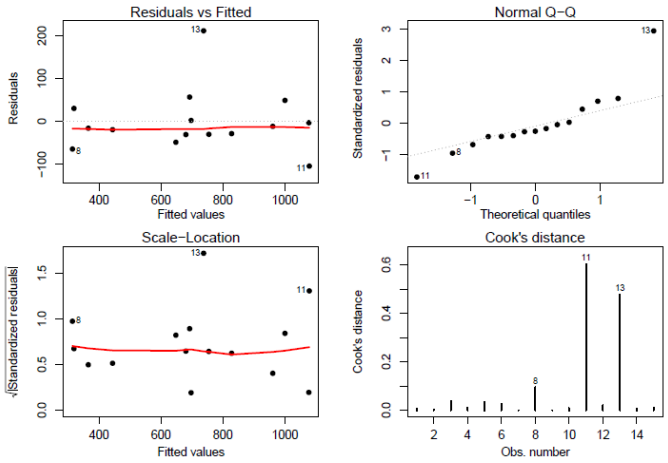
In this example, it is possible to describe the contribution of volume after fitting the intercept and then the contribution of area after fitting both the intercept and volume.

Thus the p-value for area agrees with that obtained using the t-test, since in both cases the p-values are computed in the model where volume is also included. The p-value for volume, instead, differ from that obtained using the t-test, because is computed without considering area.

In this case, actually, the two p-values for volume are very close (both around $7 \cdot 10^{-8}$), but only because the correlation between volume and area is close to zero (0.0015).

This means that the two variables carry separate pieces of information. As for simple linear regression, diagnostic plots can be considered.

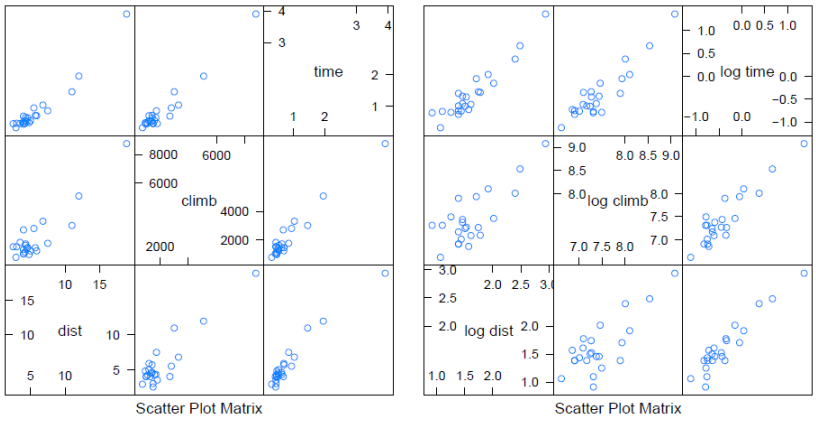
They show that observations 11 and 13 correspond to the largest residuals, and are influential points. However, they seem legitimate observations, and with only 15 points it is better not remove them:



5.4.2 Example: hill races

Data on $n = 23$ hill races in Northern Ireland; the variables are: the distance dist (miles), the heights climbed climb (ft), male record time time (hours), female record time timef (hours).

Focusing on the male times only, the scatterplot matrix for original and log data reveal some linear relationships. Taking the logs seems preferable.



These considerations suggest fitting the model

$$\log(\text{time}_i) = \beta_0 + \beta_1 \log(\text{dist}_i) + \beta_2 \log(\text{climb}_i) + \varepsilon_i$$

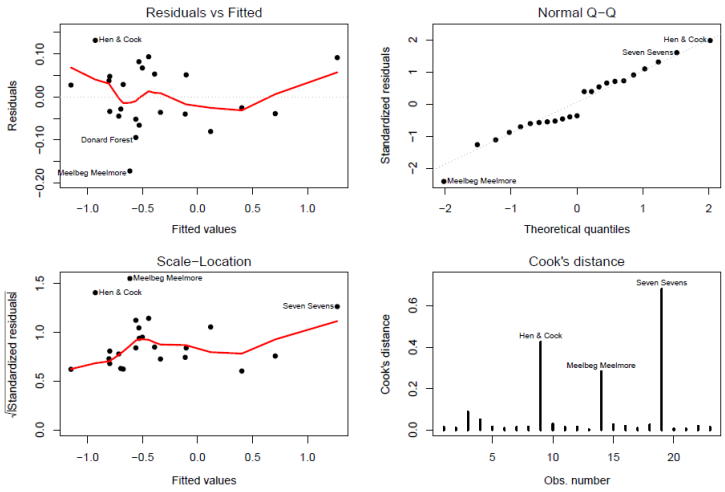
This is equivalent to a model with a deterministic part described, in the original scale, by the power relationship:

$$\text{time} = e^{\beta_0} \cdot \text{dist}^{\beta_1} \cdot \text{climb}^{\beta_2}$$

The coefficient estimates are $\widehat{\beta}_0 = -4.96$ $\left(\text{SE}\left(\widehat{\beta}_0\right) = 0.273\right)$, $\widehat{\beta}_1 = 0.68$ $\left(\text{SE}\left(\widehat{\beta}_1\right) = 0.055\right)$, $\widehat{\beta}_2 = 0.47$ $\left(\text{SE}\left(\widehat{\beta}_2\right) = 0.045\right)$, whereas the estimate for the noise standard deviation (the residual standard error) is $\widehat{\sigma} = 0.076$. The low p-values for log(dist) and log(climb) highlights that they are both important predictors of log(time). The global significance of the regression coefficients is confirmed by the F-test.

Indeed, the multiple R^2 and the adjusted multiple R^2 are, respectively, 0.983 and 0.981.

The diagnostic plots do not show any problem, apart from the moderately large residual associated with the Meelbeg Meelmore race



If the aim of the analysis is the interpretation of model coefficients, it is important to emphasizes that different formulations of the regression model, or different models, may serve different explanatory purposes. To this regard, notice that the deterministic part of the fitted model is, on the original scale,

$$\text{time} = 0.007 \cdot \text{dist}^{0.68} \cdot \text{climb}^{0.47}$$

Surprisingly, the relative rate of increase in time is 68% of the relative rate of increase in distance, holding climb constant.

This implies that for a fixed value of climb the time is smaller for the second mile than for the first mile: indeed, setting climb = 1500, the times are, respectively,

$$0.007 \cdot 1^{0.68} \cdot 1500^{0.47} = 0.218 \quad 0.007 \cdot 2^{0.68} \cdot 1500^{0.47} = 0.349$$

This seems quite strange, but the explanation comes from the meaning of keeping climb constant, since short races will be steeper than long races.

The coefficient for log(dist) is, reassuringly, greater than 1 if log(time) is regressed on log(dist) and log(climb/dist), instead of log(climb). Then,

$$\text{time} = 0.007 \cdot \text{dist}^{1.15} \cdot (\text{climb}/\text{dist})^{0.47}$$

Note that the two models provide the same fit, since they are different mathematical formulations of the same underlying model. Interpretability issues and application-specific considerations will drive the choice of a particular model form.

There is another, related, benefit in the second model. The correlation between log(dist) and log(climb/dist) is −0.065, negligible relative to the correlation of 0.78 between log(dist) and log(climb).

In designed experiments, uncorrelated regressors are usually considered. Even in observational studies, when possible, it is preferable to include terms in the model which have

small cross-correlation. In some sense, if two covariates x_1 and x_2 are correlated, the effect of x_1 on the response is not expressed only by $\widehat{\beta}_1$, but also by $\widehat{\beta}_2$ through the relation between x_1 and x_2.

5.5 Centering the covariates

Centering the covariates amounts to subtraction of their sample mean before introducing them in the model. For example, the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

can be written as:

$$y_i = \alpha + \beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2) + \varepsilon_i$$

This helps model interpretability in two ways: the estimated intercept is $\widehat{\alpha} = \bar{y}$, and it is the fitted value obtained when both the covariates are equal to their sample mean; the observed values can be written as:

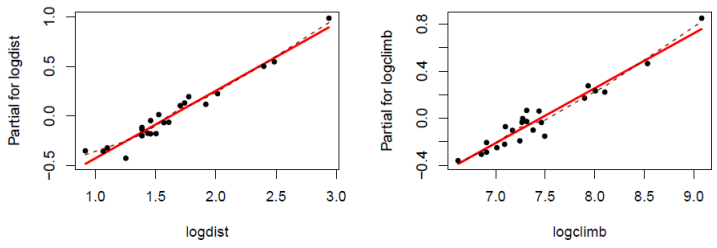
$$y_i = \bar{y} + \widehat{\beta}_1 (x_{i1} - \bar{x}_1) + \widehat{\beta}_2 (x_{i2} - \bar{x}_2) + \widehat{\varepsilon}_i = \bar{y} + t_{i1} + t_{i2} + \widehat{\varepsilon}_i$$

where t_{i1} and t_{i2} are zero-sum contributions from each covariate. The terms t_{i1} and t_{i2} can be used for the partial residual plots.

5.6 Partial residual plot

The partial residual plot for the covariate x_j, given all the others, it is a scatterplot of $t_{ij} + \widehat{\varepsilon}_i$ vs x_{ij}. It accounts for that part of the response that is not explained by the covariates other than x_j.

For example, with two covariates, the partial residual plot for x_1 graphs $t_{i1} + \widehat{\varepsilon}_i = y_i - \bar{y} - t_{i2}$ against x_{i1}. It assesses whether the part of the response that is not explained by x_2 can be approximated by a linear function of x_1. For the hill races data, using the model with log variables, the two partial residual plots, given below, are quite linear:



5.7 Quadratic effect of a covariate

In some cases, a certain covariate may have an evident nonlinear effect on the mean response; for instance, this can be highlighted by drawing a partial residual plot. In such case, it is convenient to include in the model more than a term to describe the effect of such covariate.

In the simple case of a single covariate, a model with a quadratic effect of x will specify the mean response as:

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Such model is not any longer in the class of simple linear regression models, being instead a special case of multiple linear regression, with two covariates, namely $x_{i1} = x_i$ and $x_{i2} = x_i^2$.

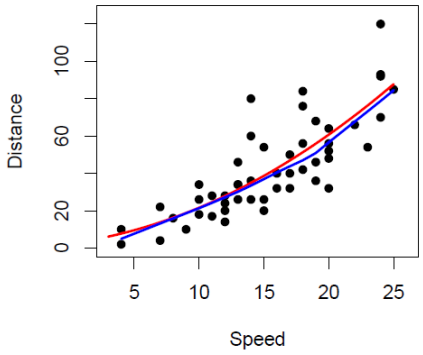
Polynomial regression models generalize, in some sense, the multivariate linear model and they may contain squared, cross-terms and higher-order terms of the original predictor variables.

5.7.1 Example: cars

For the cars data, a preliminary statistical analysis, supported by physical considerations, suggests that the distance taken to stop should be a non-linear function of the speed. Then, a plausible model is:

$$\text{distance}_{\ i} = \beta_0 + \beta_1 \ \text{speed}_{\ i} + \beta_2 \ \text{speed}_{\ i}^{\ 2} + \varepsilon_i$$

The scatterplot of the data, with the fitted regression function and the fitted smooth curve, is given below:



The estimated model parameters are:
$\widehat{\beta}_0 = 2.47 \left(\text{SE}\left(\widehat{\beta}_0\right) = 14.82\right)$, $\widehat{\beta}_1 = 0.91 \left(\text{SE}\left(\widehat{\beta}_1\right) = 2.03\right)$, $\widehat{\beta}_2 = 0.10 \left(\text{SE}\left(\hat{\beta}_2\right) = 0.07\right)$.

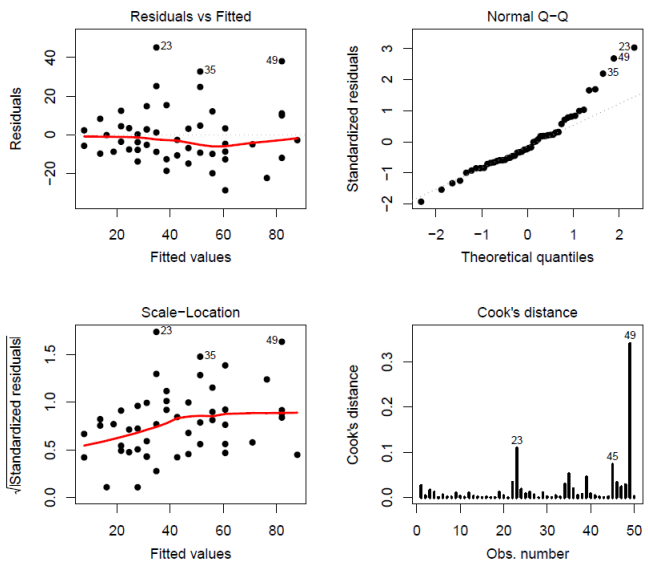
The p-values for all the model coefficients are very high, despite the fact that the p-value for the F test is $5.85 \cdot 10^{-12}$, indicating that there is a strong evidence that the assumed model is better than the constant one.

The p-values are testing whether the corresponding coefficients could really be zero given that the other terms remain in the model: they cannot be taken as an indication that all the terms can be dropped.

The point here is that the lack of independence between estimators creates difficulties in the interpretation of estimates.

The correlation between parameter estimators typically arises from correlation between the corresponding covariates.

In this case it is not possible to entirely separate out their effects on the response by examining the results of model fitting.



The diagnostic plots show some indication of non-constant variance (top left) and of a departure from normality in the residuals (top right).

An alternative model is one in which variability increases with speed; for example, assuming $\varepsilon_i \sim N\left(0, \sigma^2 \cdot \text{ speed }_i\right)$. In this case, the parameter estimates are obtained by the weighted least squares (WLS) method; that is, minimizing $\sum \varepsilon_i^2 w_i$, with $w_i = 1/\text{ speed }_i$.

5.8 Violation of model assumptions

Each of the assumptions underlying multiple linear regression may be not plausible for a certain application. This may involve both the systematic part of the model and the random term.

The assumption that the mean of Y_i is a linear combination of the xs is likely to be just an approximation.

Important deviations may be the **nonlinear effect** of an explanatory variable, and the **lack of a certain one in the data set**. The latter is more serious, and requires careful consideration of the problem under investigation.

The assumption of independence may be not plausible with **clustering** of the observations, or **repeated measurement** over time or space.

The assumptions of constant variance and normality are often violated in practice. A careful choice of the scale for the response variable may make them more plausible.

The presence of **outliers** is something to look into.

5.9 Checking on the residuals

Any of the above problems may be fixed in practice, at least to a certain extent, provided they are readily detected.

Although sometimes it is not easy or possible to detect failure of the assumptions, there are simple checks that, if the assumption fails, may indicate the nature of the failure.

Regression diagnostics are a set of methods which assist such detection.

Residual plots for multiple linear regression are interpreted in the same way as residual plots for simple linear regression. Similarly, it possible to check on **nonlinearity, constant variance** and (approximate) **normality**.

To check nonlinearity in the multiple setting, it can be useful to plot residuals against predictors, rather than relying solely on the default plot concerning fitted values vs estimated residuals.

5.10 Outliers: leverage and influence

Looking for outliers is more complicated when there are several explanatory variables. Two (or more) outliers, that are influential, may mask each other. The outliers are treated as in the simple linear regression model.

To this extent, the concepts of leverage and influence are important.

If the i-th observation y_i is changed into $y_i+\Delta_i$ (leaving all the other y-values unchanged), then the fitted value changes from $\widehat{y}_i$ to $\widehat{y}_i + h_{ii}\Delta_i$, and h_{ii} is called the leverage for the i-th point.

The h_{ii} values are the diagonal element of the **hat (influence) matrix** $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, such that:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

Large leverage values flag points away from the other points.

In a model with p coefficients $\sum h_{ii} = p$, so that values h_{ii} larger than $2p/n$ or $3p/n$ can be considered as large.

Data points that may alter the fitted values (if omitted) are **influential**. Such distortion, regarding the fitted response, is a combined effect of the size of the residual and its leverage.

A common measure of influence is given by the **Cook's distance**.

Given the **standardized residuals**

$$r_i = \frac{y_i - \widehat{y}_i}{s\sqrt{1-h_{ii}}}$$

if the model has p coefficients, the Cook's distance for the i-th observation is:

$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1-h_{ii}}$$

It measures the change in model estimates when the i-th observation is omitted.

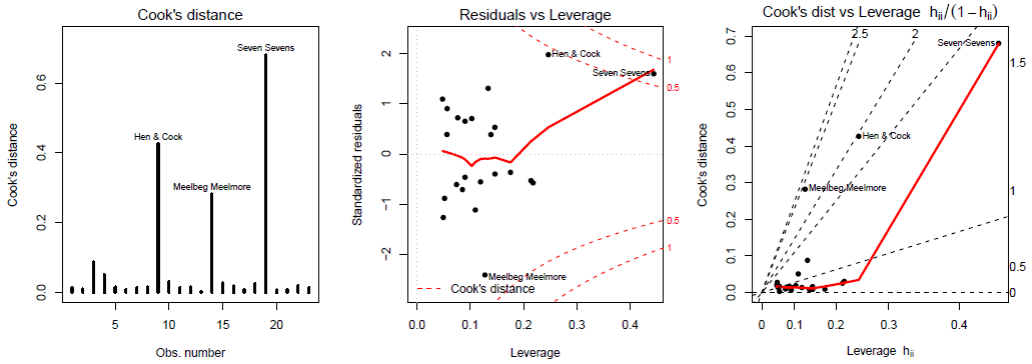
Values of D_i larger than 0.5 (or even 1.0) are suspicious, and may refer to points with a strong effect on the estimated coefficients and the estimated standard errors.

