CEE224A Transportation Data Analysis I

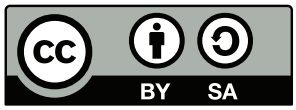
University of California, Irvine

Winter Quarter 2019

Contents

Preface

This scriptum is based on the notes taken in the lecture *CEE224A Transportation Data Analysis I* by Professor Jean-Daniel Saphores, held at the Institute of Transportation Studies at University of California in Irvine in the winter quarter 2019. This course is designed as an introduction to statistical and econometric tools for analyzing datasets for applications in the field of transportation science. It is a graduate level course with basic knowledge of probability and statistics as prerequisites. The lecture is on based on the literature recommended. The notes have been extended with further information, references and explanatory figures. Even though this notes have been gleaned with utmost care by Jens Frische, it is highly unlikely that there are no mistakes in this script. If you are aware of one or want to extend the scope further, you can find all necessary LaTeX source files here: `https://github.com/soaringhigh/NotesUciCee224`



Recommended Literature

Washington, S., Mannering, F. L., & Karlaftis, M. G. (2011). *Statistical and econometric methods for transportation data analysis* (Second edition). Boca Raton, FL: CRC Press.

Angrist, J. D., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton Univ. Press.

Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics* (Third edition update). The Pearson series in economics. Hoboken, New Jersey: Pearson.

Kennedy, P. (2008). *A guide to econometrics* (6. ed.). Malden, Mass.: Blackwell.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (Sixth edition). Boston, MA: Cengage Learning.

Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge: Cambridge University Press.

Conover, W. J. (1999). *Practical nonparametric statistics* (3. ed.). Wiley series in probability and statistics Applied probability and statistics section. New York, NY: Wiley.

Recommended Journals

Journal of Transport and Land Use.
Transportation Research Part A-F.
Transportation.
Transport Review.

I. Descriptive Statistics

Descriptive statistics give us tools to summarize and interpret data.

Definition I.1: Statistics

Analyzing characteristics of a sample to interpret characteristics of the population.

Definition I.2: Probability Theory

Knowing characteristics of a population to find the probability for a sample.

Why is it important to know your sample? Because knowing the variability in the dataset allows to find abnormalities. Such abnormalities can either be invalid data records which need to be filtered out, or remarkable observations pointing towards unexpected subpopulations, which deserve to learn more about them.

I.1. Point Estimates

I.1.A. Central tendency

Definition I.3: Arithmetic Mean

Consider a sample $x_1, x_2, ..., x_n$ with the size n from a population of the size N.

$$\text{Sample mean:} \quad \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{Population mean:} \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The *arithmetic mean* is very common because it is easy to calculate, but sensitive to extreme values which can easily influence the value in their direction. A remedy to this problem is the usage of the *median*.

Definition I.4: Median

In an ordered sample ($x_1 \leq x_2 \leq ...x_n$), the median is $x_{int\frac{1}{2}}$ such that half of the observations are larger and the other half are smaller values. It is identical with the 50^{th}-percentile.

Another useful measure is the *mode*. If the distribution of a sample data is symmetric and unimodal (thus, has one mode), then there is a unique mode which is equal to the mean and equal to the median.

Definition I.5: Mode

The mode is the value that occurs the most frequently.

I.1.B. Central relative standing

Consider an ordered sample $x_1 \leq x_2 \leq ...x_n$. We can characterize the concept of the median to calculate percentiles. For example, the 10^{th}-percentile is such that 10% of the observations are smaller and and 90% are larger. The goal is to get information about the distribution of the data.

I.1.C. Measures of variability

Measures of variability quantify the dispersion of the data (typically around the mean).

Definition I.6: Range

The range is the difference between the highest and lowest value.

$$x_{max} - x_{min}$$

Definition I.7: Interquartile Range

The interquartile range is the difference between the 3rd and the 1st quartile.

$$x_{75} - x_{25}$$

Definition I.8: Variance

The variance is the average sum of the squares of deviations to the mean.

$$\text{Sample variance:} \quad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n-1}$$

$$\text{Population variance:} \quad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Empirical experience has shown that dividing by $n-1$ gives better results for the variance with small sample sizes. The expressions $\bar{X}$ and s^2 are *estimators* for the population mean μ and variance σ^2.

Definition I.9: Estimator

An estimator is a rule which tells how to combine and conclude sample data to obtain estimate a population characteristic.

Definition I.10: Standard Deviation

The standard deviation is the square root of the variance.

$$\text{Sample standard deviation:} \quad s = \sqrt{s^2}$$

$$\text{Population standard deviation:} \quad \sigma = \sqrt{\sigma^2}$$

Definition I.11: Empirical Rule

The empirical or 68–95–99.7 rule states the approximated percentage of values that

lie within a certain range of the mean for data that is normal distributed.

$$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.6827 = 68\%$$
$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545 = 95\%$$
$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973 = 99.7\%$$

Definition I.12: z-score

A sample $x_1, x_2, ..., x_n$ has the normalizing z-scores $z_1, z_2, ..., z_n$.

$$z_i = \frac{x_i - \bar{X}}{s}$$

Example I.1: Expected Value of Random Variables

Consider a sample $x_1, x_2, ..., x_n$ as a random variable coming from $x \Rightarrow E(x) = \mu$, $var(x) = \sigma^2$.
Then z is also a random variable: $z \equiv \frac{x-\mu}{\sigma}$.

$$E(z) = E(\frac{x-\mu}{\sigma}) = \frac{1}{\sigma}(E(x) - \mu) = 0$$
$$var(z) = var(\frac{x-\mu}{\sigma}) - \frac{1}{\sigma} var(x - \mu).$$

For the random variable y: $var(y) = E((y - E(y)^2))$.

$$\begin{aligned} var(y + \alpha) &= E\left((y + \alpha) - E(y + \alpha))^2\right) \\ &= E\left((y - E(y))^2\right) \\ &= var(y) \end{aligned}$$

Definition I.13: Chebyshev's Theorem

For any sample, the proportion of observations whose z-score has an absolute value $z \leq k\, \epsilon\, \mathbb{N}^+$ is no less (so greater or equal) than $1 - \frac{1}{k^2}$.

$$k = 1 \longrightarrow 1 - \frac{1}{k^2} = 1 - \frac{1}{1^2} = 0 \longrightarrow \text{no statement. Empirical rule: } 68\%$$
$$k = 2 \longrightarrow 1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = \frac{3}{4} \longrightarrow \geq 75\%. \text{ Empirical rule: } 95\%$$
$$k = 3 \longrightarrow 1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = \frac{8}{9} \longrightarrow \geq 88.9\%. \text{ Empirical rule: } 99.7\%$$

Named after the Russian mathematician Pafnuty Chebyshev, the theorem allows to make a statement about how much values are within an certain range. This information is less specific than that coming from the empirical rule, but Chebyshev's theorem is applicable to *any* sample, regardless of its distribution whereas the 68–95–99.7 rule strictly requires normal distributed data.

