Statistical Data Analysis

Lorenzo Leuzzi

2023

Contents

1 Module 1 **2**
1.1 Gathering Data - Types of Statistical Data Studies 2
1.2 Confidence Intervals and Significance Tests 5
1.3 Methods for Comparing Two Groups 8

2 Module 2 **11**
2.1 Multivariate Descriptive Statistics 11
2.2 Principal Components Analysis 15

3 Module 3 **17**
3.1 Intro 17
3.2 Statistical Regression Models 17
3.3 Regression Diagnostics 21

4 Module 4 **25**
4.1 Time Series - Basic Concepts and Models 25
4.2 Time Series Statistical Models - Basics, Stationarity and Autocorrelations 31
4.3 Time Series - ARIMA Models 34

Module 1

1.1 Gathering Data - Types of Statistical Data Studies

1.1.1 Introduction

Statistical methods help us investigate questions in an objective manner. Statistical problem solving is an investigative process that involves four components:

- Formulate a statistical question
- Collect data
- Analyze data
- Interpret results

The three main components of statistics for answering statistical questions are:

- **Design** refers to planning how to obtain relevant data that will efficiently shed light on the problem of interest. The design often involves taking a sample of the population of interest.
- **Description** means exploring and summarizing patterns in data.
- **Inference** means making decisions or predictions based on the data. It helps in deciding whether observed patterns are meaningful. Results of a study are considered statistically significant if they would rarely be observed with only ordinary random variation.

1.1.2 Method of Comparison

To evaluate whether a treatment has an effect it is crucial to compare the outcome when treatment is applied (the outcome for the treatment group) with the outcome when treatment is withheld (the outcome for the control group), in situations that are as alike as possible but for the treatment. This is called the method of comparison.

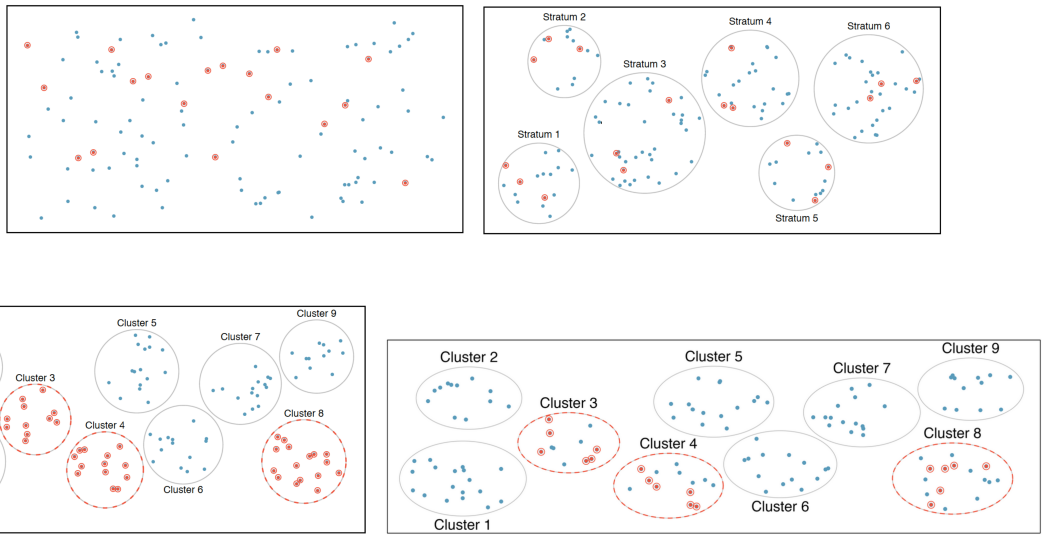
1.1.3 Observational Studies

In observation studies data is collected in a way that doesn't directly interfere with how data arise (no imposing a treatment). Researchers only observe what happened to the treatment group.
It is necessary to be careful about potential lurking variables, they are unobserved but can influence the association between the observed variables. Lurking variables have the potential for confounding, in order to prevent that it's important to match the data.

Sample Surveys

Sample surveys select a sample of people from a population and collects data from them. They take a cross section of a population at the current time. In sample surveys the sampling frame is the list of subjects in the population from which the sample is taken. Once you have a sampling frame, you need to specify a method for selecting subjects from it. The method used is called sampling design

or sampling method. Let's consider here four random sampling techniques: simple random sampling (srs), stratified sampling, cluster sampling and multistage sampling.



When results from the sample are not representative of the population, they are said to exhibit bias. Undercoverage occurs when having a sampling frame that lacks representation from parts of the population. Volunteers (Convenience samples) make sampling biased, use a random sampling design to select n subjects from the sampling frame in a way that all subjects in the population are somehow represented.
If the sample is random, the absolute size of the sample matters much more than the size relative to the population total. When using a simple random sample of n subjects in estimating a proportion, the approximate margin of error is $1/\sqrt{n}$.

Longitudinal Studies

They are studies in which an inherent temporal dimension is present. Rather than taking a cross section of a population at some time, such as with a sample survey, some studies are backward looking (retrospective) or forward looking (prospective).

- **Retrospective studies**, such as a case-control study, looks into the past; they collect data after events have taken place. Here, researchers review past events in medical records.
- **Prospective studies** identify individuals and collect information as events unfold; they follow their subjects into the future.

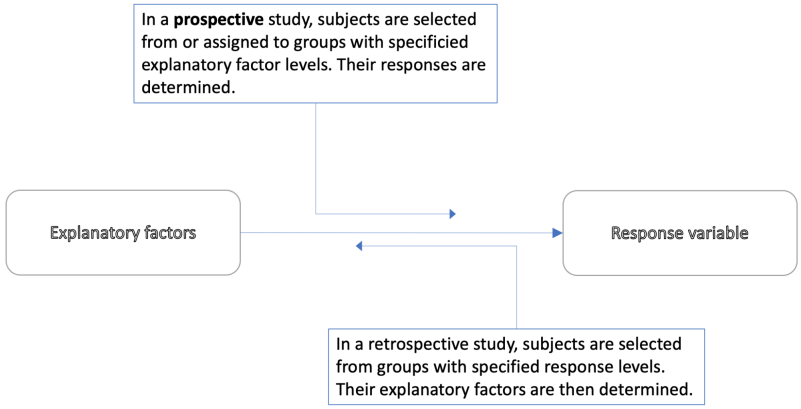


Figure 1.1: Retrospective vs. Prospective observational studies

Observational Studies and Causation

- We can never definitively establish causation with an observational study.
- As more studies are done that adjust for confounding variables, the chance that a lurking variable remains that can explain the association is reduced.
- Unless a controlled experiment is performed, causal inferences are rarely warranted.
- Confounding is the rule, not the exception, unless the assignment to treatment and control is randomized.
- Beware of Simpson's Paradox: an observed association between two variables can change or even reverse direction when there is another variable that interacts strongly with both variables.

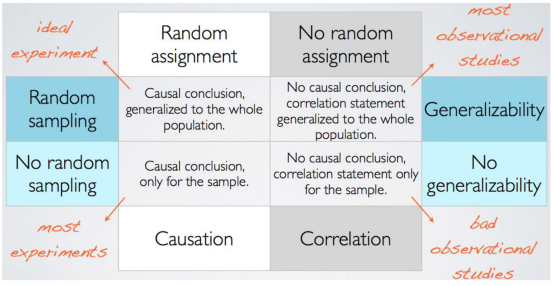
1.1.4 Experiments

When researchers want to investigate the possibility of a causal connection, they conduct an experiment. Usually there will be both an explanatory and a response variable. Experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.
An experiment deliberately imposes treatments on the experimental units (subjects: people, animals or other objects) in order to observe their responses. The goal of an experiment is to investigate the associations, how the treatment affects the response. An advantage of an experimental study over an observational study is that it provides stronger evidence for causation.

Principles of Experimental Design

- **Controlling**: researches assign treatments to cases and do their best to control any other differences in the group. A good experiment has a control comparison group. In human experiments, one way to reducing the bias is to conduct a blind experiment and a better way is to conduct a double-blind experiment.
- **Randomization**: researchers randomize patients into treatment groups to account for variables that cannot be controlled. This tends to balance the comparison groups with respect to confounding variables.
- **Replication**: the more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response.
- **Blocking**: researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into blocks and then randomize cases within each block to the treatment groups.

1.1.5 Random sampling vs Random Assignment Summary



1.2 Confidence Intervals and Significance Tests

Inference is using a data from a sample to make conclusion about the whole population. There are two types of statistical inference methods: confidence intervals for population parameters and testing hypotheses about particular parameter values.

1.2.1 Confidence Intervals (CIs)

A **confidence interval** for a parameter is an interval computed from sample data by a method that will capture the parameter for a specified proportion of all samples. It contains the most believable values for a parameter. The **success rate** is the proportion of all samples whose intervals contain the parameter and is known as the confidence level.
95% confidence interval does not mean that there is a 95% chance the confidence interval around my $\hat{\theta}$ contains the true parameter θ. Instead it means that if we were to take many random samples from the same population and calculate a 95% confidence interval for the parameter of interest in each of those samples, 95 out of 100 confidence intervals would be expected to contain the true value of the parameter.

1.2.2 Bootstrap

The primary purpose of **bootstrapping** is to make inferences about population parameters or the uncertainty of a statistical estimate without making assumptions. We estimate θ, the parameter we are interesting in, using $\hat{\theta}$ which is calculated with the gathered data. The idea is create multiple (often thousands or more) simulated datasets, each of the same size as the original sample, by randomly drawing (with replacement) from the observed data. For each resampled dataset, you calculate the statistic of interest. This gives you a collection of statistics, which can be viewed as a *bootstrap distribution*. You use the information from the bootstrap distribution to make inferences about the population parameter or the uncertainty of your original estimate.

$$\theta = T(X_1, ..., X_n) \tag{1.1}$$

$$f \longrightarrow \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \Longrightarrow \hat{f} \longrightarrow \begin{bmatrix} X^*_{1(\mathbf{1})} \\ X^*_{2(\mathbf{1})} \\ \vdots \\ X^*_{n(\mathbf{1})} \end{bmatrix}, \cdots, \begin{bmatrix} X^*_{1(M)} \\ X^*_{2(M)} \\ \vdots \\ X^*_{n(\mathbf{M})} \end{bmatrix}$$
$$\downarrow \qquad\qquad \downarrow \qquad\qquad\qquad\qquad \downarrow$$
$$\hat{\theta} \qquad\qquad \hat{\theta}^*_{\mathbf{1}}, \qquad \cdots, \qquad \hat{\theta}^*_M$$

Bootstrap distributions: Bias and standard errors

A statistic used to estimate a parameter is biased when its sampling distribution is not centered at the true value of the parameter. The bootstrap estimate of the **bias** is the mean of the bootstrap distribution minus the statistic for the original data.
The bootstrap **standard error** of a statistic is the standard deviation of the bootstrap distribution of that statistic.

1.2.3 Bootstrap confidence intervals

From the bootstrap distribution we can construct confidence intervals. There are different methods.

Interval Using a Bootstrap Standard Error

Suppose that the bootstrap distribution of a statistic from a simple random sample (SRS) of size n is approximately Normal and that the bias is small. An approximate $(1 - \alpha)100\%$ (ideal confidence

interval) for the parameter that corresponds to this statistic is: $statistic \pm t_{n-1,\alpha/2} SE_{boot}$.
Pros: simple to form and easy to understand and can be applied when the standard error of the statistic is difficult to derive.
Cons: can perform poorly if the distribution is highly skewed.

Bootstrap Percentile Interval

The interval between the $Q_{\alpha/2}$ and $Q_{1-\alpha/2}$ percentiles of the bootstrap distribution of a statistic is a ideal bootstrap percentile interval for the corresponding statistic. Use this method when the bootstrap estimate of bias is small. This method doesn't ignore skewness and it is Transformation Invariant.