The (standardized) effect of each observation on the estimates $\widehat{\boldsymbol{\beta}}$ can also be evaluated.

5.10.1 Example: hill races

For the log data on hill races in Northern Ireland, the following diagnostic plots describe Cook's distances and leverages. They do not indicate serious problems, apart a point with a Cook's distance larger than 0.5.

Evaluation of the (standardized) effect of each observation on the estimates shows that none of the three observations with the largest Cook's distance has a worrisome effect.



5.11 R^2 and adjusted R^2

Goodness of fit for linear models is usually measured in terms of the proportion of response variability explained by the model, as quantified by the **coefficient of determination** R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

The R^2 tends to overestimate the goodness of fit. The adjusted R^2 is usually preferable, since it accounts for the degrees of freedom:

$$\text{adjusted } R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 / (n - p - 1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n - 1)}$$

Both these measures are not suitable for comparisons between different studies, where the range of values of the explanatory variables may be different. They can be used instead for comparing **different models for the same data**.

Values close to 1 indicate a good fit, but low values do not necessarily indicate a poor model; the data contain a substantial random component.

5.12 Model Selection

The ANOVA results, based on suitable F tests, can be considered for comparing **nested linear models**.

An alternative procedure consists in a sequence of F tests comparing the full model with each of the models produced by dropping a single predictor.

Starting from the full model, the model term with the highest p-value is repeatedly deleted (and the new full model refitted) until all p-values are below some threshold (backward selection).

Starting from a simple model, the model term which has most evidence in an F test is repeatedly added until no more terms would lead to significant improvement (forward selection).

There are also backward-forward strategies, in which cycles of backward and forward selection are alternated until convergence.

Better alternatives, also useful for **non-nested models**, do exist and they aim at evaluating the predictive ability of alternative models.

The information criteria are particularly well suited for choosing the model with the best capability of doing prediction for unobserved data. Criteria based on the cross-validation procedures can be also considered.

The **AIC statistic** is perhaps the most commonly used, and for a linear regression model with p coefficients, is given by:

$$\text{AIC} = n \log\left(\frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n}\right) + 2p + \text{const}$$

where the constant term arises from the assumption of an i.i.d. normal distribution for the errors.

The **BIC statistic**, which replaces $2p$ by $\log(n) \cdot p$, penalizes models with many parameters more strongly.

Both the statistics are smaller for models with better predictive power.

The **Mallow's C_p statistic** (in this framework, equivalent to the AlC) differs from the AlC statistic only by subtraction of n, and by omission of the constant term:

$$C_p = n \log\left(\frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n}\right) + 2p - n$$

At times, with a moderate number of predictors, these selection statistics can be used for semi-automatic model selection procedures, such as **stepwise model selection** (backward and/or forward).

Another approach for variable selection, useful when there are large numbers of predictors relative to the number of data, involves the **lasso method**.

It penalizes the model coefficients towards zero, in such a way that as the penalization increases, many of the coefficient estimates become zero.

Alternatively, the variable selection problem can be solved using suitable **boosting algorithms**.

5.12.1 Example: cars

For the cars data, the following models are considered:

In M_1 and M_2 the WLS method, with $w_i = 1/\text{ speed }_i$, is considered.

$$\begin{aligned}
\text{M}_0: &\quad \text{distance }_i = \beta_0 + \beta_1 \text{ speed }_i + \varepsilon_i \\
\text{M}_1: &\quad \text{distance }_i = \beta_2 \text{ speed }_i^2 + \varepsilon_i \\
\text{M}_2: &\quad \text{distance }_i = \beta_0 + \beta_1 \text{ speed }_i + \beta_2 \text{ speed }_i^2 + \varepsilon_i
\end{aligned}$$

The ANOVA comparison of the nested linear models M_0 and M_2 gives a p-value $2.2 \cdot 10^{-16}$ for the F test, indicating strong evidence against M_0.

The comparison between M_1 and M_2 gives a p-value 0.046 giving some evidence against M_1. The AIC statistic suggests again the larger model, since:

$$\text{AIC}\,(\text{M}_0) = 419.16, \text{AIC}\,(\text{M}_1) = 414.80, \text{AIC}\,(\text{M}_2) = 412.26$$

Conversely, the BIC statistic points to the model M_1, penalizing the larger model M_2. Model M_1 has also the larger values for R^2 and adjusted R^2.

5.12.2 Example: hill races

For the log data on hill races in Northern Ireland, the following models are considered:

$$\mathrm{M}_1 : \log(\mathrm{time}_i) = \beta_0 + \beta_1 \log(\mathrm{dist}_i) + \beta_2 \log(\mathrm{climb}_i) + \varepsilon_i$$
$$\mathrm{M}_2 : \log(\mathrm{time}_i) = \beta_0 + \beta_1 \log(\mathrm{dist}_i) + \beta_2 \log(\mathrm{climb}_i) + \beta_3 (\log(\mathrm{dist}_i))^2 + \varepsilon_i$$

In the second one, a quadratic term for log(dist) is included. The adjusted R^2 is about the same for the two models, but both the AIC and the BIC values are smaller for M_2 :

$$\mathrm{AIC}(M_1) = -48.13, \mathrm{AIC}(\mathrm{M}_2) = -49.83$$
$$\mathrm{BIC}(\mathrm{M}_1) = -43.58, \mathrm{BIC}(\mathrm{M}_2) = -44.15$$

Including the quadratic term seems a good idea even if the p-value for the quadratic coefficient is 0.084.

5.13 Suggested steps for model fitting

Examine the distribution of each of the explanatory variables, and of the dependent variable. Look for highly skew distributions, and outlying values.

Examine the scatterplot matrix of all the explanatory variables, and also of the response variable.

Note the ranges of each of the scatterplot variables, considering if they vary sufficiently to affect the response variable and if each of the explanatory variables is accurately measured.

In case some pairwise plots hint to nonlinearity, consider the use of transformations to achieve more nearly linear relationships.

Transformation of the response is advisable in case of skew distribution. The Box-Cox transformation, already introduced for simple linear regression, can be useful for suggesting the right scale.

Pairs of explanatory variables that are so highly correlated that they appear to provide the same information should be analyzed. Background information may suggest which one should be retained.

5.14 Some diagnosis checks

Plot residuals against fitted values. Check for patterns in the residuals, and changes in the variability. Quantile-quantile plots of residuals are also useful, but they should not be taken too seriously.

For each explanatory variable, draw a partial residual plot, to check whether any of the explanatory variables require transformation.

Examine the Cook's distance statistics. In case of doubt, it may be useful to examine the standardized effect of each observation on the model parameter estimates.

In principle, outliers, influential or not, should never be disregarded. Their exclusion may be a result of use of the wrong model.

If apparently genuine outliers remain excluded from the final fitted model, they must be noted in the eventual report and included, separately identified, in graphs.

5.15 Selecting the explanatory variables

When there are several explanatory variables, selecting the variables that give the best prediction becomes an issue.

Start from an informed guess about which variables are likely to be important. Some variables ought to be included in the model even when their contribution to the prediction of the response is limited.

As a practical rule, one suggested rule is that there should be at least ten times as many observations as variables, before variable selection takes place.

Any analysis should consider an explorative investigation of the available explanatory variables, leading at times to consider suitable transformations of some or all variables.

Graphical scrutiny of the explanatory variables may lead to the omission of some variables.

Beware of spurious relationships: two or more variables seem to be related, due to either coincidence or the presence of a third, unseen variable.

Interaction effects between numerical variables (not factors) are often modeled by including pairwise products, namely $x_1 \cdot x_2$ as well as x_1 and x_2.

Multivariate techniques, such as **Principal Components Analysis**, are sometimes useful for selecting low-dimensional combinations of several explanatory variables.

These small number of combinations together account for most of the variation in the explanatory variables, thus reducing the dimension of the problem.

Other methods, such as the **lasso**, the **least angle regression** or the **boosting**, allow for semi-automatic variable selection.

Caution should be used with automatic selection techniques, such as stepwise regression and subset regression.

5.16 Multicollinearity

Some explanatory variables are linearly related to another variable, or to combinations of other explanatory variables. This is known as **multicollinearity**, and it is very common with observational data.

Multicollinearity implies that there are redundant variables. It alters the relation of the explanatory variables with the response. In extreme cases, it may lead to poorly estimated coefficients.

A measure for quantifying the severity of multicollinearity is the **variance inflation factor (VIF)**. For an explanatory variable x_j

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 coefficient for the regression model having x_j as the response variable, and all the other explanatory variables as regressors.

Thresholds usually adopted for VIF are 4 or 5, that suggest some multicollinearity. With a VIF > 10 multicollinearity is severe, and the model coefficients will be poorly estimated.

5.17 Remedies for multicollinearity

Careful initial choice of variables, based on background information and careful scrutiny of exploratory plots, often will prevent the problem.

Dropping one or more explanatory variables is the main route to address multicollinearity.

An alternative route is to obtain a combination of the original explanatory variables that summarizes them without losing too much information: this is performed by the aforementioned PCA method.

There are also some modern methods, such as **ridge regression** or the **lasso**, that are not affected by multicollinearity.

5.17.1 Example: coxite

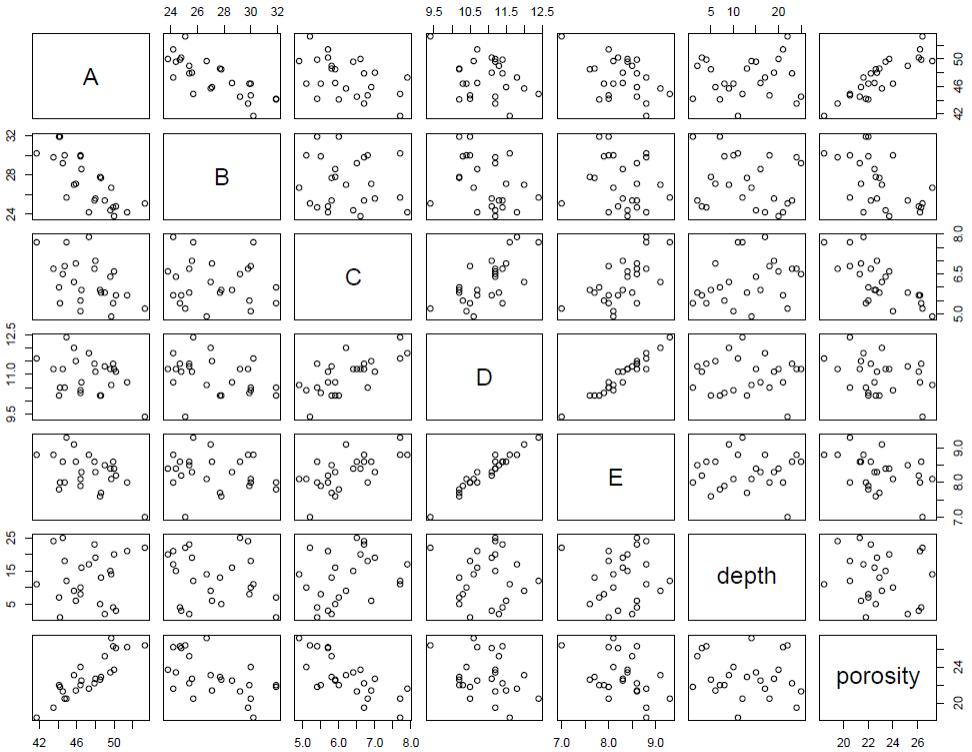
The data set coxite gives the mineral compositions of $n = 25$ rock specimens of coxite type.

Each composition consists of the percentage by weight of five minerals, namely, albite A, blandite B, cornite C, daubite D, endite E.

Indeed, the recorded depth (m) of location of each specimen and the porosity (the percentage of void space that the specimen contains) are also provided. Note that the percentages of the five minerals sum to 100.

The aim of regression analysis is to explain the response variable porosity as a function of mineral composition and depth.

The scatterplot matrix shows that D and E are strongly linearly related.



Fitting the model with all six explanatory variables gives a coefficient for E equal to zero, since the five percentage sum to 100 and then E adds no additional information.

None of the individual coefficients is significantly different from zero (all the p-values are greater than 0.3).

However, the R^2 measures are high (0.935 and 0.919, for the adjusted version), and the F test of global significance of all terms is significant (the p-value is $1.18 \cdot 10^{-10}$)

These are clear symptoms of multicollinearity: indeed, omitting E (or one of A, B, C, D), the VIF values can be calculated and they are very large:

A	B	C	D	depth
2717.8	2485	192.59	566.14	3.4166

Thus, it is unsurprising that none of the individual coefficients can be estimated meaningfully.

It is reasonable to try a model that uses those variables that, individually, correlate most strongly with porosity. Here are the correlations:

A	B	C	D	E	depth
0.869	-0.551	-0.723	-0.320	-0.408	-0.147

The model with A, B and C as explanatory variables improves the previous one, but the coefficient for A is not significantly different from zero and yet the VIFs are all larger than 4:

A	B	C
10.9360	8.5924	4.5551

The model with only B and C as explanatory variables is much better and it passes all the diagnostic checks. Indeed, the VIFs are both around 1.

Furthermore, the AIC for the full model is 56.50, whereas it is 52.47 for the final model.

5.18 Factors as explanatory variables

The model matrix X is fundamental to all calculations for a linear model. It carries the information needed to calculate the fitted values that correspond to any particular choice of coefficients.

The columns of X contain the observed values of the (numeric) explanatory variables, perhaps after transformation, whereas the first column usually corresponds to the model intercept.

However, explanatory variables are not always numeric, and actually factors are very common in many applied fields.

Factors can be included in a model in a straightforward way, and it is possible to fit their effect on the response along with the effect of numerical variables.

ANOVA models are just a special case of linear regression models where all the explanatory variables are factors.

5.19 Two-way ANOVA

One-way ANOVA models have been previously introduced in the context of simple linear regression models, to deal with the situation where there is only one factor as explanatory variable.

However, there might be more than one factor influencing the response variable, like in designed experiments.

Two-way ANOVA is suitable with two factors, while **multi-way ANOVA** accounts for an arbitrary number of factors. With more than one factor ANOVA becomes more complex, as there may be **interaction** between the factors.

Two-way ANOVA not only aims at assessing the main effect of each (categorical) variable on the response but also the potential effect due to the interaction between them.

The methodology proceeds exactly as in the one-way case, with suitable extensions for considering the presence of two factors.

5.19.1 Example: warp breaks

The data set gives, as response variable, the number of warp breaks per a fixed length of yarn during weaving.

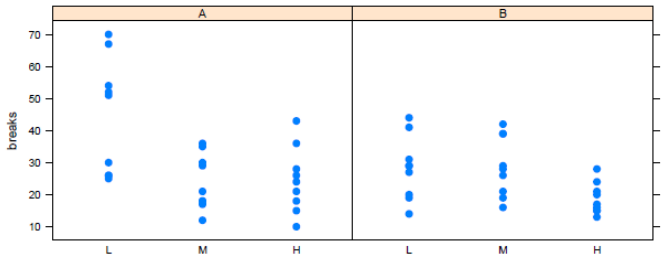
There are two experimental factors: the type of wool and the level of tension, with 2(A or B) and 3 levels (L, M, H) respectively: there are 9 replications for each of the 6 combinations of factors levels.

The total sample size is then $n = 2 \times 3 \times 9 = 54$.

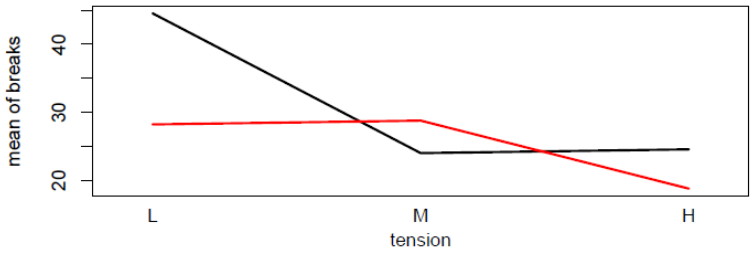
Also in this case, the data are balanced; unbalanced data are better dealt with by techniques based on linear regression models.

		replicate								
wool	tension	1	2	3	4	5	6	7	8	9
A	L	26	30	54	25	70	52	51	26	67
	M	18	21	29	17	12	18	35	30	36
	H	36	21	24	18	10	43	28	15	26
B	L	27	14	29	19	29	31	41	20	44
	M	42	26	19	16	39	28	21	39	29
	H	20	21	24	17	13	15	15	16	28

A **conditional plot** shows that the pattern of the response is not the same for the two levels of wool:



A similar result can be obtained with an interaction plot, where the means (or another summary) for each combination of two factors are displayed: **level A** and level B of the factor wool:



5.20 Statistical model for two-way ANOVA

Without interaction between the two factors: each factor has the same effect on the mean response, regardless of the level of the other.