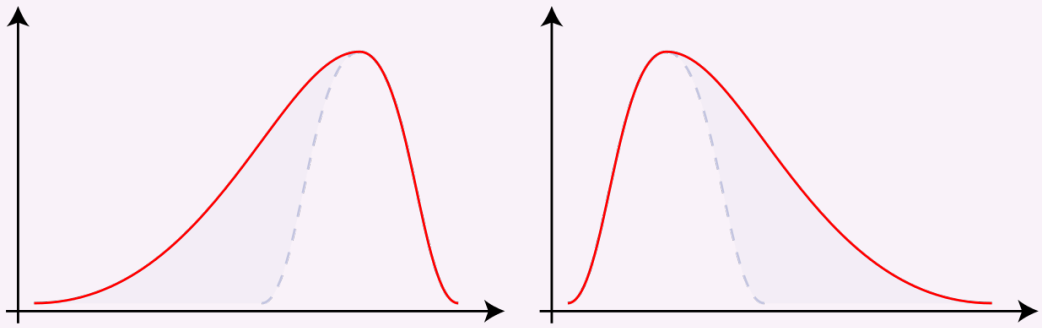
I.1.D. Measures of the shape of a distribution

Definition I.14: Skewness

The skewness is a measure for the degree of asymmetry of a distribution.

$$\text{Population skewness: } \gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E\left[(X-\mu)^3\right]}{\left(E\left[(X-\mu)^2\right]\right)^{3/2}}$$

$$\text{Sample skewness estimator: } g_1 = E = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2\right)^{\frac{3}{2}}}$$



For small sample sizes (but at least $x \geq 3$), the MATLAB-method delivers better estimates.

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

The skewness is the 3rd normalized moment when the distribution is seen as an area. The variance as shown in definition I.8 then is the 2nd central moment, whereas the mean shown in definition I.3 is the 1st moment, i.e. the center of gravity.

Definition I.15: Kurtosis

The kurtosis (from the Greek word for *curved*) is a measure about how heavy the tails of a distribution are compared to the center around the mean.

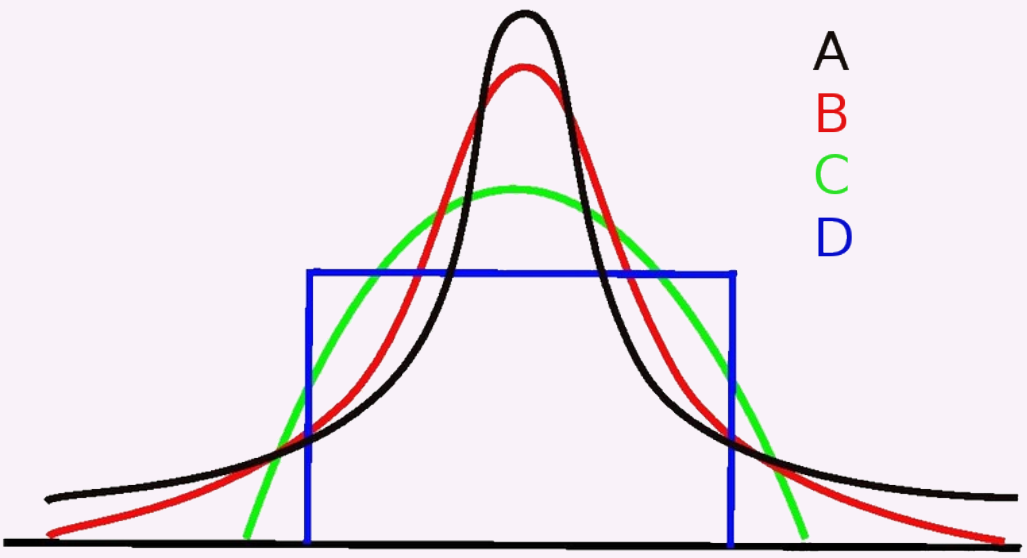
$$\text{Population kurtosis:} \quad \kappa = E\left(\left(\frac{x-\mu}{\sigma}\right)^4\right)$$

$$\text{Sample kurtosis:} \quad k_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Sample kurtosis estimator with bias correction:

$$k_0 = \frac{n-1}{(n-2)(n-3)}\left((n-1)k_1 - 3(n-1)\right) + 3$$

The estimator requires: $n \geq 4$, $\quad \lim_{n\to\infty} \frac{(n-1)(n+1)}{(n-2)(n-3)} = 3$



The graphic shows a normal distribution with different kurtosis, graph A has the highest.

The kurtosis gives us information on how prone to outliers a distribution is. Following the scheme of the skewness (I.14) as the 3^{rd} moment, the kurtosis is the 4^{th} moment of a distribution. A normal distribution has the kurtosis $\kappa = 3$.

I.1.E. Measures of association

Often, we are interested in more than one variable in a time. The *covariance* tells how two variables are related when one changes. The measure is not normalized and has an unit, so mostly informative is the sign. With a positive covariance, an increase in one value corresponds with a greater value for the other one, the covariance is positive. If the values change in different directions, the covariance is negative.

Definition I.16: Covariance

Consider two random variables x, y.

$$\text{Sample covariance:} \quad cov(x,y)_s = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{y})}{n-1}$$

$$\text{Population covariance:} \quad cov(x,y)_p = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

$$\mu_x = E(x),\ \mu_y = E(y)$$

To address the problem that the covariance only indicated the direction in which to variables are related, it can be normalized to state the degree of linear association. This measure is called the *correlation*.

Definition I.17: Correlation

Consider two random variables x, y with the covariance $cov(x,y)$.

$$\text{Sample correlation:} \quad r = \frac{cov_s(x,y)}{s_x s_y}$$

$$\text{Population correlation:} \quad \rho = \frac{cov_p(x,y)}{\sigma_x \sigma_y}$$

$$|r| \leq 1,\ |\rho| \leq 1$$

Example I.2: Showing that the correlation is not greater than 1

Let $a_i = x_i - \bar{X}$ and $b_i = y_i - \bar{y}$. Let $f(z) = \sum(a_i z + b_i)^2 \geq 0$. Expand:

$$f(z) = z^2(a_1^2 + a_2^2 + \cdots + a_N^2) + z(2a_1b_1 + 2a_2b_2 + \cdots + 2a_Nb_N) + (b_1^2 + b_2^2 + \cdots + b_N^2).$$

Consider $f(z) = 0$. There is no solution unless $a_i z + b_i = 0 \; \forall \; i, z$.

$$\alpha x^2 + \beta x + \gamma = 0$$

$$\Delta = \beta^2 - 4\alpha\gamma$$

$$\text{Cases} \begin{cases} \Delta > 0, & x^* = \frac{-\beta \pm \sqrt{\Delta}}{2\alpha} \\ \Delta = 0, & x^* = \frac{-\beta}{2\alpha}, \text{ two equal square root solutions} \\ \Delta > 0, & \text{no real solution} \end{cases}$$

In this case,

$$\Delta = 4\left(\sum a_i b_i\right)^2 - 4\left(\sum a_i^2 \sum b_i^2\right) \leq 0.$$

This is necessarily since $f(z) \geq 0$. Hence,

$$\left(\frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}\right)^2 \leq \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N} \frac{\sum_{i=1}^{N}(y_i - \mu_y)^2}{N}$$

$$\left|\frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}\right| \leq \sigma_x \sigma_y$$

$$\left|\frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}\right| \leq 1$$

$$\implies |\rho| \leq 1$$

Famous Statisticians i: John Graunt

John Graunt (April 24, 1620 – April 18, 1674) was *one of the first demographers*, though by profession he was a haberdasher (i.e., he sold small articles for sewing, such as buttons, ribbons, zips).

He was born in London, the eldest of seven or eight children of Henry and Mary Graunt. His father was a draper who had moved to London from Hampshire. In February 1641, John Graunt married Mary Scott, with whom he had one son and three daughters.