Reverse Bootstrap Percentile Interval

The reverse bootstrap interval argues that the behavior of $\hat{\theta} - \theta$ is approximately the same as the behavior of $\hat{\theta}^* - \hat{\theta}$. The interval is then:

$$(2 * \hat{\theta} - Q_{1-\alpha/2}, 2 * \hat{\theta} - Q_{\alpha/2}) \tag{1.2}$$

which is the mirror image of the bootstrap percentile interval. Be careful, it is asymmetrical in the wrong direction.

Bootstrap t-interval

Each bootstrap sample is converted into a t-score as follows:

$$t_i^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{SE_{\hat{\theta}_i^*}} \tag{1.3}$$

For $SE_{\hat{\theta}_i^*}$ we use a nested bootstrap algorithm, bootstrap the bootstrap sample, and use the variance of the estimate across bootstrapped samples as the estimate of variance. Compute the $(1-\alpha)100\%$ **studentized t intervals** as:

$$(\hat{\theta} - q_{1-\alpha/2} * SE_{\hat{\theta}}, \hat{\theta} - q_{\alpha/2} * SE_{\hat{\theta}}) \tag{1.4}$$

where q are the percentiles of our nested bootstrap t-estimates t_i^*.
Pros: simple idea with intuitive procedure; works well for location parameters; second-order accurate.
Cons: not transformation invariant; requires iterated bootstrap; doesn't work as well for correlation/association measures.

Bootstrap Bias-Corrected accelarated (BCa) Confidence Interval

BCa intervals use percentiles of the bootstrap distribution, but they don't necessarily use the $Q_{\alpha/2}$ and $Q_{1-\alpha/2}$. Depend on acceleration parameter $\hat{a}$ (estimates the rate of change of the standard error of $\hat{\theta}^*$ with respect to the true parameter θ) and on a bias-correction factor $\hat{z}$ (measures median bias of $\hat{\theta}^*$).

BCa intervals have the form $(\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2})$

- $\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\frac{\alpha}{2}}}{1 - \hat{a}(\hat{z}_0 + z_{\frac{\alpha}{2}})}\right)$
- $\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha})}\right)$
- Φ is the cdf of the normal distribution
- z_α is the 100α-th percentile of the normal distribution.
- If $\hat{z}_0 = \hat{a} = 0$ then $\alpha_1 = \Phi(z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $\alpha_2 = \Phi(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$, the BCa is the same as the percentile interval.

It is Transformation invariant, works well for a variety of parameters, second-order accurate. But requires estimation of $\hat{a}$ and $\hat{z}$.

1.2.4 Hypotheses Tests

Hypothesis tests (or significance tests) are used to investigate claims about population parameters. We use the question of interest to determine the two competing hypotheses: the null hypothesis is generally that there is no effect or no difference while the alternative hypothesis is the claim for which we seek evidence. The null hypothesis (H_0) is the default assumption; we only conclude in favor of the alternative hypothesis (H_1) if the evidence in the sample supports the alternative hypothesis and provides strong evidence against the null hypothesis. If the evidence is inconclusive, we stick with the null hypothesis.
We measure the strength of evidence against the null hypothesis using a p-value. A p-value is a statistical measurement used to validate a hypothesis against observed data. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference.

Hypothesis Tests using Permutation Tests

The key idea behind permutation tests is to use the permutation of data points to generate a null distribution of the test statistic, which can be compared to the observed test statistic to evaluate the hypothesis.

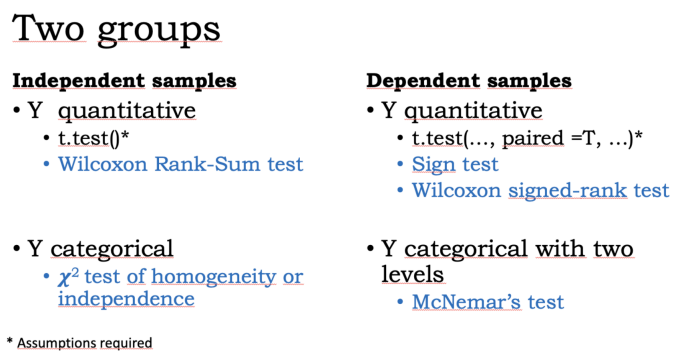
- Formulate Hypotheses: start with the null hypothesis and an alternative hypothesis that you want to test.
- Calculate the Test Statistic: compute the test statistic based on your sample data.
- Permute the Data: construct the sampling distribution that this statistic would have under the null hypothesis by randomly shuffling or permuting the data points.
- Compare and Make a Decision: compare the observed test statistic (calculated from the original, unpermuted data) to the null distribution, find the p-value by locating the original statistic on the permutation distribution. The p-value is the proportion of permuted test statistics that are as extreme as or more extreme than the observed test statistic. Finally make a decision based on the p-value: typically if it is less than your chosen significance level, often 0.05, you may reject the null hypothesis.

1.2.5 Different Distribution Recap

- A **sampling distribution** shows the distribution of sample statistics from a population, and is generally centered at the true value of the population parameter. It is theoretically determined or approximated through statistical theory. It involves repeatedly drawing samples from the population and calculating the statistic of interest for each sample. It helps assess the variability of sample statistics and is often used for making inferences about population parameters.
- A **bootstrap distribution** simulates a distribution of sample statistics for the population, but is generally centered at the value of the original sample statistic. It is created by drawing a large number of bootstrap samples (with replacement) from the observed data and calculating the statistic of interest for each sample. It provides a non-parametric method for estimating the variability of a sample statistic without assuming a specific distribution for the population. It is particularly useful when the underlying population distribution is unknown or complicated.
- A **permutation distribution** simulates a distribution of sample statistics for a population in which the null hypothesis is true, and is generally centered at the value of the null parameter. It is created by permuting the observed values and calculating the statistic of interest for each permutation. This helps create a distribution under the null hypothesis of no effect. It is often used in hypothesis testing to assess whether the observed effect is statistically significant by comparing the observed statistic to the distribution of permuted statistics.

1.3 Methods for Comparing Two Groups

Methods are classified attending to two criteria: the nature of the response variable Y (the variable being compared) and the kind of samples we have collected from the two groups (dependent or independent). In dependent samples observations in each group are meaningfully matched with one another while in independent samples values in one group are not related to or influenced by the values in another group.
For comparing two groups we have the following methods:



1.3.1 Sign Test

The sign test is a non-parametric statistical test used to compare two groups of **paired data** when the data is **quantitative** and their distribution is not assumed to be normal, and you are interested in whether one treatment is more effective than the other.
To perform the sign test, you calculate the differences between the paired measurements, which represent the impact or response to the two different treatments. These differences are obtained by subtracting one treatment's value from the other for each subject or item. The sign test focuses solely on the sign or direction of these differences. It does not consider the magnitude of the differences; it only cares about whether one treatment led to a better response than the other, worse response, or no difference. It ignores observations pairs for which the difference is 0. The null hypothesis is that the median of the differences is 0.
For **Unilateral** tests we are interested in testing the null hypothesis $H_0 : m = m_0$ against the alternative $H_1 : m > m_0$ (or $H_1 : m < m_0$), then, if the alternative hypothesis were true, we should expect the difference $X_i - m_0$ to yield more positive (or negative) signs that would be expected if the null hypothesis were true. In this case, we should reject the null hypothesis if n_- (or n_+), the observed number of negative (or positive) signs is too small, or alternatively, if the p-value define as:

$$P(N_- \leq n_-) \text{ or } P(N_+ \leq n_+) \tag{1.5}$$

In **Bilateral** tests if we are interested in testing the null hypothesis, $H_0 : m = m_0$, against the alternative hypothesis $H_1 : m \neq m_0$. We reject if n_min, which is defined as the smaller of n_- and n_+, is too small. Alternatively, we reject if the p-value as defined by:

$$2P(N_{min} \leq min(n_-, n_+)) \tag{1.6}$$

1.3.2 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a non-parametric statistical test used to compare two groups of **paired data** when the data is **quantitative**, just like the sign test. However, it provides more quantitative information by considering the magnitude of the differences between the paired responses. Similar to the sign test, the Wilcoxon signed-rank test starts by calculating the differences between the paired responses and it ignores observations pairs for which the difference is 0. The key idea is that each **difference** is assigned a **rank** based on its magnitude. The test statistic, denoted as W, is calculated as the sum of the ranks for the differences that are positive. In other words, W is the

sum of the ranks of the differences that indicate an improvement or positive change between the two conditions.
The Wilcoxon signed-rank test assumes that the population distribution of the differences is symmetric. This means that the distribution is roughly bell-shaped and not skewed to one side. The null hypothesis is that the population median of the differences is 0, implying that there is no systematic improvement or deterioration between the two conditions.
The p-value is calculated as $P(W \leq \hat{W})$.

-1	+1	+2	+4	-5	-5	+5	-8	+8	-9
1.5	1.5	3	4	6	6	6	8.5	8.5	10
-11	-12	-15	-16	-18	-21	-22	-25		
11	12	13	14	15	16	17	18		

$$W = 1.5 + 3 + 4 + 6 + 8.5 = 23$$

Figure 1.2: Example: 18 observed differences on blood pressure values after treatment - before treatment. The absolute value of the differences are ordered and ranks are assigned (in red). The mean rank is assigned when ties are present.

1.3.3 Wilcoxon Rank-Sum Test

The Wilcoxon Rank-Sum Test is a non-parametric statistical test used to compare two **independent** groups when the data is **quantitative**. It assesses whether there are statistically significant differences between the distributions of values in the two **independent** groups. The null hypothesis is that the two independent samples come from the same population. The Wilcoxon Rank-Sum Test can be used to test whether the medians (central tendencies) of the two groups are different. However, for this to be a test of medians, it is important that the distributions of values for each group are of similar shape and spread. If the distributions are not similar, the test becomes a test of stochastic (random) equality, which assesses whether one group tends to have higher values than the other, without specifically focusing on the medians.
Procedure:

- Rank all n observations (from both groups).
- The statistic W is the sum of the ranks for the first sample.
- Under the null hypothesis, its mean and standard deviation is

$$\mu_W = \frac{n_1(n+1)}{2} \text{ and } \sigma_W = \sqrt{\frac{n_1 n_2 (n+1)}{12}} \tag{1.7}$$

- The Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions when the calculated rank sum $\hat{W}$ is far from its mean.

The p-value is calculated as $P(W \leq \hat{W})$.

1.3.4 Chi-Squared Test

The Chi-Squared test is a non-parametric test commonly used to compare two **independent** groups when the data is **categorical** and counts how many observations fall into specific categories or *cells*. The chi-squared test is used to assess whether the observed **frequencies in the cells** of a contingency table are significantly different from what would be expected under the null hypothesis of independence. The chi-squared statistic and the associated p-value tell us nothing about the nature or the strength of the association (small p-values doesn't mean strong association).
For this test there's different measure of association in $2x2$ table:

- Difference of proportions.
- Ratio of proportions: Relative Risk or risk ratio. Popular measure for describing the association between a drug and a control treatment in medical studies.

- Odds and Odds ratio: The odds in a given row are defined as the ratio of the two conditional proportions in that row. The odds ratio is then the ratio of the two odds.

The Pearson goodness-of-fit test statistic for the χ^2 test compares the observed counts with the counts expected under the null hypothesis in the following way:

$$T = \sum_{j=0}^{k} \frac{(O_j - E_j)^2}{E_j^2} \tag{1.8}$$

where $O_j = X_j$ is the observed count in category j, and $E_j = np_j$ is the expected count in category j under the assumption that the null hypothesis is true. T measures how closely the model fits the observed data. It has an approximate chi-square distribution with $k - 1$ degrees of freedom when H_0 is true. This allows us to use the chi-square distribution to find critical values and p-values for establishing statistical significance. The p-value is calculated as $P(\chi^2_{k-1} \geq T)$.