The model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where $i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n$, and:

- i identifies the treatment level for the first factor;
- j identifies the treatment level for the second factor;
- k identifies the observation;
- μ is the **general mean**;
- α_i is the **main effect for the first factor**: the deviation from the general mean μ when the first factor is equal to the i-th category;
- β_j is the **main effect for the second factor**: the deviation from μ when the second factor is equal to the j-th category;
- ε_{ijk} i.i.d. $N(0, \sigma^2)$ distributed random errors.

To test the main effect of the first factor:

$$\begin{aligned} H_0: &\quad \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \\ H_1: &\quad \alpha_i \neq 0 \text{ for at least one } i \end{aligned}$$

and similarly for the other factor.

The test is performed exactly like for one-way ANOVA, by considering suitable sums of squares and d.f., that define an F test statistic.

The results of the analysis can be trusted as long as the model assumptions are reasonably satisfied.

Post-hoc analysis can also be performed, though they require more care than in the one-way case.

When possible interaction is investigated, another set of coefficients γ_{ij} is introduced in the model, that becomes:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where γ_{ij} describes the **interaction effect for the two factors**: the deviation from the general mean μ due to the interaction of the i-th category of the first factor and the j-th category of the second factor.

The test on the presence of interaction can be carried out similarly to the tests on main effects.

A hierarchical order should be followed in performing the various tests:

- first, the presence of interaction (if plausible) should be assessed;
- in case the data support the presence of an interaction effect, it makes little sense to investigate about the presence of main effects, as both the both the two factors influence the mean response;
- in case the interaction effect is not signficant, the rest on the two main effects can be performed.

5.20.1 Example: warp breaks

For the data set on warp breaks, the transformed response $sqrt(breaks)$ is considered instead of the original observations, as the conditional plot shows some evidence of non-constant variance.

Indeed, ANOVA assumes a constant variance for the observations, so in case this is doubtful a **variance-stabilizing transformation**, such as the squared root or the log, should be investigated.

The p-value for the F test on the interaction effect is $p = 0.031$, showing an appreciable interaction effect, though not very large.

Moreover, it is consequently stated that the two main effects are present as-well, since both the two factors influence the mean of the response variable warp breaks.

5.21 Regression models with dummy variables

More generally, in the context of multiple linear regression model, it is possible deal with the presence of factor explanatory variables.

To include a factor in a model it is necessary to code its levels, and one possibility is to use **dummy variables** (regressors that assume only two values, usually, zero and one).

The rule for coding a factor is quite simple. Coding a factor with h levels (categories) requires the usage of $(h - 1)$ dummy variables.

Consider, for example, the study of the relationship between personal income Y and education level X (a factor with $h = 3$ levels: middle school, high school and university).

Two dummy variables x_1 and x_2 are needed, so that

factor level	x_1	x_2
middle school	0	0
high school	1	0
university	0	1

The dummy variables used to represent the effect of different factor levels can then be included in the regression model.

In the example,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

which corresponds to specify the following expressions for the mean of Y :

- β_0 for middle-school level;
- $\beta_0 + \beta_1$ for high-school level;
- $\beta_0 + \beta_2$ for university level.

The coefficients used to represent the effect of a given factor represent the **main effect** of that factor.

Such coefficients express the differential effect on the mean response that can be attributed to the different levels of that factor.

One of the factor levels is set up as a baseline or reference, with the effects of other levels then measured from the baseline.

5.22 Models with both factors and numerical explanatory variables

Often in applications there are both categorical and numerical explanatory variables.

By using a suitable factor coding for categorical variables, it is possible to include both categorical and numerical explanatory variables in a model specification.

Regression models with both factors and numerical regressors are also known as **analysis of covariance models**, and they essentially amount to fitting multiple lines.

All the relevant inferential techniques are those already introduced for multiple regression models. Different models are compared by F tests, summarized by ANOVA-type results.

5.23 Fitting multiple lines

In the example concerning income and education level, the numerical explanatory variable x_{i3}, summarizing the number of years spent at work, is also considered.

The model including such variable and the education level is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

This is apparently similar to the model with more than one numerical predictor, but there are some important differences.

In this case, model fitting actually concerns **three parallel simple regression lines**, with intercepts changing with the level of education.

Including also an interaction term between education level and years spent at work requires the fit of **three different simple regression lines** for the three groups of units with different level of education.

5.23.1 Example: regression on age and gender

A regression model for a response variable Y related to the health status, such as cholesterol or blood pressure, is defined.

Potential predictors are the numerical variable age x and a binary factor variable gender, coded by a dummy variable g (assuming 1 for female).

A model with the main effects and the interaction effect is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 x_i g_i + \varepsilon_i$$

This model assumes that Y depends linearly on age for males, with the regression line:

$$y = \beta_0 + \beta_1 x$$

and Y depends linearly on age for females, with the regression line:

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x$$

Then **two different simple regression lines** for the groups of males and females are defined.

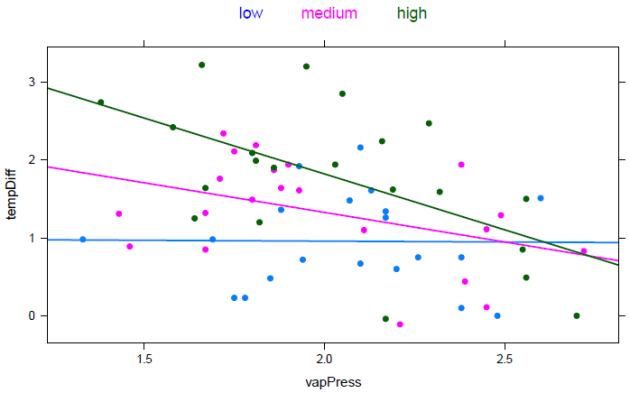
The following hypotheses may be of potential interest, and can be analyzed by means of suitable t tests or F tests.

Null/Alternative hypothesis	Translation
Relation between age and y does not depend on gender (Relation between gender and y does not depend on age)	$H_0 : \beta_3 = 0$
Gender influences the relation between y and age; (Age influences the relation between y and gender)	$H_1 : \beta_3 \neq 0$
y does not depend on age	$H_0 : \beta_1 = \beta_3 = 0$
y depends on age	$H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$
y does not depend on gender	$H_0 : \beta_2 = \beta_3 = 0$
y depends on gender	$H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$
y does not depend on age and gender	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
y depends on age or gender	$H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$

5.23.2 Example: leaf and air temperature

The data set includes $n = 62$ environmental measures regarding the vapor pressure (vapPress) and the difference between leaf and air temperature (tempDiff), for three different levels (low, medium, high) of carbon dioxide (C02level).

The scatterplot of tempDiff vs vapPress suggests that there may be three different regression lines, for the three different levels of C02level, but this has to be confirmed more formally.



Denoting by x the numerical explanatory variable vapPress, and by z_1, z_2 the two dummies for the factor c02level, four different models for the response y (tempDiff) may be defined:

M1 (constant response): $y = \beta_0 + \varepsilon$
M2 (single line): $y = \beta_0 + \beta_1 x + \varepsilon$
M3 (3 parallel lines): $y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \varepsilon$
M4 (3 separate lines) : $y = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x z_1 + \beta_5 x z_2 + \varepsilon$

Models M3 and M4 involve the factor C02level, and the level low is considered as the baseline.

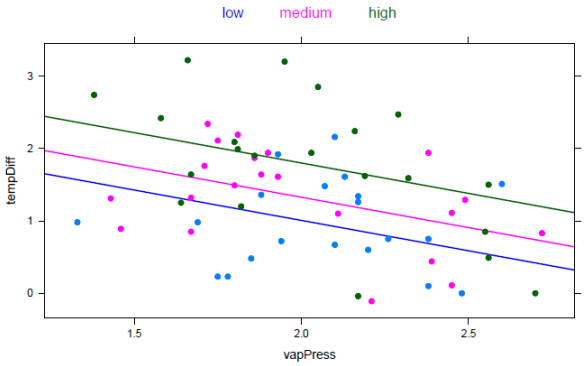
In model M3, the slope is β_1 and the three possible intercepts are β_0, $\beta_0 + \beta_2$ and $\beta_0 + \beta_3$ for the levels low, medium and high, respectively.

In model M4, the intercepts are as in M3 and the three possible slopes are $\beta_1, \beta_1 + \beta_4$ and $\beta_1 + \beta_5$ for the levels low, medium and high, respectively.

The analysis of variance table is helpful in making a choice between these model. The sequential analysis, using suitable F tests, on the four nested models in the increasing order, namely M1, M2, M3, M4, gives the p-values $0.0014, 0.0019$, and 0.1112.

The analysis of variance results suggests use of the parallel line model M3, since the reduction in the mean square from M3 to M4 has a p-value equal to 0.1112.

The diagnostic checking of M3 does not show any particular problem, and the final model fit is then:



5.24 Non-Gaussian response

Regression models may be extended to account for non-Gaussian response variables.

In particular, the response might admit binary outcomes, that is only two values (usually coded as 0 and 1) described by a Bernoulli distribution, or more generally binomial distributed outcomes.

In this case, using a linear regression model it is not attractive since predictions for the mean response, which corresponds to an outcome probability, can give off-scale values below 0 or above 1.

It makes better sense to model the probabilities on a transformed scale; this is what is done in **logistic regression analysis**.

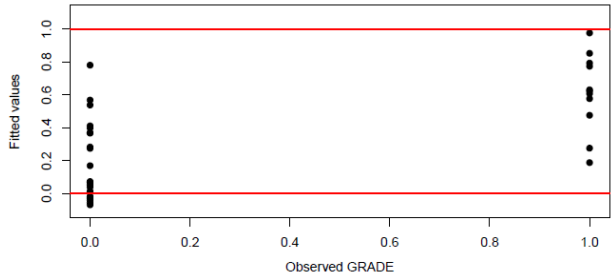
Logistic regression models belongs to the class of **generalized linear models**.

These models are characterized by their specific response distribution and by a link function, which transfers the mean value to a scale in which the relation to the explanatory variables is linear and additive.

5.24.1 Example: teaching program

Data on the effectiveness of a teaching program. For $n = 32$ students, four variables are observed: GPA (grade point average for the period), TUCE (test score on economics test), PSI (participation in program: yes, 1 , and no, 0), GRADE (grade increase, 1 , or decrease, 0 , indicator).

To measure the effect of the explanatory variables on the response GRADE, it is tempting to fit a multiple linear regression model so that $E(Y_i) = \beta_0 + \beta_1 \cdot \text{GPA}_i + \beta_2 \cdot \text{TUCE}_i + \beta_3 \cdot \text{PSI}_i$



Some fitted values are negative: it is unacceptable as $E(Y_i) = \text{P}(Y_i = 1)$.

5.25 Generalized linear models

Generalized Linear Models (GLMs) extend linear regression models so that:

- a more general form of expression for the mean response is allowed, using suitable link functions;
- various types of distributions for the response can be considered.

There naturally is quite a large overlap with the material on linear Gaussian models, but there are also some special issues concerning the specific response distribution and link function.

The main application of GLMs is for modeling **proportions** (binomial data, including Bernoulli data) or **counts** (Poisson data).

This class of models gives a unified theoretical and computational approach to models that had previously been treated as distinct.

Here the focus will be on binomial data, but similar considerations apply to count data and other GLMs. Logistic regression models are perhaps the most widely used GLMs.

In general, GLMs allow a transformation $f(\cdot)$ to the left-hand side of the regression equation. More precisely, instead of assuming:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

a linear model for $f\{E(Y_i)\}$ is specified, namely:

$$f\{E(Y_i)\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

where $f(\cdot)$ is a function usually called the **link function**, whereas $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is, as usual, the **linear predictor**.

The link function transforms from the scale of Y to the scale of the linear predictor. In case of non-Gaussian response, there is no variance parameter. Extensions with a flexible specification of the variance are possible.

Least squares estimation cannot be used for parameter estimation in GLMs. Estimation methods commonly used include maximum likelihood estimation and Bayesian methods.

Maximizing the likelihood is equivalent to minimizing the deviance, which has a role similar to the residual sum of squares.

5.26 Logistic regression: The analysis of binary data

For binary (Bernoulli) data, it is not reasonable that the expected proportion will be a linear function of the explanatory variables. Then, a suitable link function, which goes from [0,1] to the real line, can be defined.

The most commonly used one works on the log odds scale and it corresponds to the **logit (logistic) link** $f(u) = \log(u/(1-u))$.

Odds are common for bookmakers in betting: if p is a probability, the corresponding odds and log(odds) are, respectively,

$$\text{odds } = \frac{p}{1-p}, \quad \log(\text{ odds }) = \log(p/(1-p)) = \log(p) - \log(1-p)$$

Furthermore,

$$p = \frac{e^{\log(\text{ odds })}}{1 + e^{\log(\text{ odds })}}$$

GLMs for binary data that employ the logit link go under the name of **logistic (multiple) regression models**; they allow to model the log odds as a linear function of suitable explanatory variables.

Other choices do exist: function $f(u) = \log(\log(1-u))$, which has a connection to survival analysis models, or probit function $f(u) = \Phi^{-1}(u)$ (the quantile function of the normal distribution).

Similarities and differences between linear regression and logistic regression are summarized in the following table

Linear regression	Logistic regression
Estimates, std. errors, t -values	Estimates, std. errors, z -values
Sum of squares	Deviance
Residual standard error	–
Fit models by minimizing the residual sum of squares	Fit models by maximizing the log-lik. (minimizing the deviance)
Select models with smaller AIC	Select models with smaller AIC
Compare nested models via F tests	Compare nested models via χ^2 tests
Full set of diagnostic plots	Some diagnostic plots
Partial residual plots	Plots of explanatory variable contributions
R^2 and adjusted R^2	Predictive accuracy

5.26.1 Example: teaching program

Data set on the effectiveness of a teaching program with regard to $n = 32$ students. The 2×2 contingency table for the binary response GRADE (grade increase) and the factor predictor PSI (participation in program) is:

		GRADE	
		0	1
PSI	0	15	3
	1	6	8

The observed proportion of GRADE=1 is $3/18 = 0.167$, for students with PSI $= 0$, and $8/14 = 0.571$, for students with PSI $= 1$, that is:

$$\log(\text{ odds }) = \log(0.167/0.833) = -1.609 \quad \text{for PSI} = 0$$
$$\log(\text{ odds }) = \log(0.571/0.428) = 0.288 \quad \text{for PSI} = 1$$

and the corresponding logistic regression model can be written as:

$$\log(\text{ odds }) = \beta_0 + \beta_1 \text{PSI} + \varepsilon$$

where odds $= P(\text{GRADE} = 1)/P(\text{GRADE} = 0)$ and the effect due to PSI is coded using a single dummy variable.

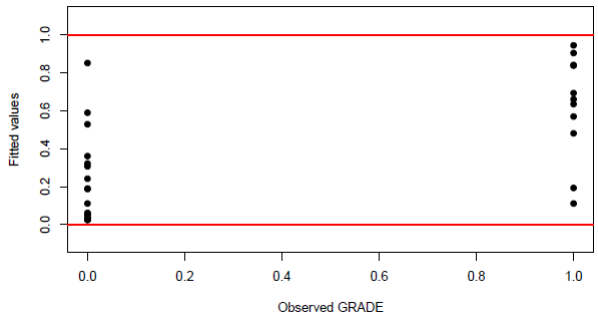
Fitting the model gives the estimates $\widehat{\beta}_0 = -1.61(0.632)$ and $\widehat{\beta}_1 = 1.90\ (0.832)$. According to the p-value for the z test, the actual significance of PSI is not so effective.

Effects due to GPA and TUCE can then be introduced by considering a logistic multiple regression model with both numerical and factor predictors.

Not all the predictors induce a significant effect. The fitted (values) probabilities from logistic regression are in [0,1], as:

$$\widehat{P}\,(\text{GRADE}_i = 1) = \frac{e^{\log\left(\widehat{\text{odds}}_i\right)}}{1 + e^{\log\left(\widehat{\text{odds}}_i\right)}}$$

with $\log\left(\widehat{\text{odds}}_{\ i}\right) = \widehat{\beta}_0 + \widehat{\beta}_1\ \text{PSI}_{\ i} + \widehat{\beta}_2\ \text{TUCE}_{\ i} + \widehat{\beta}_3\ \text{GPA}_{\ i}$.