He worked in his father's shop until his father died in 1662, and became influential in the City. He served in various ward offices, becoming a common councilman about 1669–71, warden of the Drapers' Company in 1671 and a major in the trained band.

His house was destroyed in the Great Fire of London and he encountered other financial problems leading eventually to bankruptcy. His daughter became a nun in a Belgian convent and Graunt decided to convert to Catholicism at a time when Catholics and Protestants were struggling for control of England and Europe, leading to prosecutions for recusancy. He died of jaundice and liver disease at the age of 53.

With William Petty, *he developed early human statistical and census methods that later provided a framework for modern demography.* He is credited with producing the first life table, giving probabilities of survival to each age. In addition, he is considered one of the first experts in epidemiology, since his famous book was concerned mostly with public health statistics.

His book "Natural and Political Observations Made upon the Bills of Mortality" (1663) used analysis of the mortality rolls in early modern London as Charles II and other officials attempted to create a system to warn of the onset and spread of bubonic plague in the city. Though his system was not fully created, Graunt's work resulted in the first statistically based estimation of the population of London.

He presented his work to the Royal Society and was subsequently elected a fellow in 1662 with the endorsement of the King.

Source: https://en.wikipedia.org/wiki/John_Graunt

I.2. Methods for Displaying Data

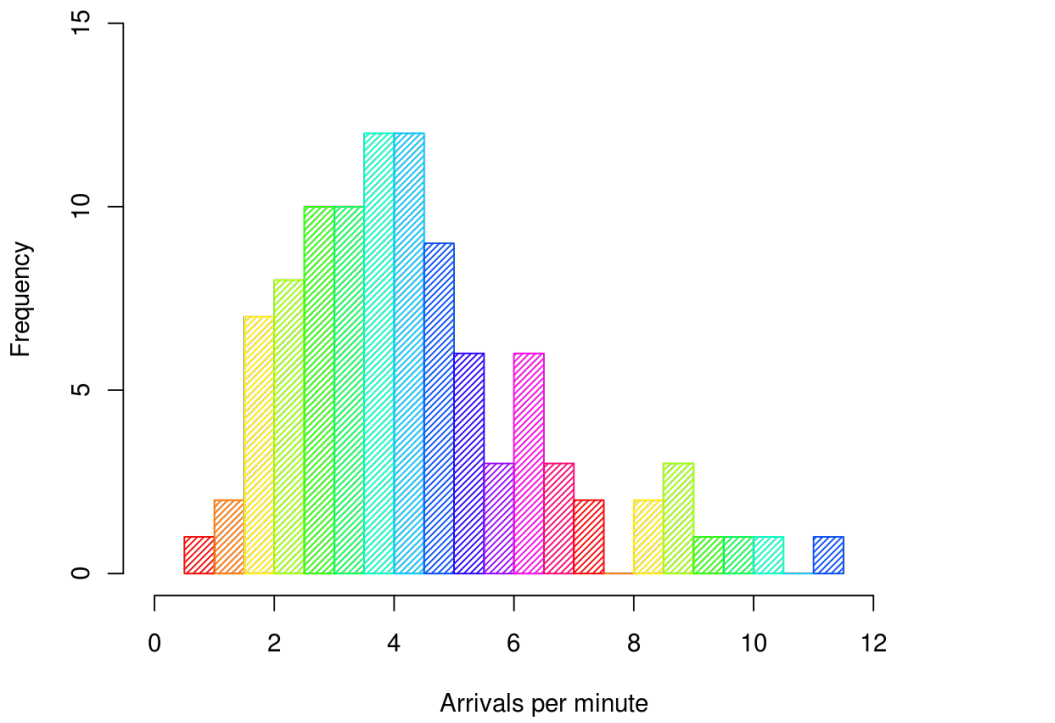
Visualizing the data is useful for

- detecting influential observations,
- identifying outliers, and
- inspecting data and understanding its distributions.

I.2.A. Histograms

Histograms as in figure I.1 are used to display the frequency or density of different values. The values are classified in *bins* representing a range of values the variable of interest can take.

Figure I.1: Histogram



Famous Statisticians ii: John Tukey

John Wilder Tukey (June 16, 1915 – July 26, 2000) was an American mathematician. Born in New Bedford, Massachusetts, he earned a B.A. in 1936 and M.Sc. in 1937, in chemistry, from Brown University, before moving to Princeton University where he received a Ph.D. in mathematics.

During World War II, he worked at the Fire Control Research Office and collaborated with Samuel Wilks and William Cochran. After the war, he returned to Princeton, where he divided his time between the university and AT&T Bell Laboratories. *He became a full professor at 35 and was founding chairman of the Princeton statistics department in 1965.*

He was awarded the *National Medal of Science by President Nixon in 1973*, and the IEEE Medal of Honor in 1982 "For his contributions to the spectral analysis of random processes and the fast Fourier transform (FFT) algorithm".

He is known for developing the FFT algorithm, the box plot, the Tukey range test, the Tukey λ distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma.

He also made many contributions and articulated the important distinction between exploratory data analysis and confirmatory data analysis. In particular, he believed that much statistical methodology placed too great an emphasis on the latter. A. D. Gordon offered the following summary of Tukey's principles for statistics:

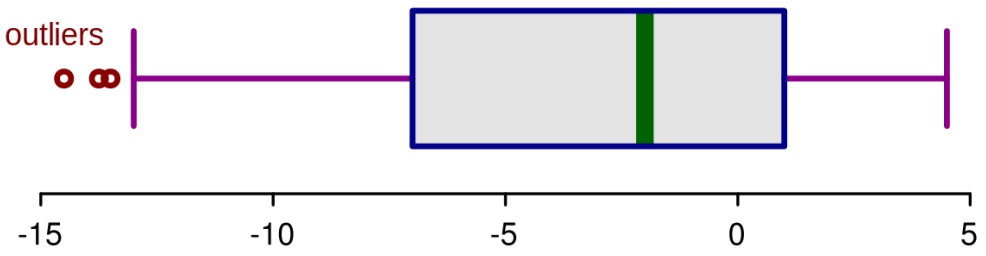
- The usefulness and limitation of mathematical statistics,
- *the importance of having methods of statistical analysis that are robust to violations of the assumptions underlying their use*,
- the need to amass experience of the behavior of specific methods of analysis in order to provide guidance on their use,
- the importance of *allowing the possibility of data's influencing the choice of method by which they are analyzed*,
- the need for statisticians to reject the role of 'guardian of proven truth', and to resist attempts to provide once-for-all solutions and tidy overunifications,
- the iterative nature of data analysis, and
- the importance of the increasing power, availability and cheapness of computers,
- with John von Neumann, he introduced the word "bit" short for "binary digit".
- Tukey's 1958 paper "The Teaching of Concrete Mathematics" contained the earliest known usage of the term "software".

Source: https://en.wikipedia.org/wiki/John_Tukey

I.2.B. Box plots

Box plots as in figure I.2 are another method to display the distribution of a dataset. They can either be shown horizontally or vertically. The box displays the *interquartile range (IQR)* range in which 50% of the data points are lying. The bar in the box is showing the 2^{nd} quartile, the median. The lines extending the box, the so called whiskers or antennas, show the area of 1.5 IQR (in the most cases, but there are also other conventions for the meaning of the whiskers). Outliers are displayed as dots for single values outside of the whiskers.