1.3.5 McNemar's Test for 2x2 Tables

McNemar's test is a statistical test used to analyze the association or dependency between two **categorical** variables measured on the same subjects or items in a **paired** or matched design. It is particularly useful for situations where you want to determine if there is a significant change or difference between two related categorical variables, often before and after an intervention or treatment. McNemar's test is a non-parametric alternative to the chi-squared test for comparing proportions in such paired data.
The data is organized into a 2x2 contingency table, where the rows represent the two categories of the first measurement (e.g., "yes" and "no"), and the columns represent the two categories of the second measurement (e.g., "yes" and "no"). The table counts how many subjects fall into each of the four possible combinations. The McNemar's test statistic is calculated based on the counts in the 2x2 table. It is computed using the formula: $z^2 = \frac{(b-c)^2}{b+c}$, where b and c are the element in the secondary diagonal. The p-value is calcualated as $P(Z \geq z), Z \approx N(0, 1)$.

Module 2

2.1 Multivariate Descriptive Statistics

2.1.1 Covariance Matrix

A covariance matrix is a square matrix that provides a summary of the covariance between multiple variables in a dataset. It contains the **variance** of each variable in the **diagonal** and **off-diagonal** entries are the **covariances** between any two variables. The covariance measures the **linear dependency** between the observations of each pair of variables. The covariance matrix contains all the information about the spread (variances) and orientation (covariance, linear dependence between each pair of variables) of our data.

$$S_{pxp} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T \tag{2.1}$$

where $\bar{\mathbf{x}}$ is the mean vector, a p-dimensional vector containing averages for every variable.

Properties

Symmetric, positive semi-definite (or nonnegative definite). Non negative trace, determinant and eigenvalues. Its rank determines the number of linearly independent variables in the data matrix and it is also the number of eigenvalues different from 0.
Dimensionality Theorem: Being $r = rank(S) \leq p$ there are r linearly independent variables in the data set. The others $p - r$ are linear combinations of these r variables.

Global Variability Measures

In the context of a covariance matrix S, $|S|$ represents the determinant of the covariance matrix. The determinant of a matrix is a scalar value that can be computed from the elements of the matrix, and it carries important information about the matrix itself.

Generalized variance
$|S|$. It has intuitively pleasant geometrical interpretations.

Effective variance (Peña and Rodríguez, 2003)
$|S|^{1/p}$.

Total variance
It's the trace of S. $T = tr(S) = \sum_{i=1}^{p} s_i^2 = \sum_{i=1}^{p} \lambda_i$.

Mean variance
$\bar{s}^2 = \frac{1}{p}\sum_{i=1}^{p} s_i^2$.

When $|S| = 0$, there are some variables that are linear combinations of the others. Some frequent specific situations would be: two or more variables sum up to a constant; two variables are identical

```
> ese1
    var1   var2   var3   var4
1 0.0947 0.0242 0.0054 0.0594
2 0.0242 0.0740 0.0285 0.0491
3 0.0054 0.0285 0.0838 0.0170
4 0.0594 0.0491 0.0170 0.0543

> det(ese1)
[1] 3.496301e-08

> eigen(ese1)
$values
[1] 1.729668e-01 8.761555e-02 4.616766e-02 4.997203e-05

$vectors
           [,1]       [,2]       [,3]         [,4]
[1,] -0.5937114  0.5270237  0.4509023 -0.407970466
[2,] -0.5046316 -0.2794621 -0.7073918 -0.408466309
[3,] -0.3019646 -0.7930913  0.5289833 -0.000437213
[4,] -0.5492459  0.1230977 -0.1282992  0.816526290
```

Figure 2.1: $|S| = 0$ example

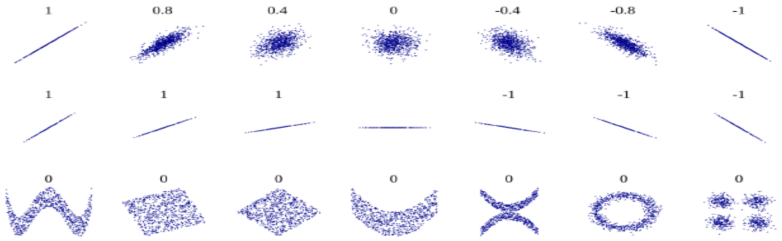


Figure 2.2: Geometric visualization of correlations

or differ merely in the mean or variance.
Whenever a non-zero vector **a** satisfies one of the following conditions, it satisfies all of them:

- **Sa** $= 0$, a is a scaled eigenvector of S with eigenvalue 0. The eigenvector **a** gives coefficients for the linear dependency of the mean corrected data.
- The linear combination of the mean corrected data, using **a**, is 0.
- The linear combination of the original data, using **a**, is constant.

2.1.2 Linear Dependency Measures

Pairwise Linear Dependence - Correlation Matrix

Covariances are difficult to interpret, so often it is useful to work with the correlation, which is always between -1 and 1. A pairwise linear dependence matrix (correlation matrix) R is a matrix that quantifies the linear relationships or dependencies between pairs of variables in a dataset. R is square, symmetric, filled with 1s in the principal diagonal and with pairwise Pearson's correlation coefficients elsewhere.

Coefficient of Determination

Squared multiple correlation coefficient R^2 between a variable X_j and a set of (independent, predictors) variables $(X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)$ is defined in the context of multiple linear regression. It measures the goodness of a linear combination of the independent variables to predict X_j. It is a vector where each elements is a value tell us how well the j-th variable is explained by the others.

$$R^2_{j.1...p} \tag{2.2}$$

Partial Correlation

Partial correlations are a measure of the relationship between two variables while controlling for the effects of one or more other variables. In a pairwise context, it assesses the relationship between two variables while holding all other variables constant. Partial correlations help reveal the unique association between two variables, taking into account the influence of other variables. They are used to examine direct relationships between pairs of variables, removing the effects of confounding variables.

For instance, P matrix for 4 variables would be:

$$P = \begin{pmatrix} 1 & r_{12.34} & r_{13.24} & r_{14.23} \\ r_{21.34} & 1 & r_{23.14} & r_{24.13} \\ r_{31.24} & r_{32.14} & 1 & r_{34.12} \\ r_{41.23} & r_{42.13} & r_{43.12} & 1 \end{pmatrix}$$

where, for example, $r_{12.34}$ is the partial correlation of (X_1, X_2) controlling for X_3 y X_4.

Effective Dependence

It is a multivariate overall measure of linear dependence. It represents the average proportion of explained variability among the variables due to linear dependencies.

$$D(\mathbf{R_p}) = \mathbf{1} - |\mathbf{R_p}|^{\mathbf{1/(p-1)}} \tag{2.3}$$

2.1.3 Distance and Outlier

To obtain a useful distance measurement in a multivariate setting, we must consider not only variances but also their covariances or correlations. **Mahalanobis Distance**: the (squared) distance between two vectors is defined as:

$$D_j = [(\mathbf{x_j} - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x_j} - \bar{\mathbf{x}})]^{1/2} \tag{2.4}$$

This distance can be used to detect outliers. If the data comes from a p-dimensional Normal variate its distribution follows a χ^2_p.

Outliers Identification

In a Univariate settings outliers can be identified using, for example, boxplots.
If data is Multivariate, using Mahalanobis distance or robust estimators.

- Compute, for each observation, the Mahalanobis distance to the mean vector, d_i^2.
- Examine these distances for unusually large values. In a chi-squared plot, these would be the points farthest from the origin. Look for values that exceed the upper fifth percentile of the chi-squared distribution.

2.1.4 Linear Transformation

Two important linear transformations that are often used with multivariate data are:

- Individual standardization of each variable: eliminates the mean and standardises the variance of each variable.
- Multivariate standardization: eliminates the mean and the correlation between the variables and standardises the variance of each one.

Box-Cox Transformation

$$x^{(\lambda)} = \begin{cases} \frac{(x+m)^{\lambda}-1}{\lambda} & (\lambda \neq 0) \quad (x > -m) \\ ln\,(x+m) & (\lambda = 0) \quad (m > 0) \end{cases}$$

with λ in range $[-5, 5]$. The goal is to find the value of λ that maximizes the normality and homoscedasticity of the transformed data.
When interpreting the Box-Cox transformation, you should consider the following: The transformed data may be more suitable for analysis and modeling since it meets the assumptions of normality and constant variance? The choice of λ reflects the degree of transformation applied to the data?

Andrew's Curve

Each multivariate observation, $(x_1, ..., x_p)$, is transformed into a curve:

$$x(t) = \frac{x_1}{\sqrt{2}} + x_2 sin(t) + x_3 sin(t) + x_4 sin(2t) + x_5 sin(2t) + ... \tag{2.5}$$

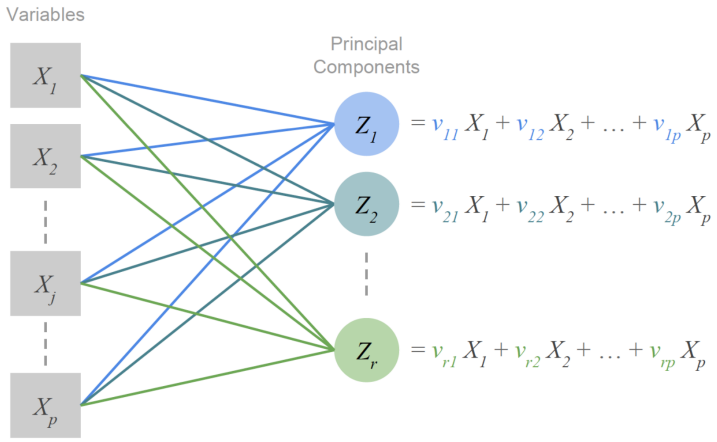
This function is plotted over the range $-\pi \leq t \leq \pi$. A set of multivariate observations will appear as a set of lines on the plot.
The distance between curves reflects correctly the pairwise distances between observations. Outliers appear as single Andrews' curves that look different from the rest. A subgroup of data is characterized by a set of similar curves. The order of variables does affect the appearance of the plot.

2.2 Principal Components Analysis

We are searching for a better low-dimensional summary of the data (more informative than traditional EDA). We are interested in studying resemblances between observations (individuals) and relationship among variables.
With PCA we seek to **reduce the dimensionality** (condense information in variables) of a data set while retaining as much as possible of the variation present in the data. It also allows us to find a graphical representation of the essential information contained in a data set (through Biplots) and to find an optimal approximation of a data set with a minimal loss of information.

2.2.1 PCA as Best Summary of Information

Given a set of p variables $X_1, ..., X_p$, we want to obtain r new variables $Z_1, ..., Z_r$, called the Principal Components (PCs) where $r < p$. We want to compute them as linear combinations of the original variables.



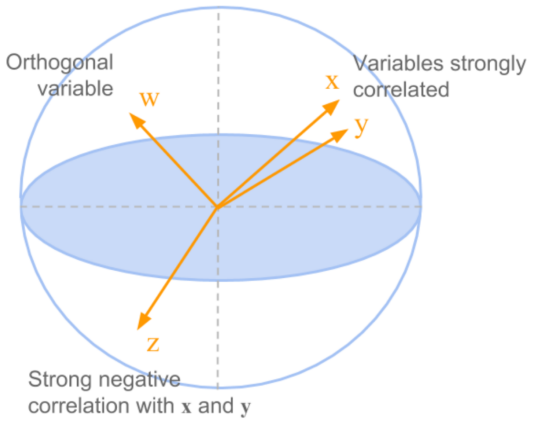
The main requirement for the Principal Components, is that they need to capture the **most variation** in the data X. To avoid a PC capturing the same variation as other PCs (i.e. avoiding redundant information), we may also requiere them to be mutually orthogonal, so they are uncorrelated with each other.
Finding the directions with maximum variance means, algebraically:

- Maximizing a quadratic form that involves the covariance matrix S of the centered data.
- Diagonalizing the covariance matrix (S) of the centered data. This means finding a set of orthogonal eigenvectors (principal components) and eigenvalues that describe the directions and the amount of variance along these directions in the data.
- The eigenvalues $(\lambda_1, ..., \lambda_r)$ of the covariance matrix represent the amount of variance explained by each corresponding principal component. Eigenvalues are positive, and larger eigenvalues indicate principal components that capture more variance in the data.
  Proportion of total Variance due i-th principal component $= \frac{\lambda_1}{\lambda_1 + ... + \lambda_r}$

PCs are just linear combinations of Xs and weights $Z_{n,r} = X_{n,p} V_{p,r}$. The weights V are given by the normalized eigenvectors of $S = X^T X$, say $(\lambda_1, e_1), ..., (\lambda_r, e_r)$, for eigenvalue-vector pairs.