5.27 Predictive accuracy

For binary data, it makes sense to compare the observed data y_i with the predictions obtained from the model, that is

$$\widehat{y}_i = \begin{cases} 1 & \text{if } \widehat{P}\,(\text{GRADE}_i = 1) \geq 0.5 \\ 0 & \text{if } \widehat{P}\,(\text{GRADE}_i = 1) < 0.5 \end{cases}$$

Predicted and observed values can be summarized in a 2×2 table. For the teaching program example, it corresponds to

		Observed values	
		0	1
Predicted values	0	18	3
	1	3	8

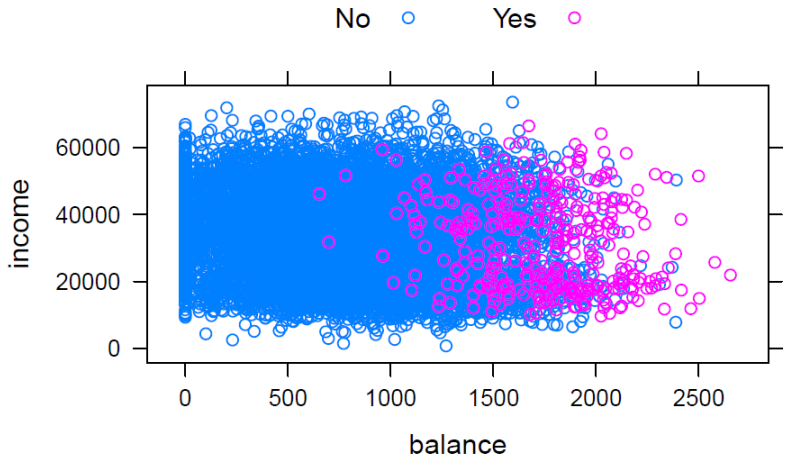
An evaluation of the **predictive accuracy** is given by the percentage of correct classifications; in this case, it is equal to $26/32 = 0.812$. As usual, when the same data are used twice, the performance of a given model is over-estimated. A more correct assessment uses a **cross-validation procedure**. For the current example, it returns a value around 0.688.

5.27.1 Example: credit card

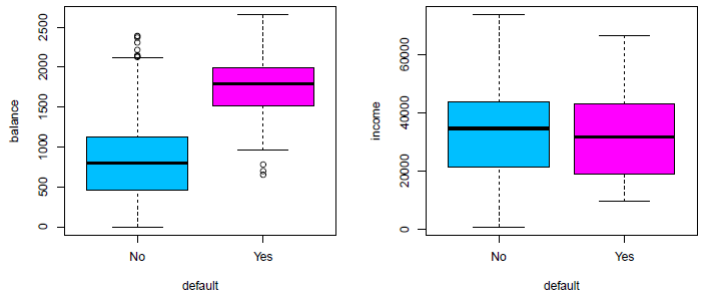
Data set on the defaults on credit card payments. For $n = 10000$ customers, four variables are observed: DEFAULT in a given period (STUDENT (yes), INCOME (annual income), BALANCE (monthly credit card balance).

To describe the effect of the explanatory variables STUDENT, INCOME and BALANCE on the binary response DEFAULT and to predict whether an individual will default.

Only about 3% of people in the data set actually default and individuals who defaulted tended to have higher credit card balances than those who did not, as shown in the following plot:



This is confirmed by the boxplots of BALANCE (left) and of INCOME (right) as a function of DEFAULT status:



Then a sensible model for DEFAULT is the (simple) logistic regression model

$$\log(\text{ odds }) = \beta_0 + \beta_1 \text{ BALANCE } + \varepsilon$$

Fitting the model gives the estimate $\widehat{\beta}_1 = 0.0055(0.0002)$. According to the p-value for the z test, the actual significance of BALANCE is effective.

An increase in BALANCE is associated with an increase in the probability of DEFAULT: a one-unit increase in BALANCE increases the log(odds) of DEFAULT by 0.0055 units.

Once the coefficients are estimated, it is possible to make predictions on DEFAULT by computing the probability of default for an individual with a given credit card balance x

$$\widehat{P}(\text{DEFAULT} = 1 \mid x) = \frac{\exp\{-10.6513 + 0.0055 \cdot x\}}{1 + \exp\{-10.6513 + 0.0055 \cdot x\}}$$

The predicted probability of default for individuals with balances of \$1000 and \$2000 are 0.00576 and 0.586, respectively.

An alternative (simple) logistic regression model can be considered using, as explanatory variable, the qualitative variable STUDENT, coded as a dummy variable

$$\log(\text{ odds }) = \beta_0 + \beta_1 \text{ STUDENT } + \varepsilon$$

The estimated coefficient associate with the dummy variable is positive is $\widehat{\beta}_1 = 0.4049(0.1150)$ and the corresponding p-value is statistically significant. since $\widehat{\beta}_1 > 0$, students tend to have higher default probabilities than non-students: 0.0431 and 0.0292, respectively.

The joint effects of INCOME, BALANCE and STUDENT can be described using the following multiple logistic regression model

$$\log(\text{ odds }) = \beta_0 + \beta_1 \text{ INCOME } + \beta_2 \text{ BALANCE } + \beta_3 \text{ STUDENT } + \varepsilon$$

The estimated coefficients are $\widehat{\beta}_1 = 3.033 \cdot 10^{-6}$($p$-value $= 0.7115$) $\widehat{\beta}_2 = 0.0057$ (p-value $< 2 \cdot 10^{-16}$) and $\widehat{\beta}_3 = -0.6468$($p$-value $= 0.0062$) since $\widehat{\beta}_3 < 0$, for a fixed value of INCOME and BALANCE, students are less likely to default than non-students. This seems in contradiction with the previous conclusion.

Although the student default rate is usually below that of the non-student default rate for every value of BALANCE, the overall student default rate is higher (confounding phenomenon).

STUDENT and BALANCE are slightly correlated. Students tend to hold higher levels of credit card balances, which is associated with higher default rates.

A student is riskier than a non-student if no other information is available. However, that student is less risky than a non-student with the same credit card balance.

5.27.2 Example: UCB admissions

Data set on the admission at the University of California at Berkeley in the fall of 1973 .

Three categorical variables: admission Admit (Admitted/Rejected), Gender (Male/Female) and department Dept (A, B, C, D, E, F). The data set is related to the well-known Berkeley gender bias case: female appear discriminated, although no single department is strongly biased against woman.

The aggregate data and the department level data tell opposite stories about gender bias. Most departments have a slight female bias, while the difference on overall application and admission rates causes the aggregate bias to point in the other direction.

A logistic regression model is defined in order to describe the probability of admission. The factor explanatory variables Dept and Gender are set in this order; a potential interaction effect is also considered.

It is important, for present purposes, to fit Dept, thus adjusting for different admission rates in different departments, before fitting Gender. The estimated coefficients are given below:

Coefficients	Estimate	SE	p-value
Intercept	0.4921	0.0717	$6.94 \cdot 10^{-12}$
DeptB	0.0416	0.1132	0.71304
DeptC	-1.0276	0.1355	$3.34 \cdot 10^{-14}$
DeptD	-1.1961	0.1264	$< 2 \cdot 10^{-16}$
DeptE	-1.4491	0.1768	$2.49 \cdot 10^{-16}$
DeptF	-3.2619	0.2312	$< 2 \cdot 10^{-16}$
GenderFemale	1.0521	0.2627	$6.21 \cdot 10^{-5}$
DeptB:GenderFemale	-0.8321	0.5104	0.1031
DeptC:GenderFemale	-1.1770	0.2996	$8.53 \cdot 10^{-5}$
DeptD:GenderFemale	-0.9701	0.3026	0.0014
DeptE:GenderFemale	-1.2523	0.3303	0.0002
DeptF:GenderFemale	-0.8632	0.4027	0.0321

Comparison of the nested models using sequential χ^2 tests shows that there is a clear effect of Dept on the admission rate, while there is no detectable main effect of Gender.

The significant interaction term suggests that there are departmentspecific gender biases, which average out to reduce the main effect of Gender to close to zero.

Concerning the individual model coefficients:

- the first six coefficients relate to overall admission rates, for males, in the six departments;
- the strongly significant positive coefficient for GenderFemale indicates that log(odds) is increased by 1.05, in department A, for females relative to males;
- in departments C and E the log(odds) is reduced for females, relative to males.

6 Predictive and classification methods

6.1 Introduction

Predictive modeling is a "process by which a model is created or chosen to try to best predict the probability of an outcome".

The model or the mathematical tool which is developed is considered for giving accurate prediction.

For example, insurance companies aim at predicting the risk of potential auto, health and life policy holders. This is crucial in order to determine if an individual will receive a policy and, if so, at what premium.

Governments use predictive models for evaluating potential risks, with the aim of protecting their citizens. For example, biometric models for identifying terror suspects and models for fraud detection.

Internet companies apply predictive models to guide consumers towards more satisfying products or more profitable investments.

Although predictive models aim at indicating more satisfying products, better medical treatments and more profitable investments, they may generate inaccurate predictions and give wrong answers.

There are a number of reasons why predictive models fail. The main culprits are:

- inadequate pre-processing of the data;
- inadequate model selection and validation;
- unjustified extrapolation (application of the model to data outside the range of the available observations);
- over-fitting the model to the existing data.

It is surely important to specify reliable and trustworthy predictive models, however the accuracy of our prediction will be affected by an irreducible error component.

This unavoidable error term is related, for example, to the fact that relevant predictor variables may be missed, that there are unmeasurable, and then not exploitable, variables (such as those related to personal human behavior) and that prediction are usually constrained by our present and past knowledge.

6.2 Prediction versus inference

Suppose that we observe a quantitative response Y and $p \geq 1$ different explanatory variables (predictors) $X = (X_1, \ldots, X_p)$ and that the following general model is defined:

$$Y = f(X) + \varepsilon$$

The fixed, unknown function $f(\cdot)$ represents the systematic information on Y provided by X and ε is a zero-mean random error term. Regression models fall into this framework.

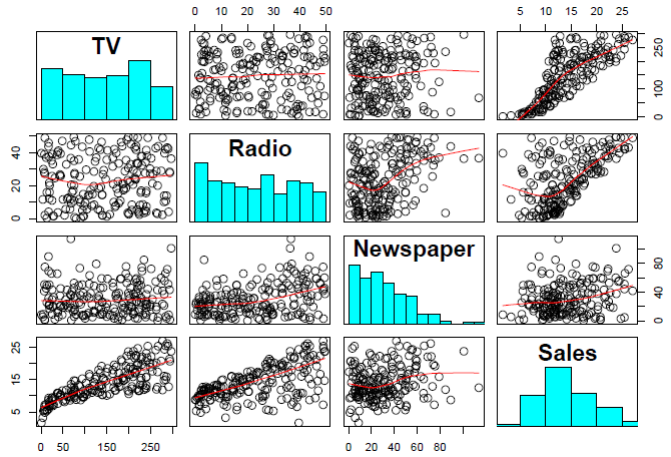
Inference: understand the relation between Y and X and, in particular, how Y changes as a function of $X_1, \ldots, X_p$; the exact form of f is usually needed and it is essential to obtain an estimate $\widehat{f}$ based on the available data.

Prediction: given a set of inputs for X, the aim is to predict the associated value for Y using a predictor $\hat{Y} = \widehat{f}(X)$ or a prediction interval; $\widehat{f}$ may be treated as a black *box*.

Applications may fall into the prediction setting, the inference setting, or a combination of the two.

6.2.1 Example: advertising data

Data set on sales of a certain product (response variable Y) along with the advertising budgets for three different media, TV, radio, newspaper (predictor variables X_1, X_2, X_3); values in thousands related to $n = 200$ different markets.



A multiple linear regression model suggest that the effect of newspaper on sales is not statistically significant:

Coefficients	Estimate	SE	p-value
Intercept	2.9389	0.3119	< 0.0001
TV	0.0458	0.0014	< 0.0001
radio	0.1885	0.0086	< 0.0001
newspaper	-0.0010	0.0059	0.8599
	$s^2 = 2.841$	$R^2_{\text{adj}} = 0.896$	

In the **inference framework** one may be interested, for example, in finding which media contributes significantly to sales, which media generates the biggest boost in sales, what is the increase in sales associated with a given increase in TV advertising.

In the **prediction framework** one may be interested, for example, in predicting the amount of sales given a fixed budget for the three media.

6.3 Measuring the quality of fit

By considering the available data (training observations) $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$, a suitable model may be fitted, obtaining the estimate $\widehat{f}$.

In order to evaluate the performance of the model, it can be useful to evaluate how well the predicted response values $\widehat{y}_i = \widehat{f}(\mathbf{x}_i)$ are close to the true response values $y_i, i = 1, \ldots, n$.

A common measure, used in the regression setting, is the **training mean squared error (MSE)** given by:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{f}(\mathbf{x}_i) \right)^2$$

The MSE is computed using the training observations, which have been previously employed for model fitting; therefore, the fitting accuracy is in fact measured instead of the predictive accuracy.

Since the same data are used twice, the training MSE would give an overoptimistic predictive assessment of the model.

In order to evaluate the predictive performance of a model, it is more convenient to consider how well the predicted response values $\widehat{y}_{0j} = \widehat{f}(\mathrm{x}_{0j})$ are close to the true response values $y_{0j}, j = 1, \ldots, m$ with regard to previously unseen observations.

While the model is fitted using the training observations, the prediction accuracy is now evaluated by considering the **test observations** $(\mathrm{x}_{01}, y_{01}), \ldots, (\mathrm{x}_{0m}, y_{0m})$, not used to train the statistical model.

Thus, a measure of predictive fit is given by the **test MSE:**

$$\text{testMSE} = \frac{1}{m} \sum_{j=1}^{m} \left(y_{0j} - \widehat{f}(\mathbf{x}_{0j}) \right)^2$$

where $\widehat{f}$ is estimated using the training observations.

In some settings, a test data set can be available; for example, when the original data set is sufficiently large.

Whenever, as usual, no test data are available, methods for estimating test MSE using the training data can be considered. One important method in **cross-validation**.

If one selects a statistical model by minimizing the training MSE there is no guarantee that the lowest test MSE is achieved.

The training MSE is usually smaller than the test MSE. Furthermore, while the training MSE declines as model flexibility increases, the test MSE initially declines and then start to increase again.

This phenomenon is known as **overfitting**: the model focuses exaggeratedly on patterns that are just caused by randomness and it misses the true properties of the unknown function f.

The test MSE may be viewed as an estimate for the **expected test MSE** for the future random response Y_0, with a given value x_0:

$$E\left[\left(Y_0 - \widehat{Y}_0 \right)^2 \right] = V\left(\widehat{f}(\mathbf{x}_0) \right) + \left[\text{Bias}\left(\widehat{f}(\mathbf{x}_0) \right) \right]^2 + V(\varepsilon)$$

where $\widehat{Y}_0 = \widehat{f}(\mathrm{x}_0)$ is the point predictor based on the fitted model.

This value can never lie below $V(\varepsilon)$, the **irreducible error term**. The minimum is achieved for models that simultaneously have low variance and low bias.

The **bias-variance trade-off**: usually, as more flexible methods are considered, the variance will increase and the bias will decrease.

6.4 Regression vs classification problems

Variables can be characterized as either quantitative or qualitative (also known as categorical).

Predictive problems with a quantitative response are usually called **regression problems**, while those involving a qualitative response are often referred to as **classification problems**.

The distinction is not always sharp, since logistic regression, which is often used as classification method, may be viewed as an extension of linear regression models with the aim of modeling probabilities on a transformed scale. The present section focuses on prediction using (multiple) linear regression models. Some notions, already discussed in the previous sections, are reviewed by studying two data sets. The next section is devoted to the analysis of classification problems.

6.4.1 Example: automobile bodily injury claims

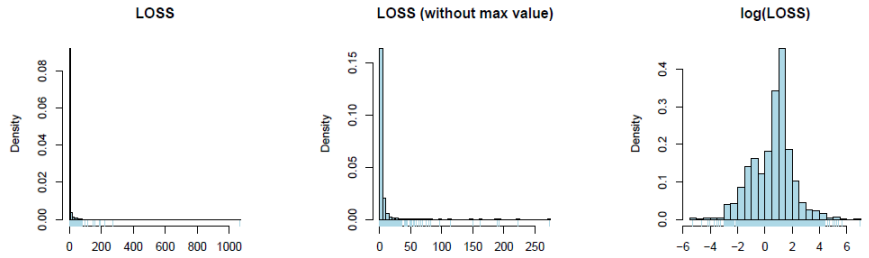
Data from the Insurance Research Council, US, collected in 2002 and regarding automobile bodily injury claims. Observations on the variables:

- ATTORNEY: whether the claimant is represented by an attorney;
- CLMSEX: claimant gender (male, female);
- MARITAL: claimant marital status (married M, single S, widowed W, divorced D);
- CLMINSUR: whether or not the driver of the claimant's vehicle was uninsured;
- SEATBELT: whether or not the claimant was wearing a seatbelt/child restraint;
- CLMAGE: claimant's age;
- AGECLASS: claimant's age split into five classes: $(-18]$, $(18, 26]$, $(26, 36]$, $(36, 47]$, $(47+)$;
- LOSS: claimant's total economic loss (in thousands).

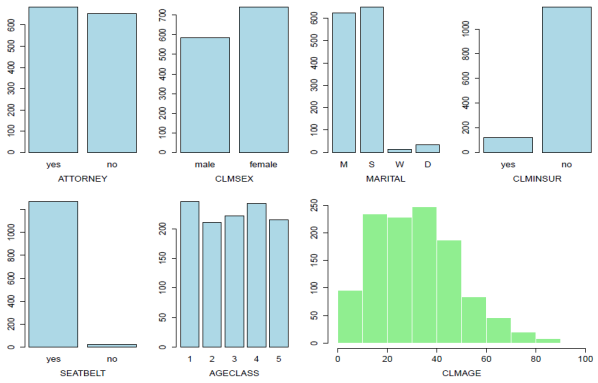
It is of interest to build a statistical model for predicting claim amounts in future policies, based on a sample of claim amounts of past policies.