Figure I.2: Box plot



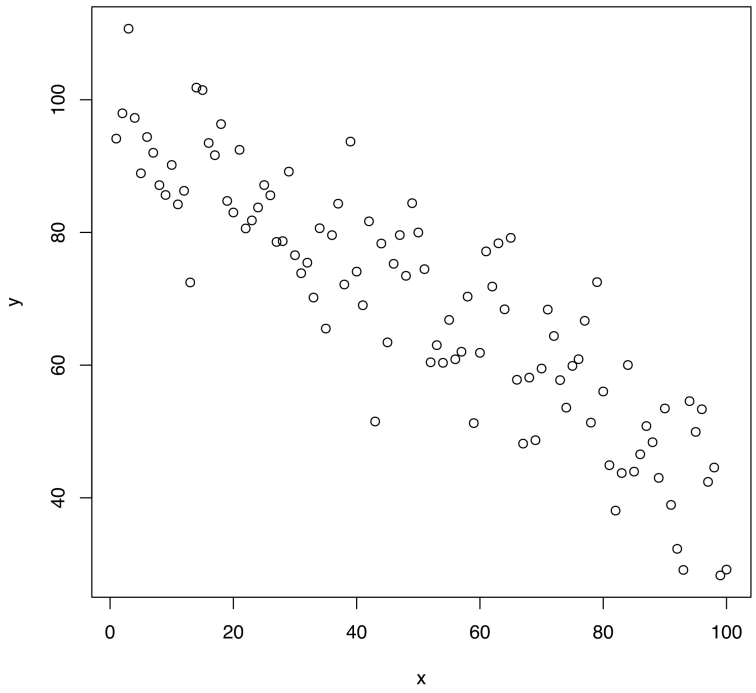
I.2.C. Pie charts

Pie charts are a useful tool to display the shares of different categories. However, they only work out well for relative values and can have a lack of clarity if many small values are included.

I.2.D. Scatter plots

Scatter plots as in I.3 display datapoints in a Cartesian coordinate system. This can be done in one to three dimensions, but it is mostly common in two dimensions. Another variable can be displayed by adding color or symbol coding. Scatter plots are helpful to find clusters and get a general idea of how the distribution of the data looks like.

Figure I.3: Scatter plot



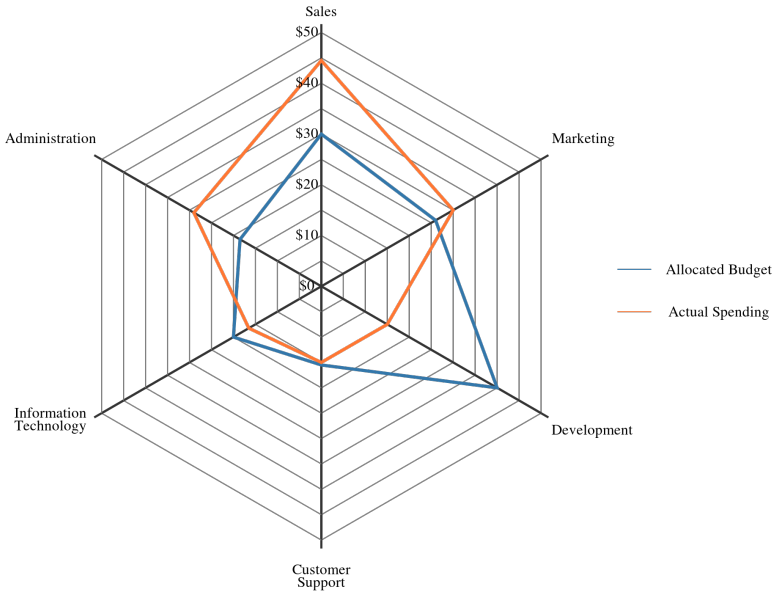
I.2.E. Radar charts

Radar charts, also known as spider plots, are shown in figure I.4. These kind of plots are useful to display how multiple variables changes in different datasets. This can be helpful to find clusters or outliers in a dataset.

I.3. Properties of Estimators

Estimators as described in definition I.9 are important to make statements about the population from a sample. There are many ways to create estimates, and some deliver better results than others.There are some properties that are important for an estimator to be a good estimator.

Figure I.4: Spider plot



Public Domain.

I.3.A. Unbiasedness

An estimator is *unbiased* if its expected value is the value of the population parameter of interest Θ. In figure I.5, the estimator T_1 is unbiased because $E(T_1) = \Theta$. The estimator T_2 has a bias which is the deviation of its expected value to the population parameter of interest.

Example I.3: Unbiased Estimator of the population mean

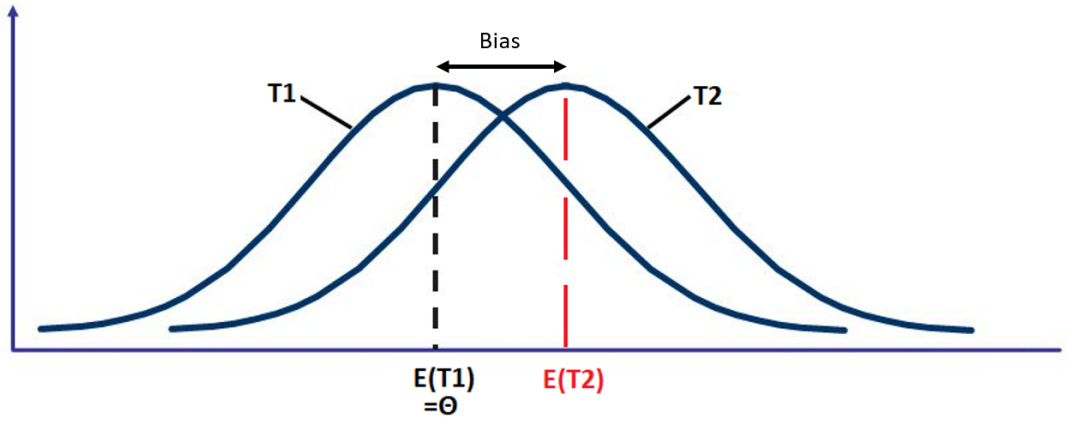
Consider a random sample

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

generated by $x_1, x_2, ..., x_n$ where the x_i have the same distribution $E(x_i) = \mu$.

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

Figure I.5: Unbiasedness of an Estimator



I.3.B. Efficiency

Consider two unbiased estimators T_1, T_2 of the same population parameter Θ. T_1 is more *efficient* than T_2 because if it has a smaller variance as shown in I.6.