2.2.2 Geometric Apporach

Under a purely geometric approach, PCA aims to represent the cloud of points in a space with reduced dimensionality in an optimal way, that is, making the dispersion of the projected points as large as possible.



2.2.3 PCA as Optimal Approximation of Data

The role of Models in Statistical Analysis, $Data = Structure + Residual \rightarrow X_{n,p} \approx Z_{n,r}V_{r,p}^T + Residual = \hat{X}$.

2.2.4 PCA Output

From any PC analysis:

- Eigenvalues provide information about the amount of variability captured by each principal component
- Scores or PCs provide coordinates to graphically represent objects in a lower dimensional space.
- Loadings provide the correlation between variables and components. They often help to interpret the components as they measure the contribution of each individual variable to each principal component. However, they do not indicate the importance of a variable to a component in the presence of others.
- Eigenvectors are unit-scaled loadings. They are the coefficients of orthogonal transformation (rotation) of variables into PC's.

Components to retain: Look for an elbow in the curve plotted in the scree plot. This point is considered to be where large eigenvalues cease and small eigenvalues begin. Kaiser's rule, which consists of retaining those PCs with eigenvalues $\lambda_k > 1$. Jollife's rule, in which we retain those PCs with eigenvalues $\lambda_k > 0.7$.

Module 3

3.1 Intro

A good statistical model is one in which f is a parsimonious function that helps us explain how Y is related to X.

$$Y = f(X;\theta) + \epsilon \tag{3.1}$$

How do we define what a good model is?

- A model that fits the data well (minimize resubstitution error)
- A model with optimal parameters (most likely coefficients)
- A model that adequately predicts new (unseen) observations (minimize generalization error)

Typically we use the third one (good predictions), this involves finding a measure of accuracy for predictions. We will use typically the **Root Mean Square Error** (RMSE) with linear models.

$$RMSE(\hat{f}, X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2} \tag{3.2}$$

The *bias-variance* trade-off refers to the idea that you can't have low bias and low variance at once. A model that overfits has low bias and high variance.

3.2 Statistical Regression Models

data = model + residual, the residual should not contain any additional pattern or structure. If it does contain additional structure, then the model associated with the data needs refinement.

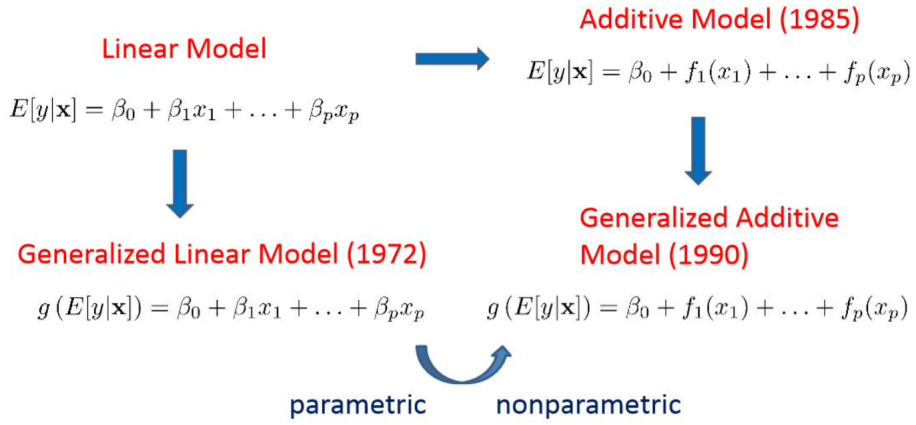


Figure 3.1: Types of Models

3.2.1 Simple Linear Regression Model

$$y = \beta_o + \beta_1 x + \epsilon,\ \epsilon \sim IIDN(0, \sigma^2) \tag{3.3}$$

For each x there is a (hypothetical) distribution of y values with means linearly related to x and constant variance. We obtain estimates for the model parameters: $\hat{\beta_0}$, $\hat{\beta_1}$ and $\hat{\sigma^2}$.
Parameter interpretation: $\hat{\beta_0}$ the intercept, is the value of the response for X = 0. Most of the time it has not a meaning or sensible interpretation itself; $\hat{\beta_1}$ the slope, represents the variation in the response Y when X is increased (or decreased) one unit.
Least Squares Estimation is a powerful tool for estimating the parameters of a regression model but it is limited in terms of providing information about the variability of errors. The residuals play a key role in understanding this variability and in checking the validity of model assumptions. Maximum Likelihood Estimation builds upon these assumptions, particularly normality, to provide a more comprehensive statistical understanding of the model, enabling inferences and predictions.
To use the model for inference and prediction, we need to impose four conditions that comprise the simple linear regression model (**LINE**).

- **L**inearity: the mean of the response $E(Y_i|X_i)$ at each value of the predictor is a linear function of the X_i.
- **I**ndependence: the residuals (errors) e_i are independent.
- **N**ormality: the residuals, for each value of the predictor X_i are normally distribuited.
- **E**qual variances (σ^2): the residuals, at each value of the predictor X_i have equal variances.
- Validity of the data for answering the research question.

Violation of most of these assumptions are not fatal and can be fixed by improving your model, by picking different or additional variables or using a different distribution or modeling framework. Residual plots are the best tool for assessing whether model assumptions have been satisfied.

3.2.2 Multiple Linear Regression Mode

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e \tag{3.4}$$

There are multiple predictors (independent vars., x) of varying kinds and a single outcome (dependent variable, y).

Statistical Assumptions

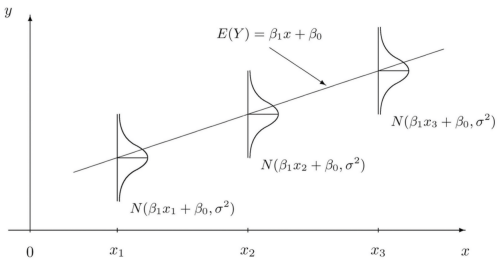
Normality (and Linearity): the conditional distribution of Y for any combinations of values of the $X_1, \ldots, X_p$ is Normal. The expected value is a linear function of the $X's$,

$$(Y|X_1 = x_{i1}, \ldots, X_k = x_{ik}) \approx N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \sigma^2)$$

$$e_i \approx N\left(0, \sigma^2\right)$$

Independence. The observations from Y are statistically independent.

Homocedasticity. The conditional variance of Y given any specific combination of values of the $X_1, \ldots, X_p$ is the same, i.e., σ^2.



Coefficient Interpretation: Each β_j is a partial regression coefficient, reflecting the expected change in the independent variable per unit change in the $j-th$ independent variable, assuming all other independent vars. are held constant i.e., with all other predictors held fixed. The coefficients of a multiple regression must not be interpreted marginally! They have to be interpreted as a ceteris paribus relationship between the Y and each X_j.

Regression Test

After fitting a model to a set of data it is necessary to asses the adequacy of the fit. From the general assumptions made on the error terms (and on the distribution of Y) different tests are developed. Is at least one of the predictors $X_1, X_2, ..., X_k$ useful in predicting the response? Do the Xs help to explain the y?

$$H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0 \tag{3.5}$$

vs

$$H_1 : \exists \beta_i \neq 0, \forall i \tag{3.6}$$

Failing to reject H_0, maybe due to no linear relationship between the response variable and the explanatory variables. Rejecting H_0 means the model with some (or all) predictors is better to explain the output Y than a model with just the mean of the y scores.

t Test

```
> summary(modelo3)

Call:
lm(formula = X2 ~ X3 + X4 + X5, data = diabetis, x = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-39.223 -10.646   0.927   8.559  71.432

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.172321   4.135995   6.328 3.14e-09 ***
X3           0.193768   0.007941  24.400  < 2e-16 ***
X4          -0.039134   0.013483  -2.902   0.0043 **
X5          -0.013996   0.022427  -0.624   0.5336
---
Signif. codes:  0 Â‘***Â’ 0.001 Â‘**Â’ 0.01 Â‘*Â’ 0.05 Â‘.Â’ 0.1 Â‘ Â’ 1

Residual standard error: 16.28 on 140 degrees of freedom
Multiple R-squared: 0.937,      Adjusted R-squared: 0.9356
F-statistic:   694 on 3 and 140 DF,  p-value: < 2.2e-16
```

The intercept and coefficients for vars. $X3$ and $X4$ may be considered different from 0. Variable $X5$ may be dropped out, given the others in the model.

Scenario	F Test	specific tests
1	Significant	Every one Significant
2	Significant	Some of them Significant
3	Significant	None Significant
4	Not Significant	All of them Significant
5	Not Significant	Some of them Significant
6	Not Significant	None Significant

- Scenarios 1 and 2: proceed with the analysis, dropping out variables if needed.

- Scenarios 3, 4 and 5: Multicollinearity problems.
- Scenario 6: No linear relationships are detected between the variables involved in the model.

3.2.3 Multiple Correlation

The correlation between the predicted scores ($\hat{Y}$) and the observed (criterion) scores (Y) is called the multiple correlation coefficient, R. It falls between 0 and 1, it cannot be negative. R^2 measures the proportion of the variance of the dependent variable about its mean that is explained by the independent, or predictor, variables. Low values may be explained because important variables have been left out of the model.

$$R^2 = \frac{SSReg}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \tag{3.7}$$

where: n is the number of observations, $\hat{y}_i$ is the predicted value for the i-th observation, y_i is the actual observed value for the i-th observation, $\bar{y}$ is the mean of the observed data.

R^2 properties

- R^2 falls between 0 and 1. The larger the value, the better the explanatory variables collectively predict Y.
- $R^2 = 1$ only when all residuals are 0, that is, when all regression predictions are perfect ($Y = \hat{Y}$).
- $R^2 = 0$ when all $\hat{Y} = \bar{Y}$. In that case, the estimated slopes all equal 0 and the correlation between Y and each explanatory variable equals 0.
- R^2 gets larger or, at worst, stays the same, whenever an additional explanatory variable is added to the multiple regression model.
- The value of R^2 doesn't depend on the units of measurement.

Adjusted R^2

As we have seen, the addition of independent variables will always cause R^2 to rise or, at least, stay the same. A modified measure, adjusted R2, takes into account the number of independent variables included and the sample size. R_a^2 pays a price for the inclusion of unnecessary variables in the model.

$$R_a^2 = R^2 - \frac{k(1 - R^2)}{n - k - 1} \tag{3.8}$$

k is the number of predictors.

3.2.4 Diagnostics and Validation

Regression model building is an iterative process. The first model we try may prove to be inadequate.

Residual Plots

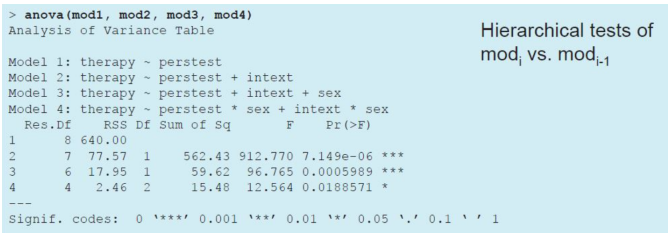
Partial residual plots (one for each regressor): show the relationship between the response variable and the corresponding explanatory variable, adjusted for the other variables in the model. They consider the marginal role of regressor X_j given other regressors that are already in the model. If they show a curvilinear band, then higher order terms in the explanatory variable X_j or a transformation may be useful. They can be used to explore explanatory variables whose effect is expected to be small relative to the effect of others. They help to find influential observations in the estimate of model coefficients.