The severity refers to the amount of claim. The response variable is quantitative and it is given by the claimant's total economic loss.

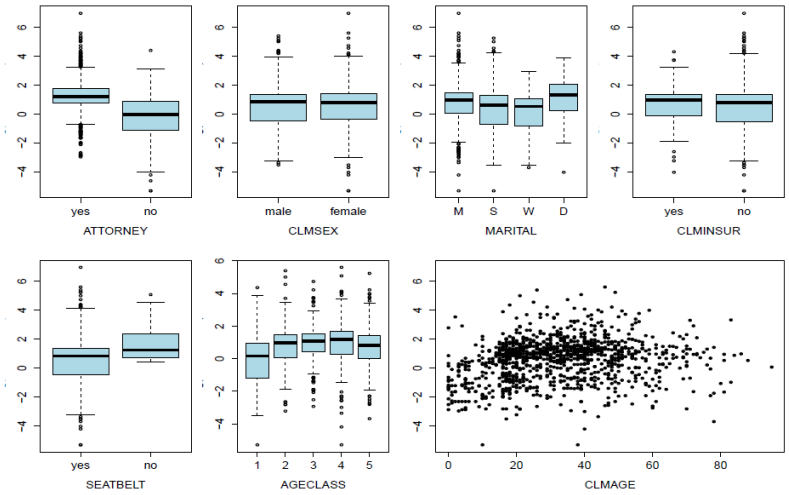
The logarithmic scale is chosen, as motivated by the following histograms, so that the response variable is logLOSS.



Information about policies enables a marginal description of the explanatory variables:



A further analysis suggests that a significant relationship exists between logLOSS and some explanatory variables:



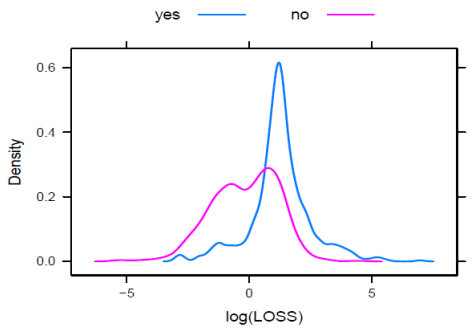
Being represented by an attorney seems to be important, as shown by considering the density estimate of logLOSS in case of ATTORNEY=yes and ATTORNEY=no

Also other variables may matter, such as SEATBELT, MARITAL and AGECLASS. On the contrary CLMSEX and CLMINSUR seem not to have an effect. The effect of CLMAGE is not clear.

The joint effects of ATTORNEY (qualitative variable coded as a dummy variable) and CLMAGE on the response logLOSS can be described using the following multiple linear regression model:

$$\text{logLOSS} = \beta_0 + \beta_1 \text{ ATTORNEY } + \beta_2 \text{ CLMAGE } + \varepsilon$$

	Estimate	SE	p-value
Intercept	0.7376	0.0851	< 0.0001
ATTORNEYno	-1.3699	0.0729	< 0.0001
CLMAGE	0.0160	0.0021	< 0.0001
	$n - p = 1148$	$s = 1.23$	$R^2_{\text{adj}} = 0.259$



Both the coefficients are strongly significant, pointing to a relevant effect of both ATTORNEY and CLMAGE on the mean response.

The dummy variable flags those observations that have the level of ATTORNEY equal to no. Since its estimated coefficient is negative, subjects without an attorney have a smaller mean response.

Note that 189 observations were deleted due to missingness.

It is easy to estimate the mean of the logLOSS and then to transform back the results on the scale of the original LOSS variable.

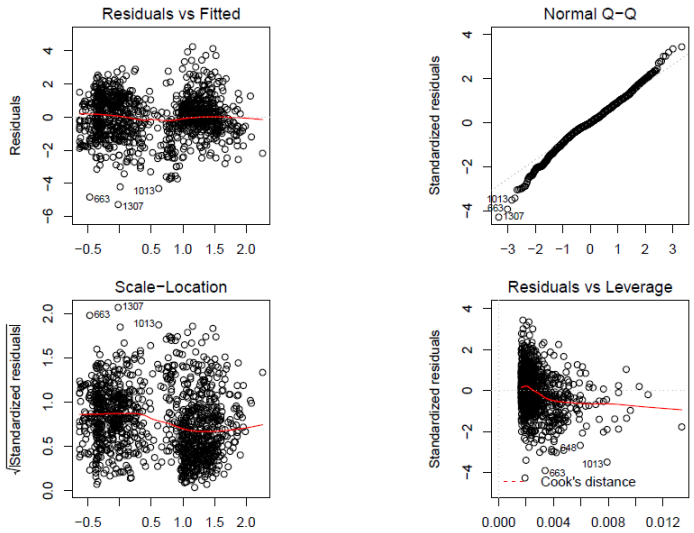
The estimated mean response for a subject of age 30, in case of ATTORNEY=yes, is $\hat{y} = 0.7376 + 30 \cdot 0.016 = 1.22$, with:

Standard error: 0.05
Confidence interval: [1.12,1.32]
Estimate on LOSS scale: $e^{1.22} = 3.38$
Interval on LOSS scale: $[e^{1.12}, e^{1.32}] = [3.06, 3.73]$

The estimated mean response for a subject of age 30, in case of ATTORNEY=no, is $\widehat{y} = 0.7376 - 1.3699 + 30 \cdot 0.016 = -0.15$, with:

Standard error: 0.05
Confidence interval: [-0.26,-0.05]
Estimate on LOSS scale: $e^{-0.15} = 0.86$
Interval on LOSS scale: $[e^{-0.26}, e^{-0.05}] = [0.77, 0.95]$

The diagnostic plots are moderately good, then the model is acceptable for practical purposes.



With the aim of looking for a better model specification, the following regression model is specified (according to R^2_{adj}, AIC and the test MSE estimated by simple Cross Validation)
$\log \text{LOSS} = \beta_0 + \beta_1 \text{ATTORNEY} + \beta_2 \text{CLMAGE} + \beta_3 \text{CLMAGE}^2 + \beta_4 \text{SEATBELT} + \varepsilon$

The qualitative variables ATTORNEY and SEATBELT are coded as dummy variables and the quadratic effect of CLMAGE is also considered.

	Estimate	SE	p-value
Intercept	-0.2249	0.1376	0.1024
ATTORNEYno	-1.3522	0.0725	< 0.0001
CLMAGE	0.0828	0.0075	< 0.0001
CLMAGE ^{2}	-0.0009	0.0001	< 0.0001
SEATBELTno	0.9241	0.2681	0.0006
	$s^2 = 1.404$		$R^2_{\text{adj}} = 0.321$

Diagnostic plots do not highlight any serious flaw in the model. One of the main usage of the model is for out-of-sample predictions.

The point predictor for the response variable Y_0, which corresponds to a certain value x_0 for the covariates, is $\widehat{Y}_0 = \text{x}_0^T \widehat{\boldsymbol{\beta}}$.

The prediction error (and then the SE of prediction) is made of two parts: the randomness of Y_0 and estimation error associated to the linear predictor $\widehat{\mu}_0 = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}$

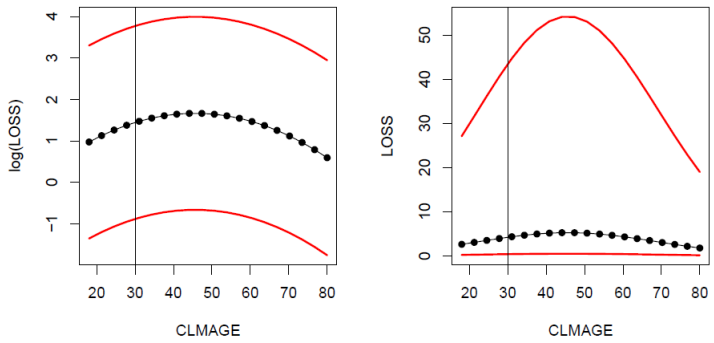
The objective is to predict the LOSS for an insured of age 30, that was represented by an attorney and that was wearing a seatbelt.

- The estimated variance for the fitted model (which correspond to the training MSE) is 1.40, which is close to 1.41, that is the test MSE assessed by cross-validation: here the overoptimism given by the training data is small.

- The 95% prediction interval for log(LOSS) is $[-0.89, 3.77]$, much wider than the confidence interval for $\widehat{\mu}_0$ given by [1.33,1.55].
- Adopting the original scale, the intervals are [0.41,43.43] and [3.80,4.72] , respectively. Using the cross-validation-based SE of prediction, the prediction interval becomes [0.41,43.62].

It is possible to compute the 95% prediction intervals for logLOSS and LOSS, for insureds represented by an attorney, wearing a seatbelt and with age ranging from 18 to 80 years.

The vertical lines identify the prediction intervals given before, for an insured of age 30.



6.5 The classification setting

In the prediction framework, the interest response variable may be qualitative (categorical) rather than quantitative. In such instances, the process of **predicting** the category of a new observation is referred to as **classification**.

There are some connections with the regression setting, as often the methods used for classification predict the probability of each of the categories of the response variable, and such probabilities are numerical scores (like the fitted values of a regression model).

The classification problem is ubiquitous in all the applications of statistics. Some examples: an email filter must classify a new email as spam or not spam; a person when arrives at the hospital emergency room with some symptoms has to be attributed to one of three medical conditions (such as stroke, drug overdose, epileptic seizure).

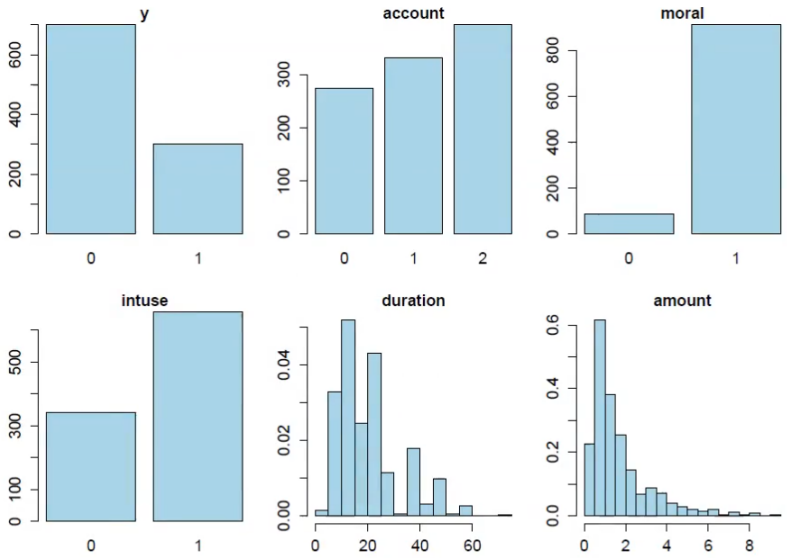
6.5.1 Example: credit scoring

A bank loan service has to evaluate the possible loan insolvency for a customer, and decide whether customer will default or customer will pay back: this is a problem which typically arises in the credit scoring setting. Data set on $n = 1000$ private credits issued by a German bank. Training data where every client is associated with a binary response:

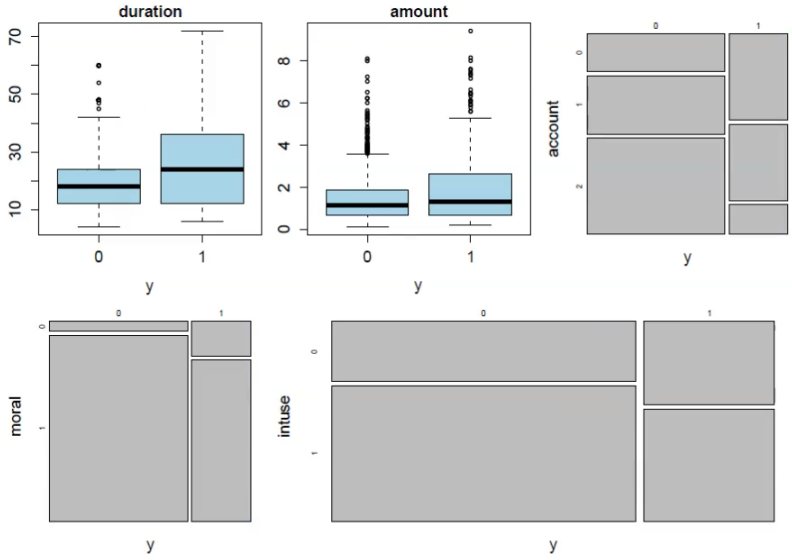
- Y: the client did not pay back his loan 1, the client paid back 0 and five explanatory variables (predictors);
- account: no running account 0, bad running account 1, good running account 2;

- duration: duration of the credit (in months);
- amount: credit amount (in thousands euros);
- moral: previous payment behavior, bad 0, good 1;
- intuse: intended use, business, 0, private 1.

The data set enables a marginal description of the response variable and of the explanatory variables:



A further analysis points to the potential relationships between the binary response and the explanatory variables:



6.6 Classification of a categorical response

Many of the concepts related to predictive model accuracy transfer over to the classification setting, with some modifications.

The case of a binary response is initially considered. The response for the i-th unit is coded as $y_i \in \{0, 1\}, i = 1, \ldots, n$. A classification method makes use of the training data $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$ to build a classifier that, for a given set of predictors x_0, returns a binary classification $\widehat{y}_0 \in \{0, 1\}$, which is very similar to what obtained for linear regression.

The most common approach for quantifying the accuracy of the classifier is the training error rate, the proportion of mistakes that are made if the classifier is applied to the training observations:

$$\text{trainingER} = \frac{1}{n} \sum_{i=1}^{n} I\left(y_i \neq \widehat{y}_i\right)$$

where $\widehat{y}_i$ the predicted i-th observation and $I\left(y_i \neq \widehat{y}_i\right)$ is an indicator variable that equals 1 if $y_i \neq \widehat{y}_i$ and 0 otherwise.

Since the same data are used twice, the training error rate would give an overoptimistic predictive assessment of the classifier. As in the regression setting, it is more convenient to evaluate the error rate on observations that were not used for training.

Thus, given the test observations $(\mathrm{x}_{01}, y_{01}), \ldots, (\mathrm{x}_{0m}, y_{0m})$, the test error rate is:

$$\text{test}ER = \frac{1}{m} \sum_{j=1}^{m} I\left(y_{0j} \neq \widehat{y}_{0j}\right)$$

which may be also computed from the training observations using cross-validation. The test error rates is an estimate of the prediction (classification) error for the future random response Y_0 corresponding to x_0:

$$E\left[I\left(Y_0 \neq \widehat{Y}_0\right)\right] = P\left(Y_0 \neq \widehat{Y}_0\right)$$

with $\widehat{Y}_0$ the associated classifier (predictor). The aim is to define a classifier which minimizes the prediction error.

6.7 The Bayes classifier

It is possible to show that the best classifier (with the smallest classification error) assigns an observation with predictor x_0 to the class 1 if:

$$P\left(Y_0 = 1 \mid \mathbf{X}_0 = \mathbf{x}_0\right) > P\left(Y_0 = 0 \mid \mathbf{X}_0 = \mathbf{x}_0\right)$$

and to class 0 otherwise. (Note that the costs of all errors are assumed to be the same).

This very simple classifier is known as the Bayes classifier, which is a gold standard achieving the best classification.

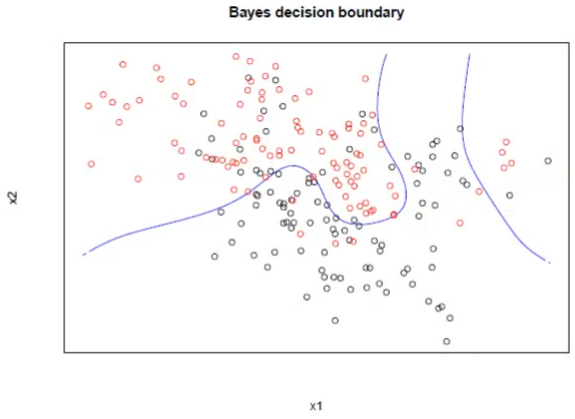
Unfortunately, for real data, the conditional probability distribution of Y_0 given $\mathrm{X}_0 = \mathrm{x}_0$ is not known, as it depends on the true distribution of the response variable. Therefore, computing the Bayes classifier is impossible.

Classification methods try to approximate such conditional probability following different assumptions and methodologies. Their performances would depend on how good this approximation is.

6.7.1 Example: two-predictors simulated data

Two continuous predictors, X_1 and X_2, and simulated binary responses (100 observations in each class, red and **black** circles).

Since the true model is known, the conditional probability is available and the Bayes classifier can be obtained. The blue line represents the points with probability 0.5: the Bayes decision boundary.



6.8 Classification based on logistic regression

Logistic regression models specify directly the conditional probability of the response, as approximated by the model and given by:

$$P\left(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i\right) = \frac{\exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\}}{1 + \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\}}$$

The decision boundary, which corresponds to the the values x such that $\widehat{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = 0.5$, has the following linear form $\mathbf{x}^T \widehat{\boldsymbol{\beta}} = 0$.