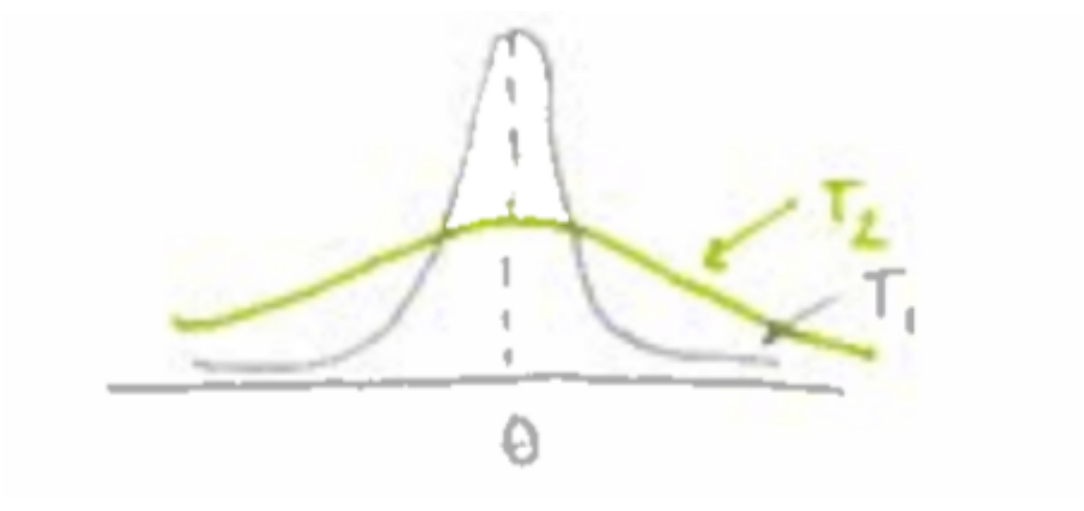
I.3.C. Consistency

An estimator is *consistent* if the probability that it generates estimates with a small error compared to the population parameter of interest increases to 1 as the sample size becomes infinitely large as shown in figure I.7.

$$\forall \varepsilon > 0, \qquad \lim_{n \to +\infty} P_s(|\hat{\Theta}_n - \Theta| > \varepsilon) = 0$$

with $\hat{\Theta}_n$ as the value of an estimate for Θ when the sample size is n.

Figure I.6: Efficiency of an Estimator



I.3.D. Sufficiency

An estimator is *sufficient* if it contains all information in a sample about the population parameter of interest.

I.3.E. Trade-off between unbiasedness and efficiency

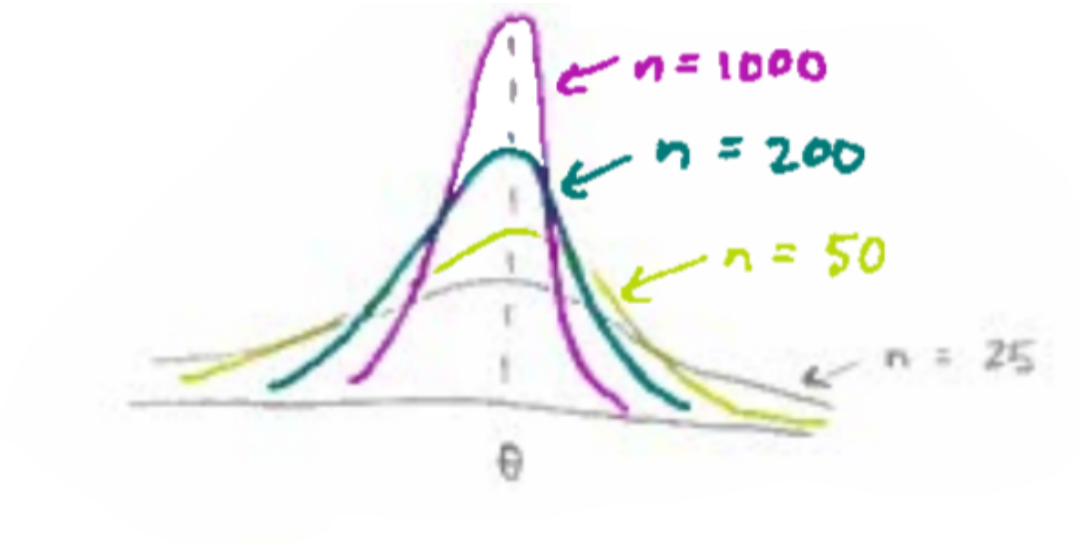
For an estimator, unbiasedness is desirable but not the only crucial factor. In figure I.8, we need to decide whether T_1 or T_2 is the better estimator. This can be determined by calculating the *mean squared error (MSE)*, and the estimator delivering the smallest MSE is the best.

$$MSE(T_1) = var(T) + bias\,(E(T), \Theta)$$

I.3.F. Central Limit Theorem (CLT)

Consider a sequence of random variables $x_1, x_2, ..., x_n$ which are independent identically distributed (i.i.d.) such that $E(x_i) = \mu$, $var(x_i) = \sigma^2$. We are interested in the random

Figure I.7: Consistency of an Estimator



variable

$$s_n = \frac{\sum_{i=1}^{n} x_i}{n}.$$

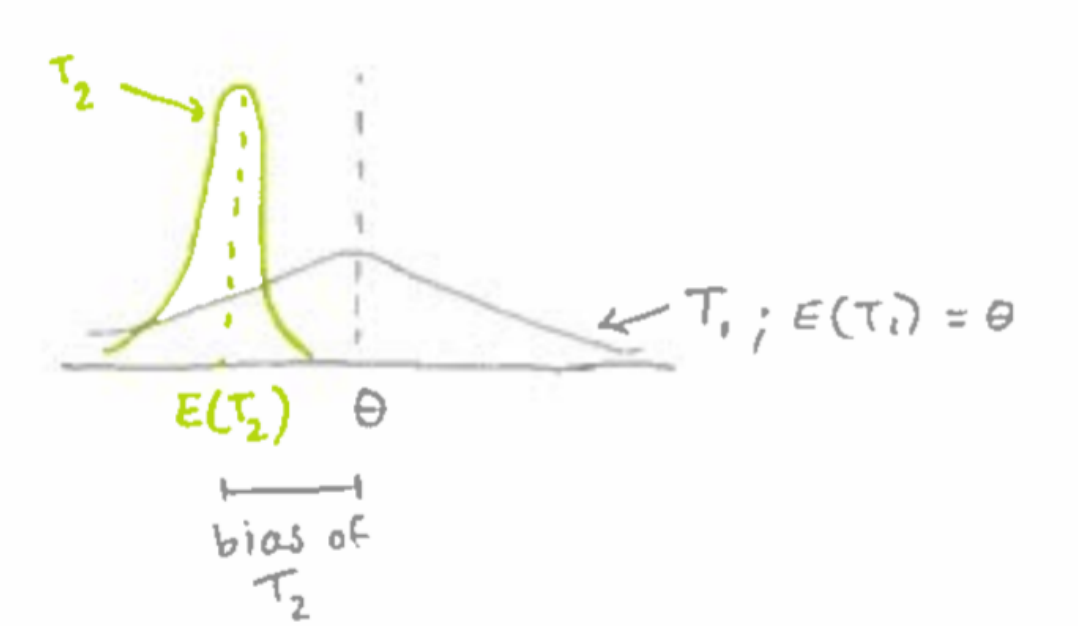
The CLT states that as $n \to +\infty$ s_n becomes approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

$$\sqrt{n}(s_n - \mu) \xrightarrow{d} N(0, \sigma^2), \qquad \xrightarrow{d} \text{ indicates in distribution.}$$

Important information we get from this:

- The variance of s_n decreases linearly with n, and
- s_n is approximately normally distributed with mean μ, so it is *unbiased*.

Figure I.8: Mean Square Error an Estimator



II. Statistical Inference

Statistical Inference is about generating and interpreting confidence intervals and setting up and testing hypotheses. Examples for questions in the field of transportation:

- Do traffic calming measures reduce speed?
- Did the deregulation of the airline industry impact the safety of flying?
- Do changeable message signs (CMS) reduce the occurrence of secondary incidents?

II.1. Confidence Intervals (CI)

Unknown parameters of the population are not known with full certainty. Working with *confidence intervals* allows to quantify the degree of uncertainty around these parameters.