3.2.5 ANOVA and ANCOVA

ANOVA (ANalysis Of VAriance) is a statistical method used to test the differences between the means of three or more independent (unrelated) groups. It is particularly useful when you want to determine whether there are any statistically significant differences between the means of several different groups. ANCOVA (ANalysis of COVAriance) is an extension of ANOVA that adds one or more continuous variables that are not of primary interest but may influence the response variable. These continuous variables are called covariates and are included in the model to statistically control or adjust for their effects.

3.2.6 Nested Model

One model is nested within another if it is a special case of the other in which some model coefficients are constrained to be zero. When two models are nested multiple regression models, there is a simple procedure for comparing them. This procedure tests whether the more complex model is significantly better than the simpler model. In the sample, of course, the more complex of two nested models will always fit at least as well as the less complex model. This is done via Partial F-tests (follow a F distribution).



The R output shown in the image is from an ANOVA (Analysis of Variance) comparing four different regression models to assess how well they explain the variability in a dependent variable. Each model includes different predictors or combinations of predictors. The output suggests that each additional predictor and the interaction terms contribute significantly to the model, with each subsequent model providing a better fit to the data than the last.

3.3 Regression Diagnostics

Model Adequacy Checking

Two important concepts regarding the application of any regression model in practice, in particular the linear regression model:

- Correctness. The linear model is built on certain assumptions. All the inferential results are based on these assumptions being true. A model is formally correct whenever the assumptions on which is based are not violated in the data.
- Usefulness. The usefulness is a more subjective concept but is usually measured by the accuracy on the prediction and explanation of the response Y by predictors $X_1, ..., X_p$.

Model adequacy checking based on residual analysis:

- Basic graphical inspection: Basic residuals scatterplots, Normal probability, plots residuals vs. fitted values, Influence plots
- Statistical inspection: check normality (Jarque-Bera test prefereable), check constant variance (Breusch-Pagan test for Heteroscedasticity), check for independence (Durbin-Watson test for autocorrelation of residuals), Influence observation measures, Multicollinearity measures.

If homocedasticity (homogeneity of variances) or independence are not fulfilled, least square estimators are still linear and unbiased, but are not the best. If autocorrelation is present in the residuals, the t and F tests can lead to wrong inferences. Although the linear model is fairly robust to the normality

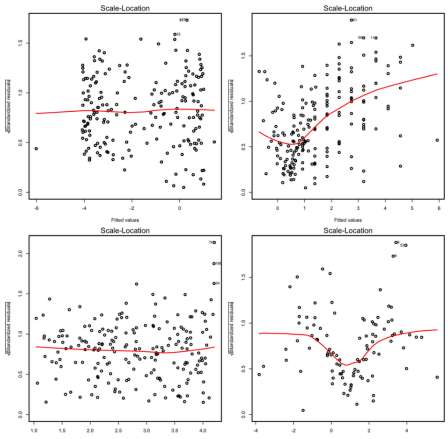


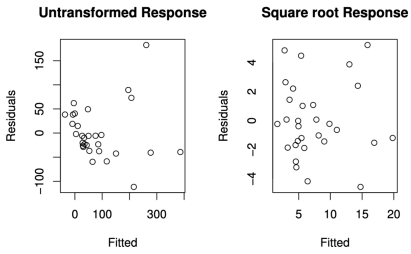
Figure 3.2: Left: valid residual plots. Right: violation of assumptions.

assumption, if normality is not adequate, the tests will only be approximately valid and it also poses problems of efficiency of estimates.

3.3.1 Ways to Treat Model Inadequacies

Transformations to stabilize variance or to achieve a linear relationship

Transformation on y: Box-Cox transformation, use the spread level plot to find a power transformation. Transformation on the regressor variables.



Model Extensions

Extensions of the estimation method: Generalized and weighted least squares (non constant variance, autocorrelation), Robust Regression (outliers, influential observations), Ridge regression partial least squares (multicollinearity).

3.3.2 Multicollinearity

Multiple regression attempts to separate out the effects of each independent variable holding the others constant. If there are important linear relationships between the X variables, we have multicollinearity. If the goal is to use a regression model to predict Y, then multicollinearity is not problematic, as long as the model accounts for a high proportion in the variation in Y , the predictions should be quite accurate. If we wish to understand how each of the explanatory variables impacts Y, multicollinearity is potentially problematic. Overall, model may be highly significant while no (or few) individual predictors are.

Detecting Multicollinearity

There's different ways:

- Start by looking at the pairwise correlations between Xs (0.8 or higher are of concern).

- Multiple correlations coefficients R^2, regressing each X on the others.
- Variance influence factors (VIFs)
- Eigenvalues from Principal Component Analysis. Perfect collinearity has an eigenvalue of 0.

The Variance Inflation Factor, for each variable X_i, is:

$$VIF_i = \frac{1}{1 - R^2_{i.others}} \tag{3.9}$$

where $R^2_{i.others}$ is the multiple correlation coefficient R^2, regressing each X on the others. $VIFs > 10$ indicate high multicollinearity. The square root of the VIF tells the factor by which the standard error and confidence interval is inflated because of multicollinearity. For instance, $\sqrt{VIF_i} = 20$ tells us that the standard error of coefficient β_i is 20 times higher than it would have been without collinearity. It indicates the impact of multicollinearity on the precision of β_i.
The Condition Index or Condition Number is based on eigenvalues of S or R (for the matrix of variables X's). Defined by:

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{3.10}$$

There can be defined one CI_i for each individual eigenvalue λ_i

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_i}} \tag{3.11}$$

$CI_i < 10$ are ok, between 10 and 30 we have to watch out and $CI_i > 30$ are trouble until reaching 100 that are disastrous.

3.3.3 Handling Multicollinearity

We could eliminate some highly correlated variables, Principal Component Regression, Ridge Regression.

Variable Selection Methods

Forward Stepwise Selection (FSS) begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.
Backward Stepwise Selection (BSS) begins with the full least squares model containing all p predictors, and then, iteratively, removes the least useful predictor, one-at-a-time.
Select a single best model from among using some criteria.

3.3.4 Influential Observations

A Regression Outlier (high residual) is an observation that has an usual value of Y, given its value of X. Neither the X nor the Y values are necessarily unusual on their own. They usually have large residuals but do not necessarily affect the regression slope coefficient.
A Leverage is an unsual X value, far from the mean of Xs, has a leverage on the regression line. The further it sits from the mean of X, the more leverage it has. However, high leverage does not necessarily mean that it influences the regression coefficients (good leverage points).
An Influential observations is an observation with high leverage that is also a regression outlier and it will strongly influence the regression line. The line (plane, hyperplane) chases the observation. $Influence = Xleverage \times Yresidual$. An influential point is one whose removal from the data set would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of those two properties.

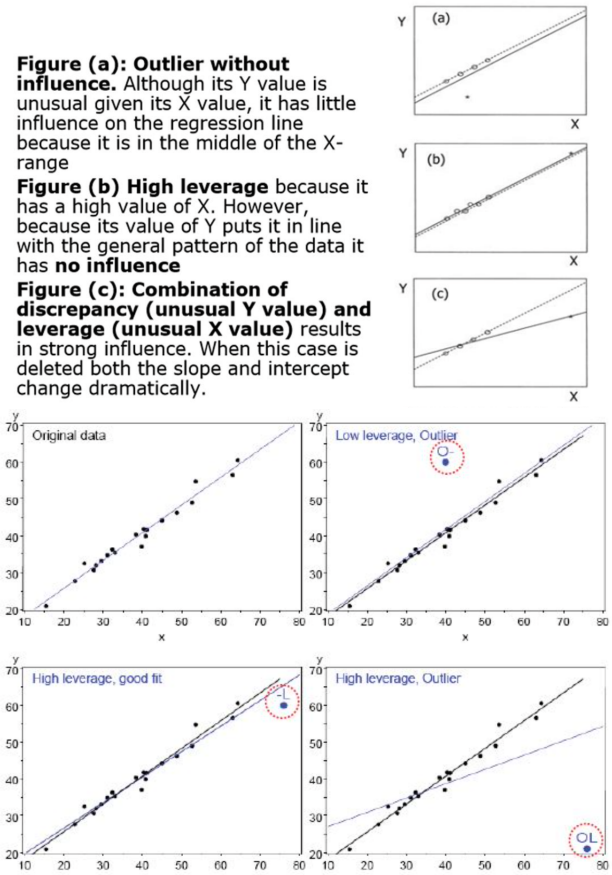


Figure (a): Outlier without influence. Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X-range

Figure (b) High leverage because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has **no influence**

Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value) results in strong influence. When this case is deleted both the slope and intercept change dramatically.

Unusual observations may reflect miscoding, then the observations can be rectified or deleted completely.

Outliers are sometimes of substantive interest: if only a few cases, we may decide to deal separately with them. Several outliers may reflect model misspecification, i.e., an important variable that accounts for the subset of the data that are outliers has been neglected.

Use methods that downweight outliers (robust methods). Often, these methods results to estimates similar to a model omitting influential cases, because they assign very low weight to highly influential cases.

Sensitivity analysis: compare results with and without the observations and analyse the effect on conclusions.

Module 4

4.1 Time Series - Basic Concepts and Models

4.1.1 Basics

Time series data are the result of observing the same phenomenon over regular time intervals. Formally, a time series $x_1, x_2, ..., x_T$ is the finite (univariate) realization of a stochastic process X_t. This stochastic process $X_t : t = 1, 2, ..., T$ is a sequence of random variables. At time t, there are $E(X_t)$, $V(X_t)$, $Cov(X_t, X_{t+k})$.
Time series analysis accounts for the fact that data points taken over time may have an internal structure (autocorrelation, trend or seasonal variation) that should be taken into account. Knowledge of the dynamic structure helps to predict better future observations and to design optimal control structures. We can study time series from two different perspective:

- **Descriptive**: time series analysis. Describe the patterns in the time series; quantify its components (patterns and noise).
- **Predictive**: time series forecasting. Predict future values of the series, create forecasts.

Graphical Representation

A particularity of time series is that the usual numerical summaries are hardly ever used. Mean, median, standard deviation, etc. do not make sense for most of the time series. We want to visualize them.
We are not able to statistically analyze any time series, because in every moment we have a random variable with a different distribution (each with its parameters), from which we have only one observed data.
What to look for in a first inspection? The main sources of non stationarity are:

- Is there a **trend**, meaning that, on average, the measurements tend to increase or decrease over time?
- Is there **seasonality**, meaning that there is a regularly repeating pattern of highs and lows related to the calendar time such as seasons, quarters, months, days of the week and so on?
- Does it seem to be a cyclical component, regular rises and falls but with a longer period that seasonal ones? This cycles usually have not a fixed period but can vary in length throughout time.
- Is there constant variance over time or is the variance non constant?

4.1.2 Classical Regression Models

Basic Decomposition

$$ObservedValues = Trend + Seasonality + IrregularComponent \tag{4.1}$$
$$Y_t = T_t + S_t + I_t \tag{4.2}$$

where:

- Trend: long term increase or decrease in the series.
- Seasonality: patterns linked to the calendar, for instance, patterns linked to the time of the year.
- Irregular component: random fluctuations around the previous components.

Time series and autocorrelation

Observations in time series tend to be correlated. The data has memory, observations today are affected by what happened in the past. Autocorrelation is the correlation coefficient between the value of the time series at time t and its value at time $t-1$. This is going to be r_1 or the autocorrelation coefficient at lag 1. But the memory of the data might go past observation at time $t-1$. We can compute autocorrelations at different lags, using, every time less pairs of observations.
The Partial Autocorrelation is also computed at different lags, but without considering the effect of the intermediate observations.

Univariate Models of time Series

$$Y_t = f(Y_{t-1}, Y_{t-1}, ...) + a_t \tag{4.3}$$
$$Y_t = Y_t^* + a_t \tag{4.4}$$

Being Y_t^* the systematic component, it is possible:

- To learn its structure from the data, **Data-driven Methods**. Exponential smoothing methods and decomposition methods: the algorithms learn patterns from the data; they are useful when model assumptions are likely to be violated; they require less input from the user, they are "user-proof"; they are preferable for series with local patterns.
- To propose a structure using a stochastic model, **Model-based Methods**. AR, MA, ARMA, ARIMA: they use a statistical, mathematical or other scientific model to approximate the time series; they are especially useful for short time series; they are preferable for forecasting series with global patterns; they are suitable to compute prediction intervals.