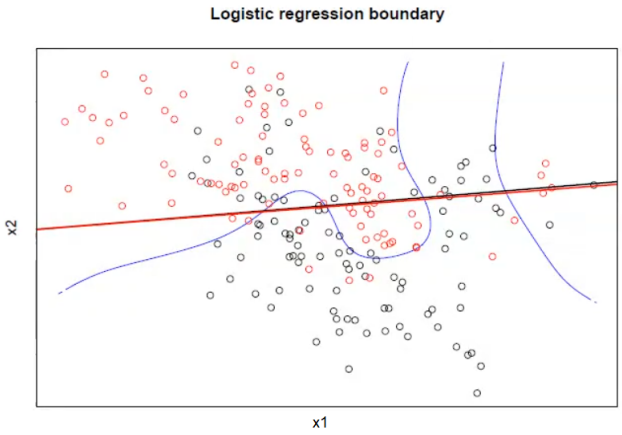
Classification based on logistic regression may be satisfactory whenever the Bayes decision boundary is roughly linear.

All the theory developed for regression models readily applies, even though, when the goal is classification, less attention is paid to inference on the coefficients, focusing instead on the classification performances.

Extension to more than two categories for the response are possible, but not used very often.

6.8.1 Example: two-predictors simulated data

The classification rule based on a logistic regression model has a linear boundary such that $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 0$. In this case, this is quite different from the Bayes decision boundary, and very similar to that one obtained using a linear regression model for the binary response, given by $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 0.5$.



6.9 Linear discriminant analysis (LDA)

Linear regression for binary (or categorical) response often gives results very similar to logistic regression, as shown in the two-predictor example.

Yet, it is fundamentally unsatisfactory, as it may give fitted probabilities outside [0,1]. Linear models for classification can be defined in an indirect way, starting from a model for the predictors and obtaining the conditional probability of interest $P(Y_i = y_i \mid \mathbf{X}_i = \mathbf{x}_i)$ by the Bayes theorem.

This is the essence of **Linear Discrimination Analysis (LDA)**, that ends up in a linear classification boundary in the space of (continuous) predictors.

Suppose that the response variable Y can take on $S \geq 2$ unordered values $y_s, s = 1 \ldots, S$, and that $\pi_s = P(Y = y_s)$ is the associated (prior) marginal probability.

Let $f_s(\mathrm{x}) = f(\mathrm{x} \mid Y = y_s)$ denote the density function of the p-dimensional vector X of continuous predictors for a response observation in the s-th category. According to the Bayes' theorem,

$$P(Y = y_s \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_s f_s(\mathbf{x})}{\sum_{r=1}^{S} \pi_r f_r(\mathbf{x})}$$

so that the Bayes classifier assigns an observation with predictor x to the category for which $P(Y = y_s \mid \mathbf{X} = \mathbf{x})$ is largest.

An approximate Bayes classifier is then defined by considering suitable estimates for $f_s(\mathrm{x})$(and, if needed, for the membership probability π_s), $s = 1, \ldots, S$, using the training observations.

The LDA requires that, for the case $p = 1$, $f_s(x)$ is the density of a Gaussian distribution $N(\mu_s, \sigma^2)$, with a class-specific mean value μ_s and a constant variance σ^2.

The approximation for the Bayes classifier is obtained by plugging into $P(Y = y_s \mid \mathbf{X} = \mathbf{x})$ the estimates:

$$\widehat{\mu}_s = \frac{1}{n_s} \sum_{i:y_i = y_s} x_i, \quad \widehat{\sigma}^2 = \frac{1}{n - S} \sum_{s=1}^{S} \sum_{i:y_i = y_s} (x_i - \widehat{\mu}_s)^2, \quad \widehat{\pi}_s = \frac{n_s}{n}$$

with n_s the number of training observations belonging to the s-class. The LDA assigns an

observation with $X = x$ to the class for which (using a simple transformation):

$$\delta_s(x) = x\frac{\widehat{\mu}_s}{\widehat{\sigma}^2} - \frac{\widehat{\mu}_s^2}{2\widehat{\sigma}^2} + \log\left(\widehat{\pi}_s\right)$$

is largest. This function is linear in x. The procedure can be easily extended to the case with $p > 1$.

If there are two categories, $\delta_1(x) = \delta_2(x)$ defines a linear decision boundary.

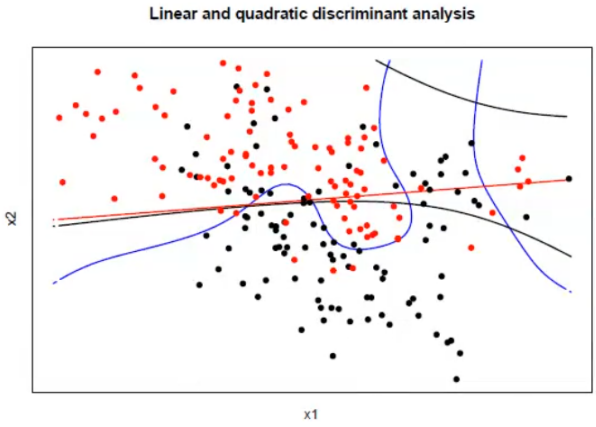
LDA and logistic regression often perform very similarly, though actually LDA is numerically more stable.

LDA assumes the normality of the continuous predictors, with class-specific means and common variance, so that it is not suitable for categorical predictors.

If the assumptions hold, and if the sample is very large (so that estimation error is very small), then the LDA classification rule approximates well the Bayes rule. Relaxing the assumption of common variance leads to Quadratic Discriminant Analysis (QDA), resulting in quadratic classification boundaries. It usually requires larger samples, and it is not so much used in practice.

6.9.1 Example: two-predictors simulated data

The classification rules based on the linear discriminant analysis and on the quadratic discriminant analysis have, respectively, a linear boundary and a quadratic boundary. In this case, they are quite different from the Bayes decision boundary. The LDA produces similar classifications to those obtained using a logistic regression model.



6.10 k-Nearest Neighbors (kNN)

The method of k-Nearest Neighbors (kNN) is a simple procedure, pertaining to the class of instance-based classification methods, that classify a new unit by using the observations in the training set with similar predictor values.

To classify a new observation with predictor x_0, the kNN classifier identifies the $k > 0$ points in the training data that are closest to x_0, forming the set $\mathcal{N}_0$. Then, the conditional

probability is estimated by:

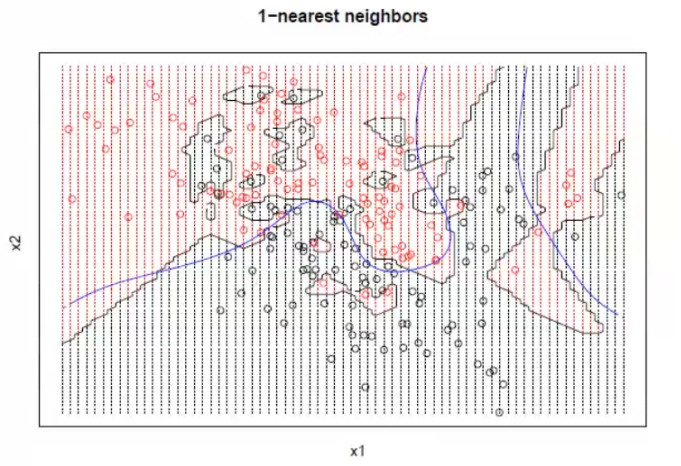
$$\widehat{P}\left(Y_0 = 1 \mid \mathbf{X}_0 = \mathbf{x}_0\right) = \frac{1}{k} \sum_{i \in \mathcal{N}_0} y_i$$

The value k is a positive integer that has a strong impact on the performance of the method. The best choices for k may lead to rather good performances (close to the Bayes classifier), despite the simplicity of the method.

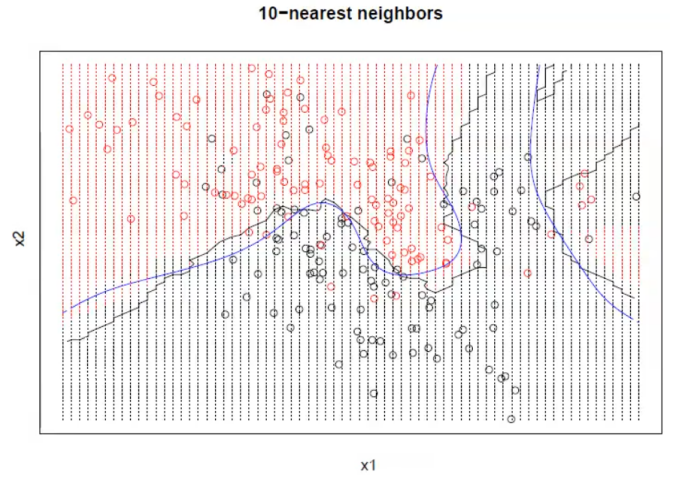
6.10.1 Example: two-predictors simulated data

The choice $k = 1$ uses only the closest point for classification. The plot below reports the classification done for a dense grid of values, giving the classification boundary.

The classification is even too detailed, with respect to that given by the Bayes decision boundary in blue.

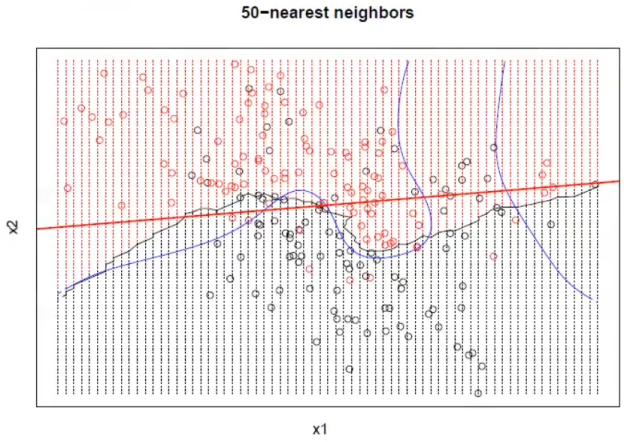


The choice $k = 10$ gives a less flexible classification, with a classification boundary much closer to the Bayes decision boundary.



The choice $k = 50$ gives a rough classification. The set of neighbors of each point is enlarged too much.

However, the classifier is closer to the Bayes classifier (blue) than the linear boundary (red) obtained via logistic regression.



6.11 Confusion matrix

The predictive performance of a binary classifier can be summarized using a confusion matrix, which cross-classifies the observed frequencies and the predicted ones:

	$Y = 0$(obs)	$\wedge Y = 1$(obs)
$Y = 0$(pred)	True Negative (TN)	False Negative (FN)
$Y = 1$(pred)	False Positive (FP)	True Positive (TP)

The percentages of correct classification corresponds to the:

true positive rate (sensitivity): TP/ (TP+FN);
true negative rate (specificity): TN/(FP+TN);
total accuracy rate: (TP+TN)/(FP+TN+TP+FN)

Further useful measures are the:

positive predictive value: TP/(FP+TP)
negative predictive value: TN/ (TN+FN)
log-odds ratio: log(TP*TN/(FN*FP))

These quantities are obtained using the same data for fitting the predictive model and for evaluating the predictive accuracy. A less optimistic, and more realistic, evaluation can be obtained using CV.

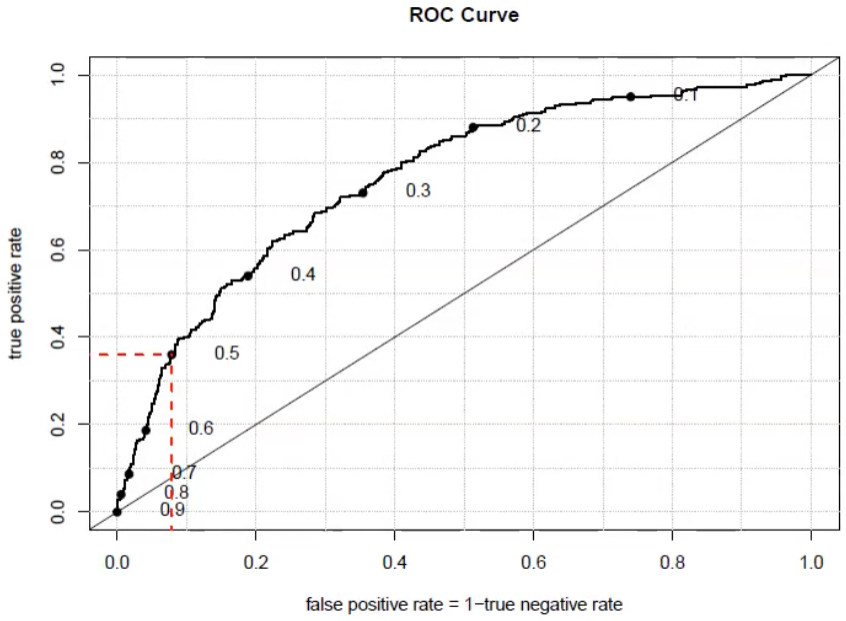
6.12 ROC curve

The choice of a classifying threshold equal to 0.5 is one possible choice. It could be useful to evaluate the global performance of a binary classifier taking into account all possible thresholds.

The **Receiver Operating Characteristic (ROC) curve** is a popular graphic created by plotting the true positive rate against the false positive rate (given by 1 - true negative rate) at various threshold settings. The overall performance of a classifier can be described by the area under the ROC curve (**AUC**).

An ideal ROC curve will pass near the top left corner, so that the larger the AUC the better the classifier. Roc curves are useful for comparing alternative binary classifiers, since they take into account all possible thresholds.

The threshold 0.5 for classifying an observation is not necessarily the best choice, especially if the two errors have different cost. The ROC curve illustrates the performance of the binary classifier for all possible thresholds.



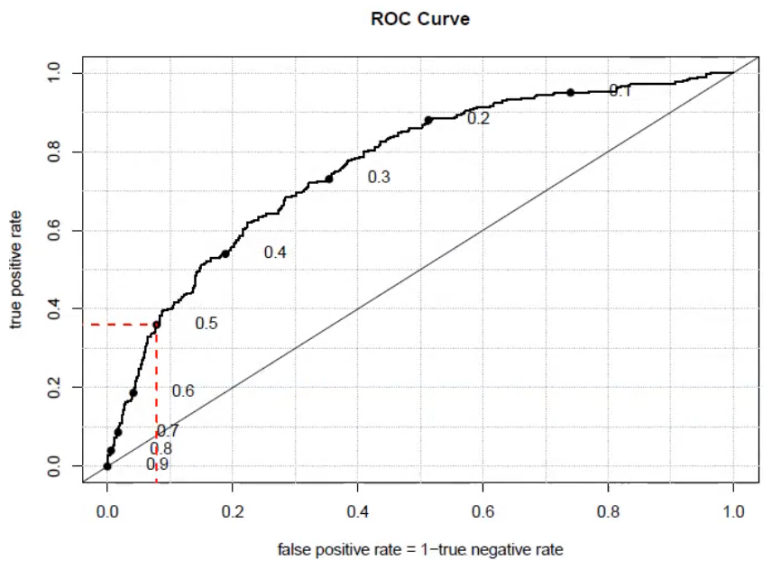
6.12.1 Example: credit scoring

A multiple logistic regression model is fitted, including all the five predictors, and it is used for classification on the same training data. To summarize its predictive performance, the confusion matrix is calculated:

	not defaulting (obs)	defaulting (obs)
not defaulting (pred)	645	192
defaulting (pred)	55	108
Total	700	300

The predictive model seems to be satisfactory only for predicting the customers not at risk of defaulting: the true negative rate and true positive rate are $645/700 = 0.921$ and $108/300 = 0.360$, respectively; the total accuracy rate is $(645 + 108)/1000 = 0.753$.

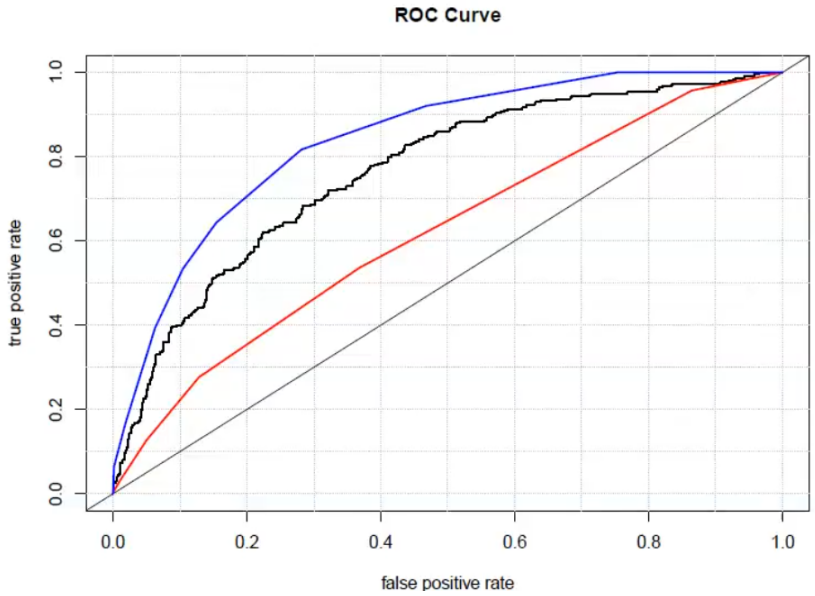
A more reliable evaluation is obtained via (10-folds) cross-validation, but in this case it makes little difference: the rates are 0.920,0.350 and 0.749, respectively.