II.1.A. CI for the population mean with known standard deviation

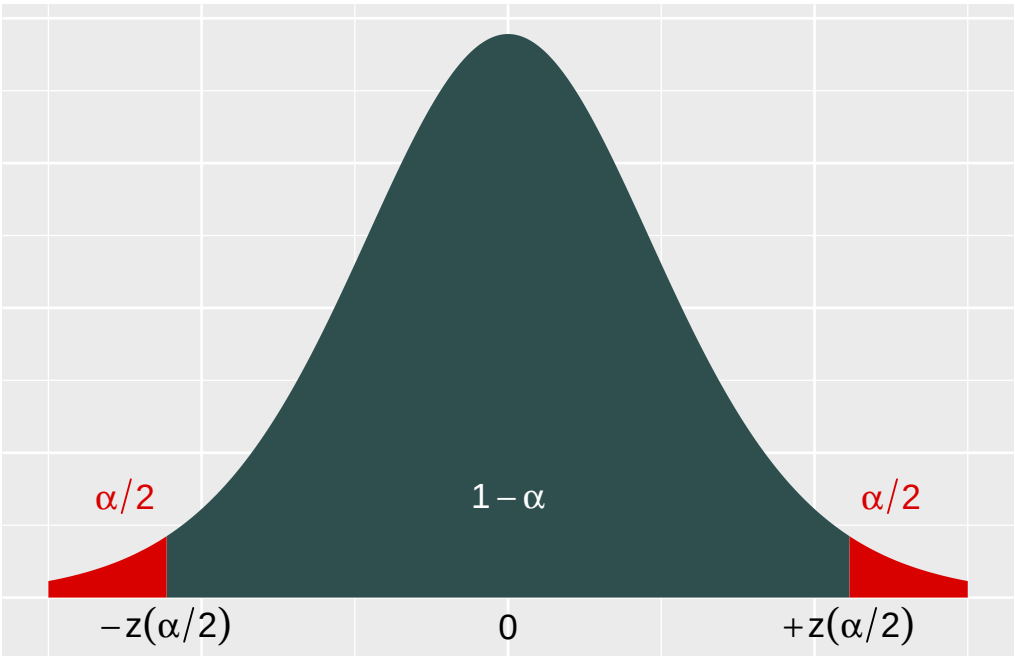
When σ is known, we can set up a confidence interval for μ using the CCT (I.3.F). If a sample is large enough, then approximately $\bar{x} \sim N(\mu, \frac{\sigma^2}{N})$. So, for approximately normal distributed data, for $\alpha \epsilon (0,1)$, a CI with confidence level α is $\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$ with n as the sample size.

II.1.B. CI for the population mean with unknown standard deviation

When σ is unknown, the notation for the confidence interval changes because we will use the sample's standard error s as an estimator for σ.

$$CI(\alpha) = \left[\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

Figure II.1: CI for the population mean with known standard deviation



By ARAKI Satoru at Wikipedia Commons. CC BY-SA 4.0. Edited.

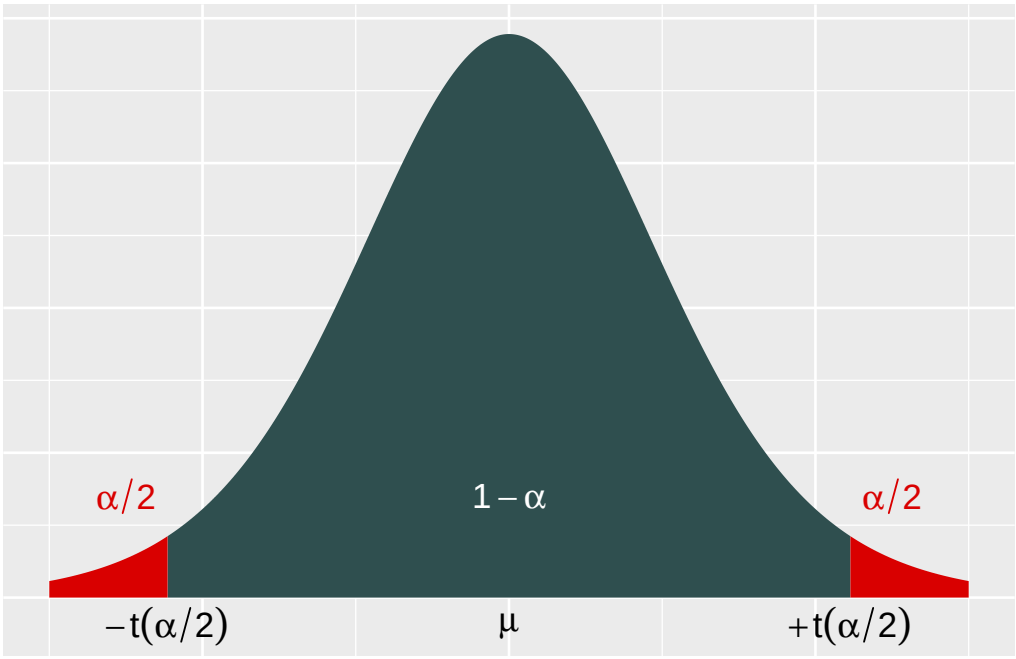
II.1.C. CI for a population proportion

We can use a sample estimate $\hat{p}$ for a population proportion, e.g. to find out the true percentage of commuters who carpool.

$$CI(\alpha) = \left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Valid if $n_p \geq 5 \bigwedge n(1-p) \geq 5$. Based on the assumption that approx. $\hat{p} \sim N\left(p, \frac{p(1-p)}{n})\right)$.

Figure II.2: CI for the population mean with unknown standard deviation



By ARAKI Satoru at Wikipedia Commons. CC BY-SA 4.0. Edited.

II.1.D. CI for the variance of a population

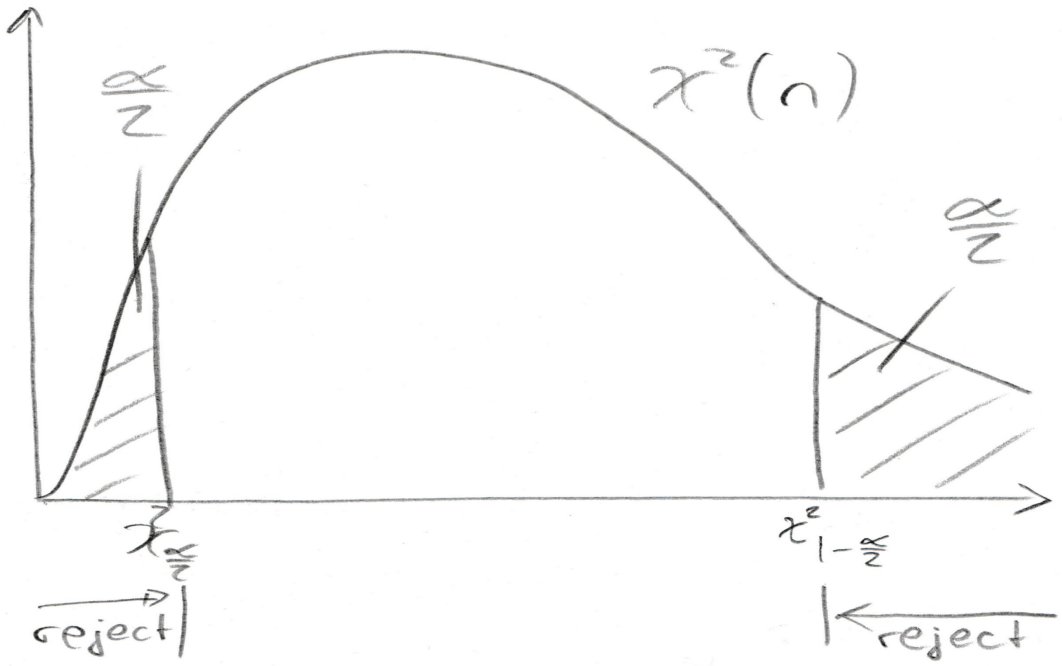
The test for variances can be used for example to find out whether high speed differences are linked to accidents occurring. Given $\alpha\epsilon(0,1)$, $\alpha(1-\alpha)\%$ CI for σ^2 is:

$$CI(\alpha) = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n}}\right]$$

II.2. Hypothesis Testing

With *hypothesis testing*, we can assess if the value for a parameter we got from an estimator differs from a set value by chance because it is truly different, so it is not related to the