4.1.3 Forecasting

Depending on the goal (analysis vs. forecasting) model validation will be different: Residual analysis vs Evaluating forecasting accuracy through different metrics (RMSE). Should be careful of common machine learning issues as overfitting.

Validation

The forecast error (residual) for time period t, denoted e_t, is defined as the difference between the actual value y_t and the forecast value at time t, F_t:

$$e_t = y_t - F_t = y_t - \hat{y}_{t|t-1} \tag{4.5}$$

Considering a test period of v observations, a popular measure of predictive accuracy is the root mean square error, RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{v} e_i^2}{v}} \tag{4.6}$$

Residual Analysis

A good forecasting method will yield residuals with the following properties: the residuals are uncorrelated, if there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts. The residuals have zero mean.
In addition to these essential properties: the residuals have constant variance and they are normally distributed.

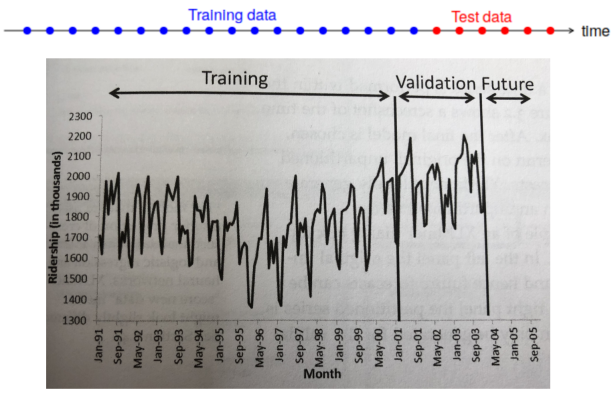


Figure 4.1: Fixed partition

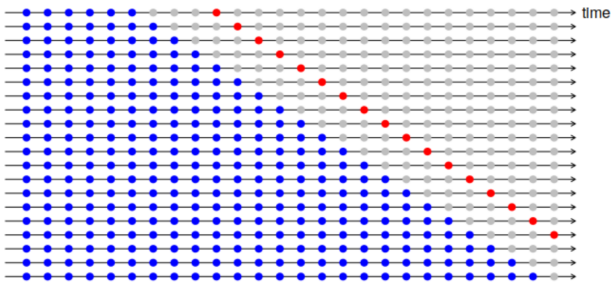


Figure 4.2: Evaluation on a rolling forecasting origin, using 4-step-ahead forecasts.

4.1.4 Smoothing Methods

In time series analysis, a smoothing method is used to reduce noise and reveal patterns (trend, seasonality, etc.) in the data, such as trends and seasonality. The main idea behind smoothing is to create a smoothed or less noisy version of the original time series, which can help understand the underlying structure of the data and make better forecasts. It can be a good first step in describing various components of the series. It is a way to extrapolate, in a way that recent data points have more weight in forecasting than older values. They are used mainly for prediction. Models have deterministic parameters that should be calibrated to suit the evolution of the series. Fitting the model is made by recursive methods.

Simple exponential smoothing: series with no trend or seasonality. It uses one parameter $0 \leq \alpha \leq 1$.

Holt double exponential smoothing: series with trend and no seasonal variation. It uses two parameters α, β. If α and $\beta \approx 0$ the evolution of the time series is near constant. If $\beta \approx 0$ slope is near constant.

Holt-Winters triple exponential smoothing: series with linear trend and seasonal variation. It uses three parameters α, β and γ.

Figure 4.3: Smoothing Methods

Exponential Smoothing

Simple Exponential Smoothing: The idea is to forecast future values using a weigthed average of all previous values in the series. It is suitable for series with no trend or seasonality. The key assumption in simple exponential smoothing is that the time series does not exhibit a trend or seasonality. It assumes that the time series is essentially a flat line (the level) with random noise around it.

For $t = 1, ..., T$, the smoothed values, the forecasts at each time value are obtained recursively:

$$\hat{y}_{t+1|t} = \alpha y_t + (1 - \alpha)\hat{y}_{t|t-1} \tag{4.7}$$

These are one-step-ahead forecasts of the training data. The process starts at L_0 which must be given and recursively the rest of the values are obtained. Then we obtain:

$$\hat{y}_{T+1|T} = \sum_{j=0}^{T} \alpha(1 - \alpha)^j y_{T-j} + (1 - \alpha)^T L_0 \tag{4.8}$$

For T large, the last term is very small. This gives the weighted average form: forecasts are calculated using weighted averages where the weights decrease exponentially as observations come from further in the past. The smallest weights are associated with the oldest observations.

Forecast

Estimated level at most recent time point.

$$F_{T+k} = L_T$$

$$\hat{y}_{T+k|T} = \hat{y}_{T+1|T} = L_T, k = 2, 3, \ldots$$

Updating equations

The level updating equation:

$$L_t = \alpha Y_t + (1 - \alpha)L_{t-1}$$

The level updating equation can be rewritten as (error correction form):

Error correction form

$$\begin{aligned} L_t &= L_{t-1} + \alpha(Y_t - L_{t-1}) \\ &= L_{t-1} + \alpha e_t \end{aligned}$$

The training data errors (in red) lead to the adjustment/correction of the estimated level (in blue) throughout the smoothing process. With this formulation, simple exponential smoothing can be seen as an adaptive learning process where α controls the degree of learning.

Table 7.1: Forecasting the total oil production in millions of tonnes for Saudi Arabia using simple exponential smoothing.

Year	Time	Observation	Level	Forecast
	t	y_t	ℓ_t	$\hat{y}_{t\|t-1}$
1995	0		446.59	
1996	1	445.36	445.57	446.59
1997	2	453.20	451.93	445.57
1998	3	454.41	454.00	451.93
1999	4	422.38	427.63	454.00
2000	5	456.04	451.32	427.63
2001	6	440.39	442.20	451.32
2002	7	425.19	428.02	442.20
2003	8	486.21	476.54	428.02
2004	9	500.43	496.46	476.54
2005	10	521.28	517.15	496.46
2006	11	508.95	510.31	517.15
2007	12	488.89	492.45	510.31
2008	13	509.87	506.98	492.45
2009	14	456.72	465.07	506.98
2010	15	473.82	472.36	465.07
2011	16	525.95	517.05	472.36
2012	17	549.83	544.39	517.05
2013	18	542.34	542.68	544.39
	h			$\hat{y}_{T+h\|T}$
2014	1			542.68
2015	2			542.68
2016	3			542.68
2017	4			542.68
2018	5			542.68

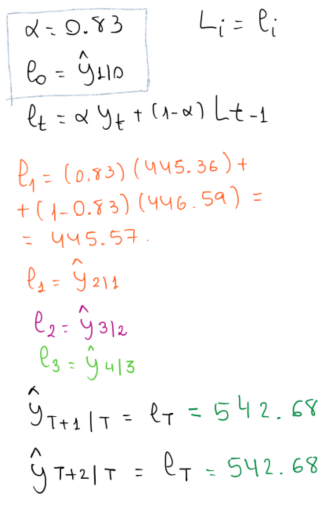
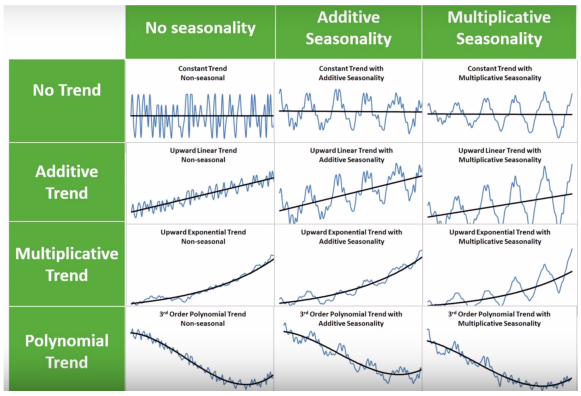


Figure 4.4: Simple exponential Smoothing Example.

Advanced Exponential Smoothing The idea is to extend SES to capture trend and/or seasonality. Let's see the types of trends and seasonalities that are popular approximations for time series:



Holt's Exponential Smoothing (or double ES): Holt's method can capture additive trend or multiplicative trend. This trend can vary adaptively over time. The smoothing constants (α, β) control the speed of learning (for level and trend). Software implementation matters: default values, different initializations, different optimization algorithms. It is simple and cheap to compute. Easy to automate.
The main assumption is that the series has only level (L_t), trend (T_t) and noise (unpredictible).

Forecast (additive model)

Estimated level + trend at most recent time point.

$$F_{T+k} = L_T + kT_T$$

Updating equations

The level updating equation and the trend updating equation:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

Error correction form

$$\begin{aligned} L_t &= L_{t-1} + T_{t-1} + \alpha e_t \\ T_t &= T_{t-1} + \alpha\beta e_t \end{aligned}$$

Table 7.2: Applying Holt's linear method with $\alpha = 0.8321$ and $\beta^* = 0.0001$ to Australian air passenger data (millions of passengers).

Year	Time	Observation	Level	Slope	Forecast
	t	y_t	ℓ_t	b_t	$\hat{y}_{t\|t-1}$
1989	0		15.57	2.102	
1990	1	17.55	17.57	2.102	17.67
1991	2	21.86	21.49	2.102	19.68
1992	3	23.89	23.84	2.102	23.59
1993	4	26.93	26.76	2.102	25.94
1994	5	26.89	27.22	2.102	28.86
1995	6	28.83	28.92	2.102	29.33
1996	7	30.08	30.24	2.102	31.02
1997	8	30.95	31.19	2.102	32.34
1998	9	30.19	30.71	2.101	33.29
1999	10	31.58	31.79	2.101	32.81
2000	11	32.58	32.80	2.101	33.89
2001	12	33.48	33.72	2.101	34.90
2002	13	39.02	38.48	2.101	35.82
2003	14	41.39	41.25	2.101	40.58
2004	15	41.60	41.89	2.101	43.35
2005	16	44.66	44.54	2.101	44.00
2006	17	46.95	46.90	2.101	46.65
2007	18	48.73	48.78	2.101	49.00
2008	19	51.49	51.38	2.101	50.88
2009	20	50.03	50.61	2.101	53.49
2010	21	60.64	59.30	2.102	52.72
2011	22	63.36	63.03	2.102	61.40
2012	23	66.36	66.15	2.102	65.13
2013	24	68.20	68.21	2.102	68.25
2014	25	68.12	68.49	2.102	70.31
2015	26	69.78	69.92	2.102	70.60
2016	27	72.60	72.50	2.102	72.02
	h		ℓ_T	b_T	$\hat{y}_{T+h\|T}$
	1				74.60
	2				76.70
	3				78.80

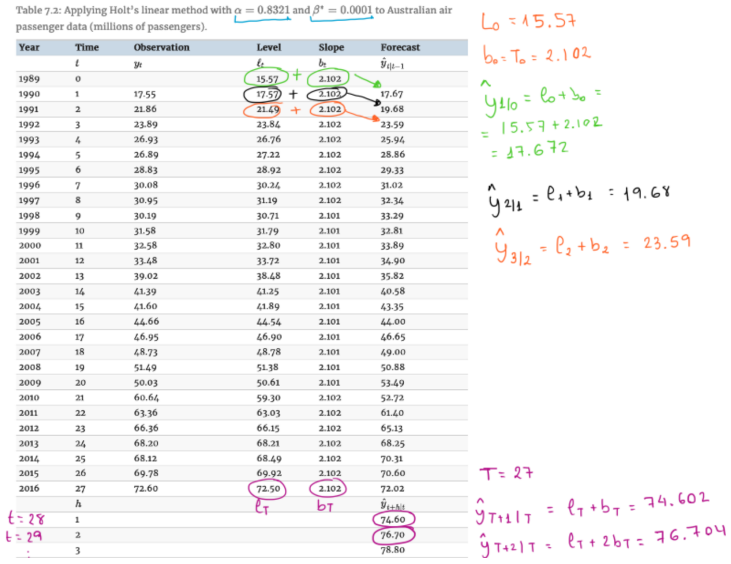


Figure 4.5: Holt's exponential Smoothing example

Holt-Winter's Exponential Smoothing (or triple ES): Holt-Winter's method can capture local additive or multiplicative trend and/or seasonality. The smoothing constants (α, β, γ) control the speed of learning (level, trend and seasonality). Software implementation matters: default values, different initializations, different optimization algorithms. It is simple and cheap to compute. Easy to automate.
The main assumption is that the series has level (L_t), trend (T_t), seasonality with M seasons and noise (unpredictible).