The threshold 0.5 for classifying an observation is not necessarily the best choice, especially if the two errors have different cost. The ROC curve illustrates the performance of the binary classifier for all possible thresholds.

ROC curves are useful for comparison, with the best classifier producing the highest curve.

Here the comparison regards the classifiers based on logistic regression, on LDA (red) using only the numerical predictors and kNN (blue) with optimal K.



With regard to the credit scoring data, the optimal K value for the kNN classifier is $K = 9$.

This value has been chosen by a leave-one-out cross-validation procedure. For such a non-linear classifier, the assessment based only on the training data is indeed over-optimistic.

	tot. accuracy rate	true negative rate	true positive rate
training data	0.79	0.90	0.53
cross-validated	0.75	0.89	0.43

A further nonlinear classifier could be obtained by extending logistic regression, including nonlinear terms for the continuous predictors in the same way illustrated for linear regression models. The additional gain of such extension is, however, rather limited in this case.

6.13 Further methods

Classification is a very broad area, and the methods presented here are just the simplest ones. A (partial) list of other useful methods may include: Naive Bayes: simple classifier which treats the predictors as independent random variables Elassification trees and decision stumps: very general, simple methods providing results easy to understand Ensemble methods (such as Boosting, Bagging, Random forests): usually based on iterated applications of a simple classifier Support vector machines: combine linear models with instance-based methods. Classification is a fundamental problem in statistical learning and there are multiple methods available, each with some pros and cons. No matter which method is used, it is always essential to estimate the classification accuracy, avoiding over-optimistic assessments based only on the training data.

7 Unsupervised methods

Most statistical learning problems fall into one of two categories: supervised or unsupervised.

The problems discussed so far belong to the **supervised learning** framework: for each observation $\mathbf{x}_i, i = 1, \ldots, n$, of the p predictor variables there is an observation y_i for the associated response variable.

A model that relates the response to the predictors is fitted and it can be considered for both interpretation and prediction purposes. Linear regression and logistic regression models operate in this context.

The **unsupervised learning** framework is, in some sense, more challenging, since for every observation $i = 1, \ldots, n$, only a vector of measurements x_i is given, with no associated response y_i.

This situation is unsupervised because a response variable, that can supervise the analysis, is not available. There is no way to check the results by seeing how well the model predicts a response.

The focus here is on unsupervised learning, which is concerned with the joint study of a set of p variables, $X_1, \ldots, X_p$, observed for a sample of size n.

The **data matrix** is then given by X and it has size $n \times p$; the element x_{ij} corresponds to the i-th observation on the j-th variable. There is no response variable, and all the p variables are treated on an equal footing.

The goal is to discover interesting things about the measurements on the p variables $X_1, \ldots, X_p$ and, in particular, if there is an informative way to visualize and to summarize the data and/or if there are subgroups among the variables or among the observations.

The statistical techniques considered in this framework are typicalllly **explorative** in nature. They were classically referred to as multivariate analysis techniques, but this name is perhaps too vague.

Unsupervised learning is often performed as part of an exploratory data analysis.

This exercise tends to be somewhat subjective, since there is no simple goal for the analysis, such as prediction of a response. Like any other statistical techniques, the first step for analyzing multivariate data is given by data description and visualization, that can be far from easy in high dimensions.

For a first look at the data, scatterplot matrices and low-dimensional view are always useful, yet to consider a number of plots can miss important structures in the data. More advanced techniques, such as dynamic graphics, can be quite effective, as they allow for rotation of the point cloud and projection in lower dimensions.

There are many unsupervised techniques, and the focus here on two classes of methods: **Principal Components Analysis** and **Cluster analysis**.

7.1 Principal Components Analysis

Principal Components Analysis (**PCA**) replaces the input variables by a set of new derived variables, called principal components.

These components are ordered according to the amount of variation of the original variables they are able to explain.

The method is useful for understanding multivariate data and it also serves as a tool for data visualization.

Plots of the first principal components are often insightful, leading to effective dimension reduction, useful in particular when p is large.

PCA gives a low-dimensional representation of the data, based on small number of interesting dimensions, that captures as much of the information as possible.

This idea is at times useful in regression, where a large number of candidate explanatory variables may be replaced by the first few principal components, provided they synthesize adequately the information in the candidate variables.

Given a set of p variables $X_1, \ldots, X_p$, the **first principal component** Z_1 is the normalized linear combination:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p$$

with the largest variance.

The weights $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})^T$ are called **loadings** and they are normalized, namely $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

This constraint is required since setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

The principal components depend on the scaling of the variables, then it is important that the variables are in comparable units. Thus, it is typically recommendable to **standardize** the variables before applying the method.

Obtaining the loadings for the first principal component is a simple linear algebra task.

Given the $n \times p$ data matrix X, the aim is to derive the loadings $\phi_{11}, \ldots, \phi_{p1}$ such that the linear combination of the n sample values:

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots + \phi_{p1} x_{ip}$$

has largest variance, subject to the constraint $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ since the variables are centered (that is the column means of X are zero), the sample variance is (the sample mean is zero):

$$\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2$$

After the loading vector ϕ_1 is found, the observed values for the first principal component Z_1 are obtained.

These values, one for each observation, are $z_{i1}, i = 1, \ldots, n$, and they are referred to as the scores of the first principal component.

The second principal component Z_2 is the normalized linear combination of $X_1, \ldots, X_p$ that has maximal variance out of all linear combinations that are uncorrelated with Z_1.

The second principal component scores $z_{12}, \ldots, z_{n2}$ take the form:

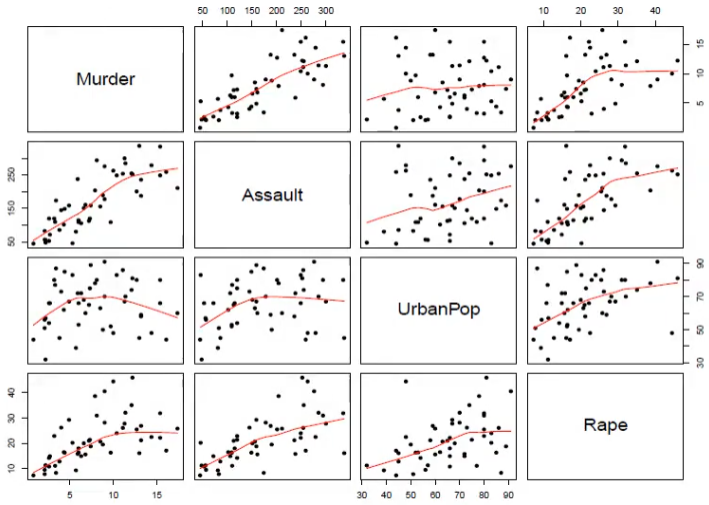
$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \ldots + \phi_{p2} x_{ip}$$

where the second principal loading vector $\phi_2 = f\,(\phi_{12}, \ldots, \phi_{p2})^T$ is such that $\sum_{j=1}^{p} \phi_{j2}^2 = 1$ and $\sum_{j=1}^{p} \phi_{j1} \phi_{j2} = 0$ (uncorrelation). The procedure can be iterated, up to a $m = \min(n -$

$1, p$) principal components. Geometric interpretation: the first loading vector ϕ_1 defines a direction in the variable space along which the data vary the most; projection of the n data points onto this direction gives the first principal component scores. The second loading vector ϕ_2 defines a direction, orthogonal (perpendicular) to the direction ϕ_1, along which the variability is maximized; data projections onto this further direction gives the second principal component scores.

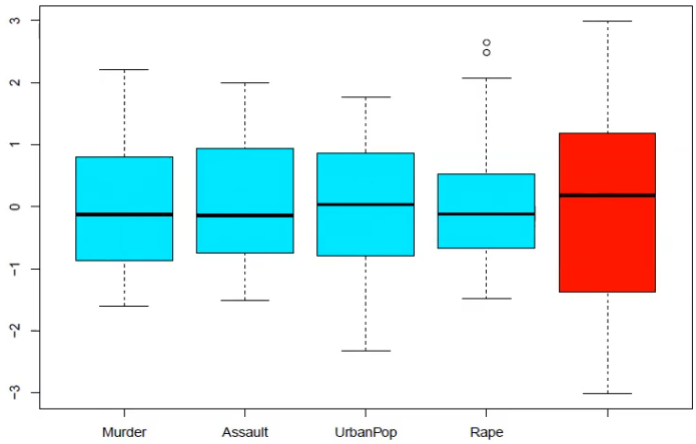
7.1.1 Example: US arrest data

Data set containing statistics, in arrests per 100,000 residents for Assault, Murder, and Rape in each of the 50 US states in 1973, and the percent of the population (UrbanPop) living in urban areas (4 variables and $n = 50$ observations).



A boxplot of the four (standardized) variables compared with the scores from the first principal component readily shows that the latter is more variable.

Using the first principal component, some global information has been extracted.



The first two principal component loading vectors, ϕ_1 and ϕ_2, are given below:

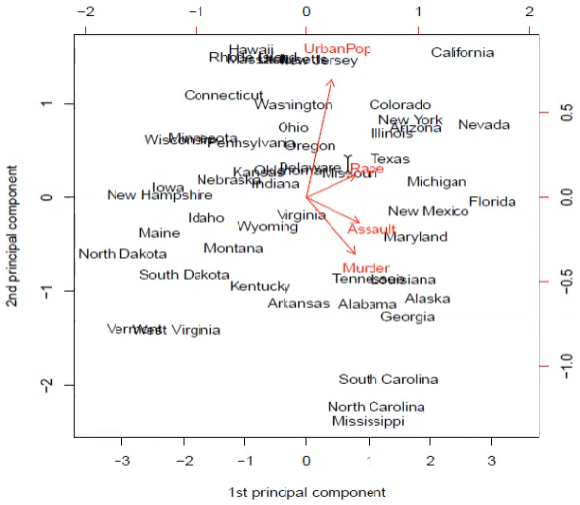
	Murder	Assault	UrbanPop	Rape
PC1	0.5358995	0.5831836	0.2781909	0.5434321
PC2	-0.4181809	-0.1879856	0.8728062	0.1673186

The first loading vector places approximately equal weight on Murder, Assault, and Rape, with much less weight on UrbanPop. Then the first principal component Z_1 roughly corresponds to a measure of overall rates of serious crimes.

The second loading vector places most of its weight on UrbanPop and much less weight on the other three variables.

Hence, the second principal component Z_2 roughly corresponds to the level of urbanization of the state.

The biplot is a graphical summary which displays both the scores and the loadings associated to the first two principal components. The state names represent the scores for the first two principal components (left and down axes), while red arrows indicate the first two loadings associated to the four variables (top and right).



The crime-related variables (Murder, Assault and Rape) are located close to each other, and the UrbanPop variable is far from the other three.

This indicates that the crime-related variables are correlated with each other (states with high murder rates tend to have high assault and rape rates) and that the UrbanPop variable is less correlated with the other three. States with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates, while states like North Dakota, with negative scores on the first component, have low crime rates.

California also has a high score on the second component, indicating a high level of urbanization. States close to zero on both components, such as Indiana, have approximately average levels of both crime and urbanization.

7.1.2 Further comments

Uniqueness of the principal components: each component is unique up to a sign flip. Flipping the sign to all the estimated loadings gives totally equivalent results, since the

direction specified in the p-dimensional space is the same. Similarly, the score vectors are unique up to a sign flip, since the variance of Z is the same as the variance of $-Z$.

Alternative geometrical interpretation: the first r principal components score and loading vectors provide the best r-dimensional approximation (using the Eucligean distance in $\mathbf{R}^p$) to the i-th observation, namely $x_{ij} \approx \sum_{s=1}^{r} z_{is}\phi_{js}$; the approximation is exact with $r = \min(n-1, p)$

How many principal components: the first component is the most informative one-dimensional linear summary, and the first two provide the most informative two-dimensional graphical summary. Under this respect, a key point is how much of the variance in the data is extracted by a given number of principal components.

7.2 The proportion of variance explained

Assuming that the observed variables have mean zero, the proportion of variance explained by the s-th component is:

$$\text{PVE}_s = \frac{(1/n)\sum_{i=1}^{n} z_{is}^2}{\sum_{j=1}^{p}(1/n)\sum_{i=1}^{n} x_{ij}^2}$$

The numerator is the (estimated) variance of the s-th component and the denominator estimates the total variance $\sum_{j=1}^{p} V(X_j)$.

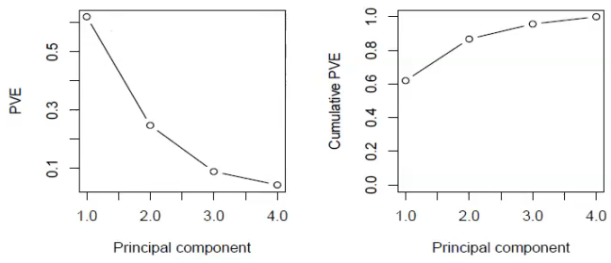
The cumulative proportion of variance explained for the first r components is then $\sum_{s=1}^{r} \text{PVE}_s$, and if $r = \min(n-1, p)$ such value is just 1. A useful graphical representation is given by the scree plot, which describes the values of PVE_s, for $s = 1, \ldots, \min(n-1, p)$.

There is no well-accepted objective way to decide how many principal components to consider, since it depends on the specific application.

It is customary to consider the scree plot and to look for a point where the proportion drops off, the so-called elbow.

7.2.1 Example: US arrest data

The screeplot (left panel) and the cumulative proportion of variance explained (right panel) are given below:



The first component explains more than 60% of the total variance, and the second one about 25% : they explain almost 87% of the variance in data, and provide a very useful summary using only two dimensions. The scree plot displays an elbow after the second principal component.

7.3 Clustering methods

Clustering methods refer to a very broad class of techniques for finding subgroups, or clusters, in a data set. When the aim is to cluster the observations of a data set, the task is to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.

For example, in a market segmentation study the available data concerns p characteristics (variables) for n potential customers and the aim is to classify the customers into different groups, such as big spenders versus low spenders, without the knowledge of the customer's spending patterns. Indeed, the n observations may correspond to tissue samples for patients with breast cancer, and the p variables are measurements collected for each tissue sample. Clustering could be used to find subgroups related to different unknown subtypes of breast cancer.

Clustering aims at discovering groups in data, and in particular in the n rows of the $n \times p$ data matrix X. In general, the objective is to cluster the observations on the basis of the features (variables) in order to identify subgroups among the observations, but it is as well possible to cluster features on the basis of the observations in order to discover subgroups among the features. In what follows, the focus is on clustering observations, though the converse can be performed by simply transposing the data matrix. Both clustering and PCA seek to simplify the data via a small number of summaries, but their mechanisms are different. Cluster analysis is widely used, but it is useful to remember that, in many cases, visualization based on specialized software may be a compelling alternative.

7.4 Clustering algorithms

Clustering methods are usually based on a measure of **dissimilarity** between units, but other than that they greatly differ among them. A rough classification of clustering methods is as follows.

1. **Hierarchical clustering**, that seeks to build a hierarchy of clusters. There are agglomerative (bottom-up) or divisive (top-down) approaches. They typically produce a graphical output called dendrogram, and need to be tailored to the data at hand.

2. **Partitioning clustering**, which tries to iteratively optimize the allocation of units to clusters. The most commonly used method of this class is K-means clustering, which requires to specify the number of clusters in advance and it is suitable for continuous data.

3. **Model-based clustering** fits an actual model to the data. Methods of this kind are usually computationally harder, but when the model is sensible for the data at hand they produce a quite reliable output.

Different clustering methods may give different answers, and there is some risk of over-interpretation. Clustering methods are better used in conjunction with visualization techniques, and principal component analysis is very useful in that respect.

7.5 Measure of dissimilarity

Many methods for cluster analysis starts from a measure of dissimilarity (or of similarity) between observations or cases (rows of the data matrix X).

A dissimilarity coefficient d has the following three properties. For each $\mathrm{a} = (a_1, \ldots, a_p)$, $\mathbf{b} = (b_1, \ldots, b_p)$ and $\mathbf{c} = (c_1, \ldots, c_p)$,

- $d(\mathbf{a}, \mathbf{b}) \geq 0$ (non-negativity);
- $d(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$ (identity of indiscernibles);
- $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$(symmetry).

Furthermore, for a metric dissimilarity (**distance**):

$$d(\mathbf{a}, \mathbf{c}) \leq d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c})$$

and for an ultrametric dissimilarity:

$$d(\mathbf{a}, \mathbf{c}) \leq \max\{d(\mathbf{a}, \mathbf{b}), d(\mathbf{b}, \mathbf{c})\}$$

A dissimilarity need not be a distance.