Figure II.3: CI for the variance of a population



actual value of interest. The case of no relation existing is called *null hypothesis*. Hypothesis testing always follows the same framework of the following three steps:

1. Formulate a null hypothesis H_0 and an alternative hypothesis H_1 or H_a.

2. Pick a test statistic (a function based on the sample data) which helps to decide between H_0 and H_1, and gives a distribution if H_0.

3. Calculate the test statistic, pick a probability α of being wrong if H_0 is true, "a burden of proof". State and apply a decision rule (i.e. a rejection region). Conclude the results clearly.

Depending of the question the test is carried out for, H_0 can be described with an $=$, $\leq$, or $\geq$ sign. H_1 is the negated version of H_0, so not less or equal which means greater than. The mentioned distribution of the test statistic usually will be Normal (N), $F(p, q)$ or Chi-Squared ($\chi^2(n)$). The results can either indicate that we have to *reject* H_0, or if we cannot confirm that the null hypothesis is wrong, we *fail to reject it*.

Definition II.1: Errors in hypothesis testing

There are two types of errors which occur when conducting hypothesis testing.

actual truth → test result ↓	H_0 is true	H_0 is false
reject H_0	Type I error. Control error size by choosing α.	OK
fail to reject H_0	OK	Type II error. $P \sim (E_{II}) = \beta$.

Example II.1: Hypothesis statements

We want to asses whether increasing the landing fees at John-Wayne Airport (SNA) has an impact on the number of monthly flights as SNA?

1. $N = \#$flights at SNA. Current #flights is N_0.

$$H_0 : \ N = N_0 \qquad vs. \qquad H_1 : \ N \neq N_0$$

2. Will a 10% increase in landing fees show in a decrease in #flights?

$$H_0 : \ N \geq N_0 \qquad vs. \qquad H_1 : \ N < N_0$$

II.2.A. Interference about a single population

We assume that the samples we have come from approximately normal distributed populations. The results are robust against *small* deviations from this assumption in reality.

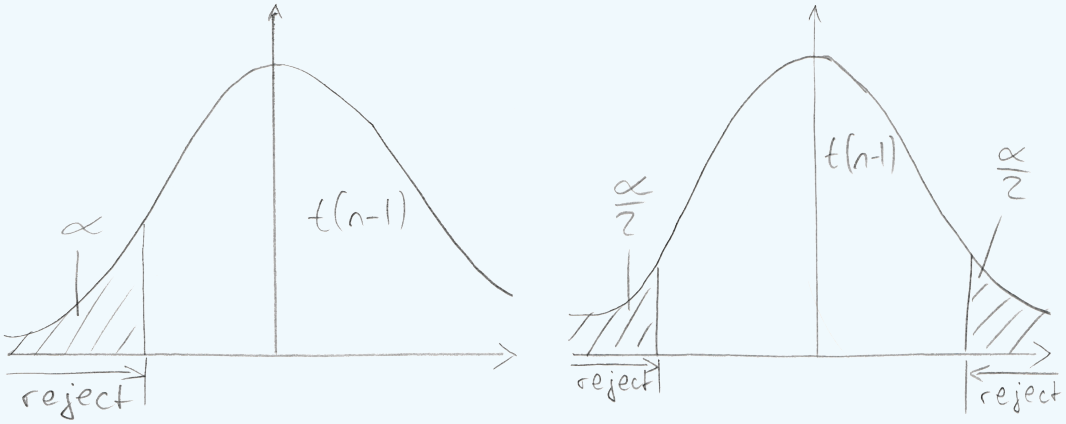
Test a population mean with unknown variance

1. State H_0 and H_1 context specific.
2. Pick the test statistic $T^* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$. If H_0 is true, $T^* \sim t(n-1)$.
3. Calculate T^*. Pick the confidence level (and probability for a type I error) α. Common values for α are 0.5%, 1%, 5%, or 10%.

The idea is to reject H_0 if the observation is highly unlikely to be based on this hypothesis.

Example II.2: One- and two-tailed hypothesis tests

One-tailed: $H_0: \mu = \mu_0$ *vs.* $H_1: \mu \neq \mu_0$ Two-tailed: $H_0: \mu \geq \mu_0$ *vs.* $H_1: \mu < \mu_0$



Test a population variance

The same 3-step-framework is used. For testing a null hypothesis of σ_0^2 from a sample size n, the test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$

If H_0 is true, $\chi^2 \sim \chi^2(n-1)$.

Example II.3: Hypothesis testing for variance

Testing if the variance of vehicle speeds on the I-405 on weekday afternoons equal to $10MPH^2$? We have a random sample with $n = 100$ from which we calculate $s^2 = 9MPH^2$.

1. $H_0: \sigma^2 = \sigma_0^2 = 10MPH^2$ *vs.* $H_1 \sigma^2 \neq 10MPH^2$
2. $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$, $\chi^2 \sim \chi^2(n-1)$
3. $\alpha = 5\% = 0.05$, $\chi^2 = \frac{99 \cdot 9}{10} = 89.1$ This value is within the confidence interval of $[73.36, 128.42]$. Therefore, we fail to reject the null hypothesis.

Test a population proportion

We use the same framework to test if a population has a certain proportion p. The following

z test statistic is used:

$$Z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}. \text{ If } H_0 \text{ is true, } z \sim N(0,1)$$

with $\hat{p}$ as a random variable that gives us the sample proportion, p_0 as the proportion specified in H_0 and sample size n.