Forecast (additive trend and seasonality)

Estimated level + trend + seasonality at most recent time point.

$$F_{t+k} = L_t + kT_t + S_{t+k-M}$$

Forecast (additive trend and multiplicative seasonality)

$$F_{t+k} = (L_t + kT_t) \cdot S_{t+k-M}$$

Updating equations

The level updating equation, (additive) trend updating equation and (multiplicative) seasonality:

$$L_t = \alpha \frac{Y_t}{S_{t-M}} + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \gamma \frac{Y_t}{L_{t-1} + T_{t-1}} + (1-\gamma)S_{t-M}$$

4.1.5 From Methods to Models

Exponential smoothing methods can be rewritten into a fully specified statistical model using their error correction form, together with a statistical distributions on the errors. These are called innovations

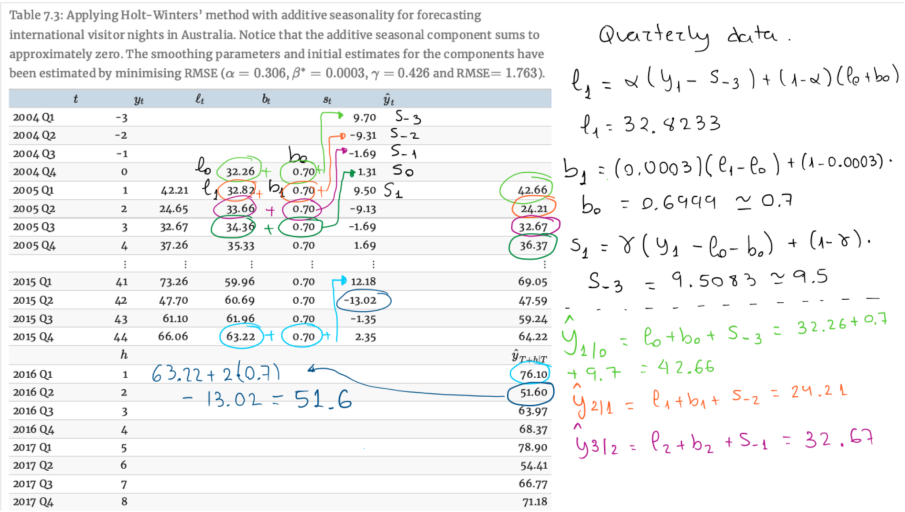


Table 7.3: Applying Holt-Winters' method with additive seasonality for forecasting international visitor nights in Australia. Notice that the additive seasonal component sums to approximately zero. The smoothing parameters and initial estimates for the components have been estimated by minimising RMSE ($\alpha = 0.306, \beta^* = 0.0003, \gamma = 0.426$ and RMSE= 1.763).

	t	y_t	ℓ_t	b_t	s_t	$\hat{y}_t$
2004 Q1	-3				9.70	
2004 Q2	-2				-9.31	
2004 Q3	-1				-1.69	
2004 Q4	0		32.26	0.70	1.31	
2005 Q1	1	42.21	32.82	0.70	9.50	42.66
2005 Q2	2	24.65	33.66	0.70	-9.13	24.21
2005 Q3	3	32.67	34.39	0.70	-1.69	32.67
2005 Q4	4	37.26	35.33	0.70	1.69	36.37
	⋮	⋮	⋮	⋮	⋮	⋮
2015 Q1	41	73.26	59.96	0.70	12.18	69.05
2015 Q2	42	47.70	60.69	0.70	-13.02	47.59
2015 Q3	43	61.10	61.96	0.70	-1.35	59.24
2015 Q4	44	66.06	63.22	0.70	2.35	64.22
	h					$\hat{y}_{T+h\|T}$
2016 Q1	1					76.10
2016 Q2	2					51.60
2016 Q3	3					63.97
2016 Q4	4					68.37
2017 Q1	5					78.90
2017 Q2	6					54.41
2017 Q3	7					66.77
2017 Q4	8					71.18

Figure 4.6: Holt-Winter's exponential Smoothing example

state-space models. These statistical models generate the same point forecasts as the methods already considered, but they can also generate prediction (or forecast) intervals. Even when you have formulas, if your series violate every kind of assumption, you must check the residuals!

4.1.6 Decomposition Methods

Addictive decomposition: $x_t = Trend + Seasonality + Irregular = T_t + S_t + I_t$.
Multiplicative decomposition: which is equivalent to the additive decomposition after Logarithmic transformation. $x_t = T_t \times S_t \times I_t$

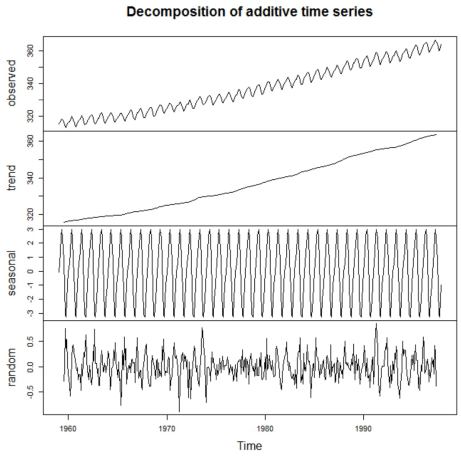


Figure 4.7: Using decompose() function in R: classical decomposition.

4.2 Time Series Statistical Models - Basics, Stationarity and Autocorrelations

4.2.1 Stationary Time Series

A time series is strictly stationary if the joint probability distribution of the observations stay the same across time.

$$P(Y_{t+1}, ..., Y_{t+k}) = P(Y_{t+1+s}, ..., Y_{t+k+s}), \forall t > 0, k > 0, s > 0 \tag{4.9}$$

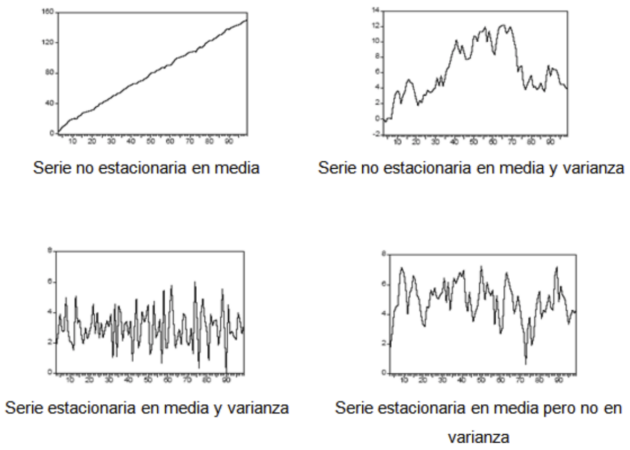


Figure 4.8: Top left: series non stationary in mean. Top right: series non stationary in mean and variance. Bottom left: series stationary in mean and variance. Bottom right: series stationary in mean but not in variance.

Strict stationarity is a strong requirement and often too stringent for practical purposes because it implies that every aspect of the probability distribution of the series must be constant over time, including mean, variance, and all autocorrelations for any number of lags.
A time series is considered weakly stationary (or second-order stationary) if it meets the following three conditions:

- Constant Mean: The mean of the series is constant over time. This means that the expected value of the time series at any point does not depend on the time at which it is measured.
- Constant Variance: The variance of the series is constant over time, meaning that the time series does not exhibit periods of varying volatility. The series will fluctuate within a consistent range of values.
- Constant Autocorrelation Structure: The autocovariance (and thus the autocorrelation) between two time points only depends on the distance or lag between those two points and not on the actual time at which the measurements are taken. In other words, the correlation between observations a fixed number of periods apart remains the same regardless of the time at which the correlation is computed.

Weak stationarity is a less restrictive assumption than strict stationarity and is more commonly used in time series analysis because it only requires constancy in the first two moments and a consistent autocorrelation structure. It does not require higher moments, such as skewness and kurtosis, to be constant. Most time series modeling techniques, such as ARIMA models, assume weak stationarity because it is usually sufficient for modeling and forecasting purposes. If a time series is not weakly stationary, it may need to be transformed (e.g., by differencing or taking a logarithm) to become stationary before applying these models.
Summarizing: In Stationary Time Series the mean, variance and autocorrelation structure do not change over time. This means a flat looking series, without trend, with constant variance over time, with constant autocorrelation structure and no periodic fluctuations (seasonality). The future is similar to the past, in a probabilistic sense.
In Non-Stationary Time Series the mean, variance and/or the relationships between equally spaced data change over time. Seasonal fluctuations are considered a kind of non-stationarity.

4.2.2 Transformations to Achieve Stationarity

If the series presents non-constant variance, increasing with the level of the series: log transformation. If the series presents a linear trend: take one regular difference. If the series presents a quadratic trend: take two regular differences. If the series presents a seasonal component: take seasonal differences.

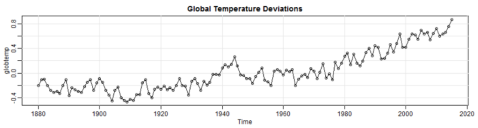
Differencing

Differencing is a method used in time series analysis to transform a non-stationary time series into a stationary one. The main idea is to compute the differences between consecutive observations in the series, which often helps to stabilize the mean of the time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

- First order differences: this involves subtracting the previous observation from the current observation, it is useful for removing a trend. $d_t^1 = y_t - y_{t-1}$
- Second Differencing: if the time series still exhibits non-stationarity after first differencing, second differencing may be applied. This involves differencing the first-differenced series, can help to remove a quadratic trend. $d_t^2 = d_t^1 - d_{t-1}^1$
- Seasonal difference: this involves subtracting the observation from the same season in the previous cycle, useful for removing seasonality. For monthly data with a yearly cycle, you would subtract the value from 12 months ago. $d_t^{12} = y_t - y_{t-12}$

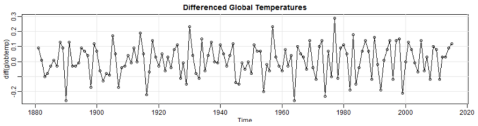
X_t

```
globtemp
```



$X_t - X_{t-1}$

```
diff(globtemp)
```



4.2.3 Time Series and Autocorrelation

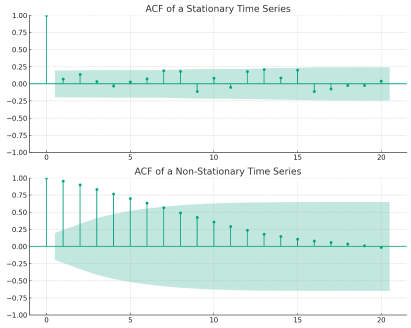
Observations in time series tend to be **correlated**. The data has memory, observations today are affected by what happened in the past.

Correlogram

A correlogram, also known as an autocorrelation plot, is a visual tool used in time series analysis to show the correlation of a time series with itself at different lags. It helps to identify patterns in the data, such as seasonality or autoregressive behavior. Remember that lag means how far apart are the values we are using to compute the correlation.
The correlogram of a stationary time series goes to zero quickly.

- Slow decrease of autocorrelations indicates that series is non stationary.
- Slow decrease of autocorrelations in the seasonal lags indicates that series is non stationary with seasonality. Values in the same season tend to be correlated.
- For non-stationary data, the value of the autocorrelation coefficient at lag-1 is often large and positive.
- A large positive spike at lag-1 is called stickiness. It means that high values follow high values and low values follow low values.
- A large negative spike at lag-1 is called swings. It means that the series swings between high and low values.