Given two numeric vectors $\mathbf{a} = (a_1, \ldots, a_p)$ and $\mathbf{b} = (b_1, \ldots, b_p)$, the following distance measures can be defined:

- **Euclidean distance:**

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

- **Manhattan (taxicab) distance:**

$$d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{p} |a_i - b_i|$$

- **Maximum distance:**

$$d(\mathbf{a}, \mathbf{b}) = \max_i |a_i - b_i|$$

For non-numeric vectors $\mathbf{a} = (a_1, \ldots, a_p)$ and $\mathbf{b} = (b_1, \ldots, b_p)$, the following distance (dissimilarity) measures can be defined:

- **Binary distance:** (for binary vectors, with only 0 s and 1 s) : it is the percentage of nonzero coordinates (namely, other than (0,0)) that differ.

It is a special case of the Jaccard distance (for categorical variables with a preferred level): the proportion of such variables with one of the cases at the preferred level (level 1 in case of binary distance) in which the cases differ. In general, given the sets A and B, it is

$$d(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

where $|\cdot|$ is the size of a given set.

- **Hamming distance:** (for strings or categorical data): it is the number of coordinates where the two strings differ.

- **Gower dissimilarity:** (for mixed, numeric-categorical, data): it is based on a more involved formula. It is a non-metric dissimilarity, very useful for real data sets.

7.6 Hierarchical clustering

Hierarchical clustering algorithms connect "objects", to form "clusters", based on their distance. It results in an attractive tree-based representation of the observations, called a dendrogram.

There are agglomerative hierarchical methods (bottom-up) and divisive hierarchical methods (top-down). Agglomerative methods produce a set of clusterings, starting with one cluster for each observation, and then merging pairs of clusters as moving up the hierarchy.

It is the most common type of hierarchical clustering. Divisive methods also produce a set of clusterings, starting from a single cluster and making successive splitting.

They are computationally harder, and may be attractive when grouping into a few large clusters is of interest. In some situations, the assumption of a hierarchical structure might be unrealistic (for example, a group of people classified by gender or by nationality).

7.7 Linkage criteria

Hierarchical clustering is a whole family of methods that differ by the way dissimilarities are computed. Furthermore, in order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a suitable dissimilarity measure between sets of observations is employed. Linkage criteria define the dissimilarity between two groups of observations, starting from a notion of dissimilarity between a pair of observations. Some common choices are:

- **complete linkage:** the dissimilarity between clusters is the maximum of the dissimilarities between their members;
- **single linkage:** the dissimilarity between clusters is the minimum of the dissimilarities between their members;
- **average linkage:** the dissimilarity between clusters is the average of the dissimilarities between their members;
- **centroid linkage:** the dissimilarity between clusters is defined as the dissimilarity between their centroids (mean vectors).

7.8 Dendogram

Hierarchical methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters.

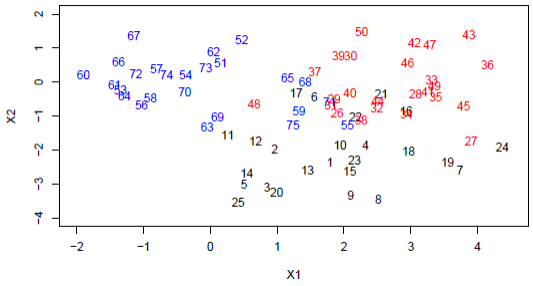
They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (namely, the "chaining phenomenon", in particular with single linkage).

The result of hierarchical clustering is typically displayed by means of tree diagrams referred to as dendrograms. In a dendrogram, the y-axis indicates the distance at which the clusters merge, while the objects are placed along the x-axis. The hierarchy of clusters in a dendrogram is obtained by cutting it at different heights.

There are many proposals, but no general and effective guidelines for performing such a task: cluster analysis is exploratory in nature, and the optimal number of clusters is context-specific.

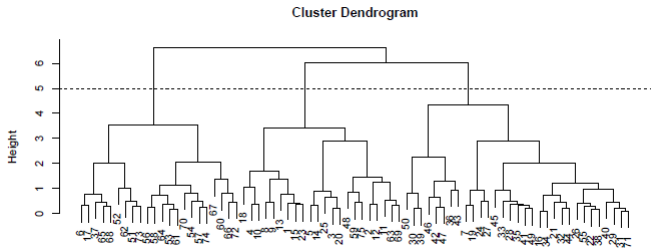
7.8.1 Example: three clsuters simualted data

Simulated data set with $n = 75$ observations of the variables X_1 and X_2; 25 bivariate observations in each cluster: black, red and blue numbers.



The class labels are treated as unknown and the aim is to cluster the observations in order to discover classes from the data.

Hierarchical clustering, with complete linkage and Euclidean distance, is performed. Different results can be obtained with alternative dissimilarity and linkage criterion.



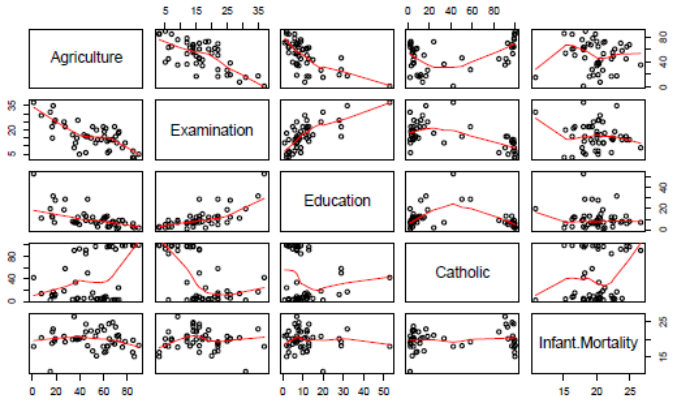
Each leaf of the dendrogram represents one of the 75 observations. When moving up the tree, some leaves begin to fuse into branches. These correspond to observations that are similar. The height of this fusion is measured on the y-axis. Thus, observations that fuse at the very bottom of the tree are quite similar.

Conclusions about the similarity of two observations should not be based on their proximity along the x-axis. Rather, on the location on the y-axis where branches containing those two observations first are fused. If the dendrogram is cut at the height of 5, three distinct clusters are specified. Alternative cut points may give different clustering results.

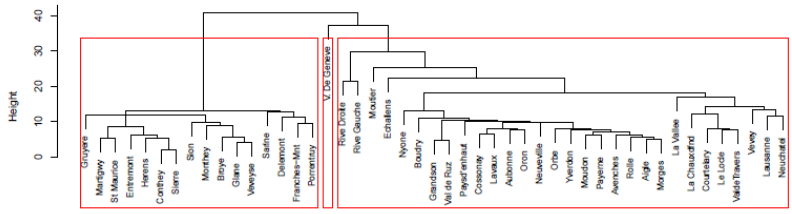
7.8.2 Example: Swiss socioeconomic indicators

Data set on socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888; $n = 47$ observations on $p = 5$ variables, each of which is in percent.

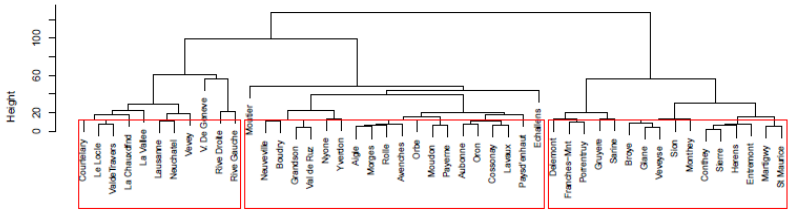
Since data are percentages, then the Euclidean distance is a reasonable choice as dissimilarity between pairs of observations.



Single linkage, agglomerative clustering is performed. The tree is cut into three clusters, indicated by red rectangles. There are two main groups, with a single point, well separated from them.



Divisive clustering is also performed, giving the following three clusters.



7.9 Partitioning clustering

Partitioning methods aim at classifying the observations into $K > 0$ distinct, non-overlapping clusters. These methods require to fix in advance the number K of clusters. Although there are criteria for choosing K in an iterative fashion, this selection problem is far from simple.

In partitioning clustering, it is required the decision on how many clusters are expected in the data, as in hierarchical clustering it is necessary to cut the dendrogran, in order to obtain clusters. An initial cluster assignment is required for the observations.

There are different partitioning algorithms and several alternative optimality criteria, some of them are based on probabilistic models.

7.10 K-means clustering

K-means is by far the most commonly used partitioning clustering method. It aims at minimizing the within-cluster variation. It is most appropriate for continuous variables, suitably scaled; customary implementations use the Euclidean distance. Given the sets (clusters) $C_1, \ldots, C_K$, containing the indices of observations and satisfying these two properties:

- $C_1 \cup \cdots \cup C_K = \{1, \ldots, n\}$ (each observation belongs to at least one cluster);
- $C_r \cap C_s =$, for all $r \neq s$ (no observation belongs to more than one cluster);

the aim is to solve the following optimization problem:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{r=1}^{K} W(C_r) \right\}$$

where $W(C_r)$ measures the within-cluster variation of C_r (the amount by which its observations differ from each other). There are many possible ways to define this concept, but by far the most common choice is:

$$W(C_r) = \frac{1}{|C_r|} \sum_{r,s \in C_r} \sum_{j=1}^{p} (x_{rj} - x_{sj})^2$$

namely, the sum of all of the pairwise squared Euclidean distances between the observations in C_r, divided by the total number of observations in the cluster.

The K-means clustering algorithm involves the following steps:

1. a number from 1 to K is randomly assigned to each of the observations (initial cluster assignments for the observations);

2. until the cluster assignments stop changing:

2a. the centroid (the vector of the p feature means) for each of the K clusters is computed;

2b. each observation is assigned to the cluster whose centroid is ckisest (according to the Euclidean distance).

Alternatively, the starting point of the algorithm may be given by the centroids identified by group-average agglomerative hierarchical clustering. Differently from hierarchical clustering, K-means requires the entire data matrix, not just the matrix of dissimilarities.

The optimization problem is usually very difficult to solve precisely: there are almost K^n ways to partition n observations into K clusters. The K-means algorithm finds a local rather than a global optimum: the result depends on the initial (random) cluster assignment.

For this reason, it is important to run the algorithm multiple times from different random initial configurations. Then one selects the best solution, namely, that for which the objective function is smallest.

7.11 K-medoids clustering

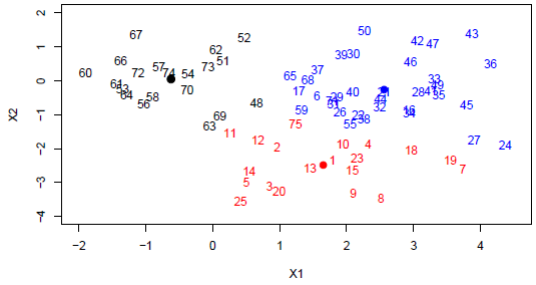
K-medoids methods are a variation of K-means clustering. As K-means, K-medoids algorithms are partitioning methods, attempting to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster.

In contrast, K-medoids methods choose datapoints as centers, and work with an arbitrary dissimilarity between points, rather than the Euclidean distance.— Then, only the matrix of dissimilarities is required and the results are more robust to noise and outliers, as compared to those obtained with K-means.

There are several K-medoids variants. The most common is the Partitioning Around Medoids (PAM) algorithm, with notable applications to genomic data. It uses a greedy search procedure which may not find the optimum solution, but it is faster than exhaustive search algorithms.

7.11.1 Example: three clusters simulated data

K-means clustering, with $K = 3$, is applied to the simulated data set with $n = 75$ bivariate observations. The resulting partition is given below, with indication of the centroids of the three clusters.



Multiple (in this case 20) initial random cluster assignments are considered and the best solution is selected. This procedure is strongly recommend, since otherwise an undesirable local optimum may be obtained.

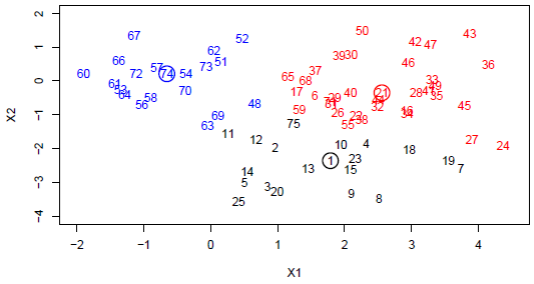
Alternatively, K-medoids methods can be considered. In this case, the three clusters corresponds exactly to those ones obtained using the K-means procedure.

The PAM algorithm is considered. It is based on the search for $K = 3$ representative objects or medoids among the observations of the data set. These observations should represent the structure of the data. Then the clusters are constructed by assigning each observation to the nearest medoid. The clusters are given below, with indication of the three medoids (observations used as cluster centers)

7.12 Model based clustering

Model-based clustering provides a more thorough methodology, which includes criteria for choosing the number of clusters and for assessing the goodness of the solution found.

For continuous data, the employed models are mixture of multivariate normal distributions, which are capable of approximating well a broad array of multivariate distributions.

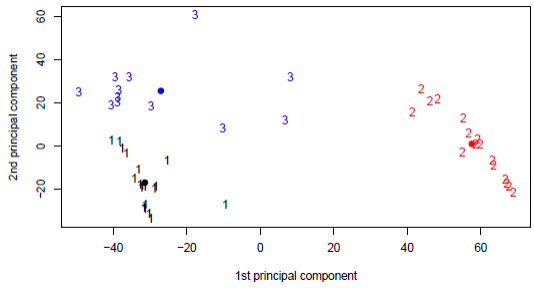


The model is estimated by maximum likelihood estimation or by the Bayesian approach. Unfortunately, there are many algorithmic parameters to tune in order to use the method successfully.

Furthermore, the method needs a deeper understanding of the underlying theory with respect to other clustering methods, and the theory for model-based clustering is far more complex.

7.12.1 Example: Swiss socioeconomics indicators

With regard to the data set on Swiss provinces, the K-means clustering, with $K = 3$, is considered. As a starting point, the means of the three clusters identified by hierarchical clustering, with average linkage, are taken into account. The result is given below, with indication of the centroids of the three clusters. since there are $p > 2$ variables, the observations are plotted by considering their first two principal components.



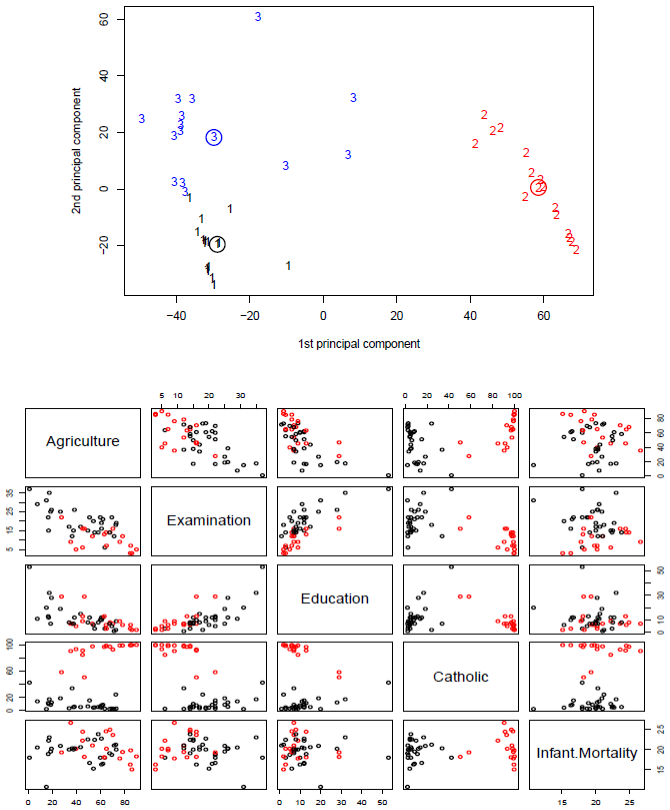
A similar result is obtained by considering multiple initial random cluster assignments. Furthermore, K-medoids methods can be applied. Using the PAM algorithm, with $K = 3$, the following clusters are specified, with indication of the three medoids (observations used as cluster centers)

All the obtained clusters tend to isolate the counties with a prevalence of Catholics, splitting the remaining counties in two groups.

The scatterplot matrix describes the relationships among the 5 variables. In red the counties with Catholic majority, which displays a peculiar behavior, as pointed out using clustering techniques.

7.13 Practical issues

In order to perform clustering, some crucial decisions must be made, with particular regard to: - the standardization of the variables; the choice of the dissimilarity, the linkage criterion



and the position where to cut the dendrogram (hierarchical clustering); the number K of clusters to be considered (partitioning clustering). Each of these decisions can have a strong impact on the results obtained and, usually, there is no single right answer.

In practice, several different choices should be tried in order to look for the one giving the most useful or interpretable result. There are advanced methods for selecting the number of clusters (and the most effective ones are based on simulations), or for comparing the similarity of two alternative cluster solutions.

However, it is very hard to try to understand whether the clusters that have been found represent true subgroups in the data, or whether they are simply a result of clustering the noise.

In addition, clustering methods generally are not very robust to perturbations to the data. since clustering methods force every observation into a cluster, the clusters found may be heavily distorted due to the presence of outliers, not belonging to any cluster.

Model-based clustering and mixture models are an attractive approach for accommodating the presence of such outliers. Cluster analysis is a very useful methodology, which can often add some information to other statistical analyses, even if it is not prudent to view the results as the absolute truth about a data set.

As there are so many clustering methods available, the recommended strategy is to try more than one method, and assess whether the resulting outcomes are similar.