We are going to use ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots to help us check whether a series is stationary. Once we have a stationary series we are going to look for the remaining patterns of information.



ACF of a Stationary Time Series: In the top plot, the ACF of a stationary series is shown. You'll notice that the autocorrelations quickly drop off to near zero and stay within the confidence interval bounds (the blue shaded area). This behavior is typical of a stationary series, where autocorrelations do not persist at higher lags. The lack of a significant pattern or trend in the autocorrelations and their diminishing values suggest that the time series does not depend heavily on its past values and is relatively stable over time.
ACF of a Non-Stationary Time Series: The bottom plot displays the ACF of a non-stationary series. Here, the autocorrelations decrease very slowly as the lags increase. This slow decay is indicative of a non-stationary series, often seen in processes like a random walk. The persistent autocorrelation over several lags suggests a strong dependency on past values and indicates that the mean and variance of the series are not constant over time.

4.3 Time Series - ARIMA Models

Auto-Regressive Integrated Moving Average (ARIMA) models are, in theory, the most general class of models for forecasting a time series which can be made to be **stationary** by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging.

4.3.1 Autoregressive Process

The idea of AR models is to capture autocorrelation in a series in a regression-type model and use it to improve short-term forecasts. It is similar to linear regression where the predictors are lagged versions of the series. It is estimated by the maximum likelihood method. They are used when the stationary assumption is met.
An autoregressive (AR) model is a type of statistical model used in time series analysis, where the current value of the series is explained as a function of past values. It's based on the concept that past values in the series can provide information about future values.
AR model of order p, AR(p):

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + e_t \tag{4.10}$$

X_t is the time series and e_t is white noise (random drawings for the same distribution: 0 mean and constant variance).

ACF and PACF for AR

For an AR process, the ACF gradually decreases and may oscillate, reflecting the sustained influence of past values, while the PACF shows significant correlations up to the order of the AR process and then abruptly becomes insignificant. These patterns are used to identify the appropriate AR model for a given time series.

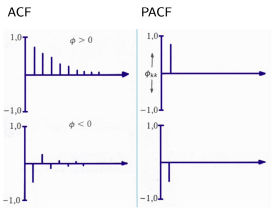


Figure 4.9: ACF and PACF for AR(1)

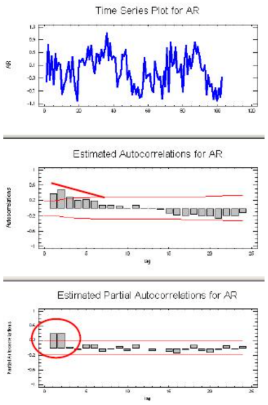


Figure 4.10: AR signature

4.3.2 Moving Average Process

A Moving Average (MA) process in time series analysis is a model where the current value of the series is defined as a linear combination of past error terms. The model is used to describe time series data that exhibit patterns not captured by an autoregressive (AR) model. Each value X_t is a weighted average of the past q forecast errors.
Moving Average Models, $MA(q)$, include lagged terms of the residuals or noise (or innovations). Current value of X can be found from past errors $e_{t-1}, ..., e_{t-q}$ (unevenly weighed) plus a new error. The time series is regarded as a moving average of a random error series e_t. Don't confuse this with moving average smoothing!

$$X_t = \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} + e_t \tag{4.11}$$

The MA signature in the ACF and PACF plots helps in identifying MA processes. The ACF's sharp cutoff helps determine the order of the MA process (the number of lagged error terms), while the PACF's gradual decline provides evidence of the indirect effects of past errors on the current value.

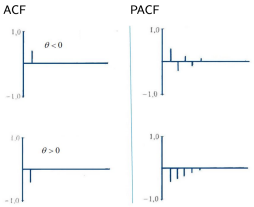


Figure 4.11: ACF and PACF for MA(1)

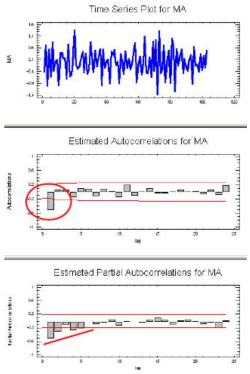


Figure 4.12: MA signature

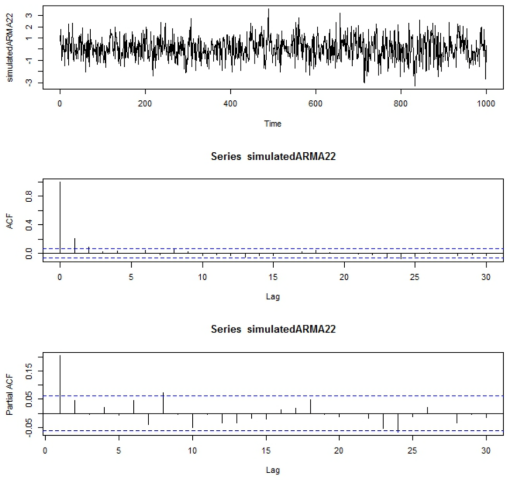
Wether a series displays AR or MA behaviour often depends on the extent to which it has been differenced. An underdifferenced series has an AR signature (positive autocorrelation). After one or more orders of differencing, the autocorrelation will become more negative and a MA signature will emerge. Don't go too far: if a series already has zero or negative autocorrelation at lag 1, don't difference again.

4.3.3 Autoregressive Moving Average Process

It is a combination of AR and MA models. Current observation linearly depends on the last p observations and on the last q error terms (also called innovations).

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_1 p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} \tag{4.12}$$

Predictors include both lagged values of X_t and lagged errors.
For ARMA models the ACF plot presents exponential decay from lag $p + q - 1$, or damped sine. The PACF has exponential decreasing or sinusoidal pattern. Diminishing slowly. ARMA models (including both AR and MA terms) have ACFs and PACFs that both tail off to 0. These are the trickiest because the order will not be particularly obvious. Basically, you just have to guess that one or two terms of each type may be needed and then see what happens when you estimate the model.



4.3.4 ARIMA

ARIMA stands for Autoregressive Integrated Moving Average. It is a class of statistical models for analyzing and forecasting time series data. ARIMA models are capable of capturing a range of different standard temporal structures in time series data. An ARIMA model is described by three terms p, d, and q:

- p: The number of autoregressive terms (AR part).
- d: The degree of differencing (the number of times the data have had past values subtracted to make the series stationary).
- q: The number of moving average terms (MA part).

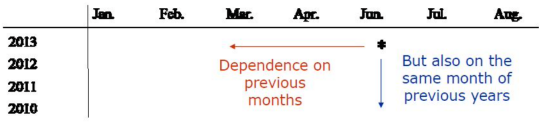
The model is generally referred to as ARIMA(p, d, q). Here's a brief explanation of each component:

- $AR(p)$ - Autoregressive: This part involves using the dependencies between an observation and a certain number of lagged observations (the p past values).
- $I(d)$ - Integrated: To make the time series stationary, which means that its statistical properties such as mean, variance, and autocorrelation are constant over time, the data may need to be differenced one or more times (d times). This differencing step is a way of transforming a non-stationary series into a stationary one.
- $MA(q)$ - Moving Average: This aspect models the error term as a linear combination of error terms at various times (q previous error terms).

The ARIMA model can be used for both fitting and forecasting time series data. Once an ARIMA model is fitted to a time series, it can be used to forecast future values. The ARIMA model is flexible and can include only the AR part (making it an AR model), only the MA part (an MA model), or both.

Seasonal ARIMA Models

The monthly, quarterly, or daily series usually have dependency with previous observations, but also with those ocurred a year or a week ago. Given this seasonal dependence, the autocorrelation function of the models shows a long time dependence that will lead to models with high values of p and q (high number of parameters). Models will be simplified using Multiplicative Models.



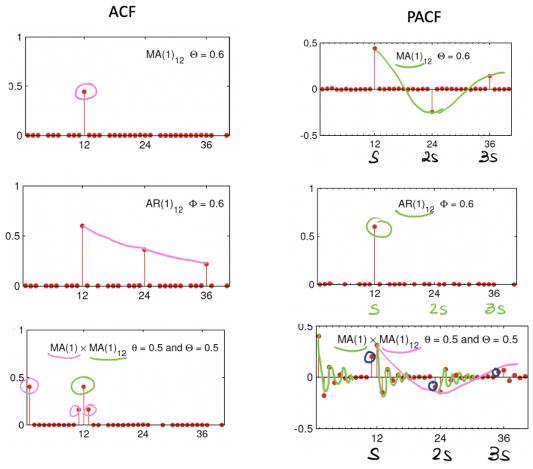
The same principles as for non-seasonal models, except focused on what happens at multiples of lags s in ACF and PACF.

Multiplicative ARMA Models

They are considered for series with trend and seasonality. In general, a time series may show: Regular dependence, described by a regular ARMA model, Seasonal dependence, described by a seasonal ARMA model. Multiplicative ARMA model multiplies the regular structure by the seasonal one, using less parameters. The multiplicative form implies that the effects of past values and past errors are not simply added together but are instead combined in a way that the impact of one component can amplify or diminish the impact of another.

$$X_t = (RegularComponent) \times (SeasonalComponent) \times e_t \tag{4.13}$$

e.g. $MA(1)MA(1)_{12}$.
ACF (autocorrelation function): In the first lags, the regular part is observed. In seasonal lags, the seasonal part is observed. Around the seasonal lags, it is shown the repetition of the regular part of the autocorrelation function on both sides of each seasonal lag. Specifically, if the regular part is a moving average of order q, on both sides of each non-null seasonal lag there will be q coefficients different from 0. If the regular part is autoregressive, we will observe the decreased imposed by the AR structure on both sides of the seasonal lags.
PACF (Partial autocorrelation function): The PACF of a multiplicative seasonal process is complex because it depends on the PACF of the regular and seasonal parts as well as on the ACF of the regular part. In the first lags, the PACF of the regular part is observed and in the seasonal lags, the PACF of the seasonal part appears. To the right of each seasonal coefficient, the PACF of the regular part will appear. If the seasonal coefficient is positive, the regular PACF appears inverted in sign. It it is negative, the regular PACF appears with its sign. To the left of the seasonal coefficients, we observe the ACF of the regular part.



Multiplicative Seasonal ARIMA Model

A Multiplicative Seasonal ARIMA model, often abbreviated as SARIMA or Seasonal ARIMA, extends the ARIMA model to account for seasonality in time series data. This model is designed to capture both the non-seasonal (ARIMA) and seasonal components of a time series.
A Multiplicative Seasonal ARIMA model is typically denoted as ARIMA(p, d, q)(P, D, Q)s:

- p: number of autoregressive terms (AR) for the non-seasonal component.
- d: number of regular differences to obtain stationarity.
- q: number of moving average terms (MA) for the non-seasonal component.
- P: number of seasonal autoregressive terms.
- D: degree of seasonal differencing required to make the series stationary.
- Q: number of seasonal moving average terms.
- s: seasonality period (e.g., 12 for monthly data with an annual cycle).

Often, you find that the correct order of differencing is $d = 1$ and $D = 1$. With one difference of each type, the autocorrelation is often negative at both lag 1 and lag s. This suggests an $ARIMA(0, 1, 1)(0, 1, 1)$ model, a common seasonal ARIMA model. It is similar to Holt-Winter's model in estimating a time-varying trend and a time-varying seasonal pattern.
When fitting a time series with a strong seasonal pattern, you generally should try:

- $ARIMA(0, 1, q)(0, 1, 1)$ model $q = 1, 2$
- $ARIMA(p, 0, 0)(0, 1, 1)$ model $p = 1, 2, 3$

Seasonal ARIMA models compare favorably with other seasonal models and often yield better short-term forecasts.
Advantages: solid underlying theory, stable estimation of time-varying trends and seasonal patterns, relatively few parameters.
Drawbacks: no explicit seasonal indices, hard to interpret coefficients or explain how the model works, danger of overfitting if not used with care.