II.2.B. Comparison of two populations

Compare two means from independent samples
Assume that the CLT (I.3.F) applies. Using the same framework, we test for the sample means as estimators of the population mean.

$$Z^* = \frac{\bar{x_1} - \bar{x_2} - (\mu_{01} - \mu_{01})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Example II.4: Hypothesis testing average on-time flights

We test if *1: Lufthansa* or *2: Air France* is more on time in average. Therefore, we test for H_0 of Lufthansa being more on time.

$$H_0: \ \mu_{10} - \mu_{20} \leq 0 \qquad vs. \qquad H_1: \ \mu_1 - \mu_2 > 0$$

If H_0 is true, for normal and small samples $n_1, n_2 \leq 25$, $T \sim t_{dof}$ with dof degrees of freedom (rounded to the next integer).

$$dof = INT\left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}\right]$$

If the populations have the same variances, the following test statistic can be used and compared to a t-distribution with $dof = n_1 + n_2 - 2$.

$$\frac{\bar{x_1} - \bar{x_2} - (\mu_{10} - \mu_{20})}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Compare two means from paired observations
For two paired datasets of the size n_d and the difference parameters marked by $\bar{x_d}, s_d$, we use the following test statistic:

$$T^* = \frac{\bar{x_d}\mu_0}{s_d/\sqrt{n_d}}. \text{ If } H_0 \text{ is true, } T^* \sim t(n_d - 1)$$

Evaluate the difference between two population proportions
For large enough sample size, we can test differences as $H_0 : \ p_1 - p_2 =, \leq, \geq p_{d0}$.

$$Z^* = \frac{\hat{p}_1 - \hat{p}_2 - p_d}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_2)}{n_1} + \frac{\hat{p}_1(1-\hat{p}_2)}{n_2}}}. \text{ If } H_0 \text{ is true, } Z^* \sim N(0,1)$$

Compare two population variances
They can be compared using the following f-test:

$$F^* = \frac{s_1^2}{s_2^2}. \text{ If } H_0 \text{ is true, } F^* \sim F(n_1 - 1, n_2 - 1)$$

Further information on hypothesis testing can be found in Conover (1999) and Washington, Mannering, and Karlaftis (2011, p. 28ff.).

Famous Statisticians iii: Karl Pearson

Karl Pearson (March 1857 – 27 April 1936) was an English mathematician and biostatistician. He established the discipline of mathematical statistics, and contributed significantly to biometrics, meteorology. Pearson was a protégé and biographer of Sir Francis Galton, who coined the term "regression".

After a private education at University College School, he went to King's College, Cambridge in 1876 to study mathematics. He then traveled to Germany to study physics and metaphysics in Heidelberg. He attended lectures on Darwinism, but also Roman law, medieval and 16th century German literature, and Socialism in Berlin.

After returning to London, he studied law until 1881 but never practiced. He then returned to mathematics, first at King's College, London in 1881 and then at University College, London in 1883. After his appointment to the professorship of Geometry at Gresham College, he met Walter Frank Raphael Weldon, a zoologist with whom he developed a fruitful collaboration. Weldon introduced him to Darwin's cousin Francis Galton. Pearson became Galton's protégé, and after Galton's death in 1911, he wrote his biography. He formed the Department of Applied Statistics, into which he incorporated the Biometric and Galton laboratories. In 1890, he married Maria Sharpe, and they had three children.

Unfortunately, Pearson was racist, anti-Semitic, and a proponent of eugenics, i.e., a social philosophy advocating the improvement of human genetic traits through the promotion of higher rates of reproduction for people with desired traits, and reduced rates of reproduction and sterilization of people with less-desired or undesired traits.

Pearson's work embraced wide applications and the development of mathematical statistics, with contributions to *biology, epidemiology, anthropometry, medicine, psychology and social history.* In 1901, with Weldon and Galton, he founded the journal Biometrika that focuses on statistical theory. Pearson's thinking underpins many of the 'classical' methods which are in common use today. Some contributions are:

- Correlation coefficient. He studied its relationship with linear regression,
- method of moments: Pearson introduced the concept borrowed from physics,
- Pearson's system of continuous univariate probability distributions that came to form the basis of the now conventional continuous probability distributions,
- foundations of hypothesis testing and of the statistical decision theory,
- use of P-values and Pearson's chi-squared test,
- Principal component analysis.

Source: https://en.wikipedia.org/wiki/Karl_Pearson

III. Linear Regression (LR)

Linear regression is widely used because it is a method which is easy to calculate. Also, linear models are easy to interpret and therefore chosen in many different contexts. However, it is important to remember the underlying assumptions when using LR.

III.1. Assumptions

Example III.1: Linear and non-linear models

Which of the following models is linear?

1. $y = \alpha_0 + \alpha_1 x_1^2 + \alpha_2 x_2 + \alpha_3 x_3 + \epsilon$
2. $y = \alpha x_1^{\beta} \mu x_2^{\gamma}$
3. $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \eta$
4. $y = \alpha + \frac{tan(\beta x_1 + \gamma x_2^2)}{1 + 2x_1^3} + \epsilon$

Solution:

1. Is linear because all parameters α are linear, the regressors x can be transformed. ϵ is an error term.
2. Is non-linear, but can be transformed to a linear model by an easy logarithmic (ln) transformation.
3. Is linear.
4. Is non-linear because the parameters are in non-linear form which cannot easily be transformed to a linear function.

III.1.A. Linearity in the unknown parameters (A1)

The unknown parameters linking the dependent variable y with the known explanatory variables or regressors x have to be linear.

$$\overbrace{y}^{\text{dependent}}_{n\times 1} = \overbrace{x}^{\text{regressor}}_{n\times k} \times \overbrace{\beta}^{\text{parameter}}_{k\times 1} + \overbrace{\epsilon}^{\text{error}}_{n\times 1}$$

Where $n = [1, 2, ..., n]$ is the sample size and $\beta = [0, 1, ..., k-1]$ are the number parameters we are looking for.

$$\begin{cases} y_1 & = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_{k-1} x_{1k-1} + \epsilon_1 \\ & \vdots \\ y_n & = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{k-1} x_{nk-1} + \epsilon_n \end{cases}$$

III.1.B. Unbiasedness of the error term (A2)

The error term consists of random errors and variables that have been excluded in an act of simplifying. However, the error term shall not have any systematic bias but be independently and identically distributed (i.i.d.) for all $\epsilon_i \sim (0, \sigma^2)$ with mean 0 (unbiased) and an unknown variance. This implies that $E(\epsilon_i) = 0$.

III.1.C. Homoscedasticity of the error term (A3)

The distribution has the variance $var(\epsilon_i) = \sigma^2$. This variance is the same for all observations.

III.1.D. Independence of the error term (A4)

The errors of different observations are not related: $cov(\epsilon_i, \epsilon_j) = 0 \; \forall \; i \neq j$.

III.1.E. Exogeneity of the regressors (A5)

The values of the explanatory variables or regressors are determined solely by influences *outside* of the environment depicted by the model. This means, y shall not have any direct influence on the value of any x.

$$\forall \; (i, j) \varepsilon \left\{0, ..., k-1\right\} \times \left\{1, ..., n\right\}, \quad cov(x_i, \epsilon_j) = 0$$

III.1.F. Normal distribution of error terms (A6)

For the purpose of testing the model, we also require the error terms ϵ_i to be approximately normal distributed.

$$\forall\,(i)\varepsilon\,\{1,...,n\}\,,\quad \epsilon_i \sim N(0,\sigma^2)$$

From these assumption we can conclude:

$$y_i \sim N\left((x)_i\beta,\sigma^2\right)$$

with $(x)_i$ as the i^{th} row (dimension $1 \times k$) of the regressors x and β with dimension $k \times 1$. This means, we assume all x_{ij} are known.

III.2. Estimation

To find estimates for the unknown β and σ^2, we use the method of *linear regression*. It helps to understand how the regressors x affect the dependent variable y.

III.2.A. Least Squares

The *method of least squares*, also known as ordinary least squares (OLS), is one of the most common estimations used in linear regression. The method tries to draw the best fitting straight line into the datapoints (x_n, y_n). By minimizing sum of the square of the distance between the line and each datapoint (the error or residuum), we will get the *BLUE*, the best linear unbiased estimator.

$$\min_{\beta_0,...,\beta_{k-1}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

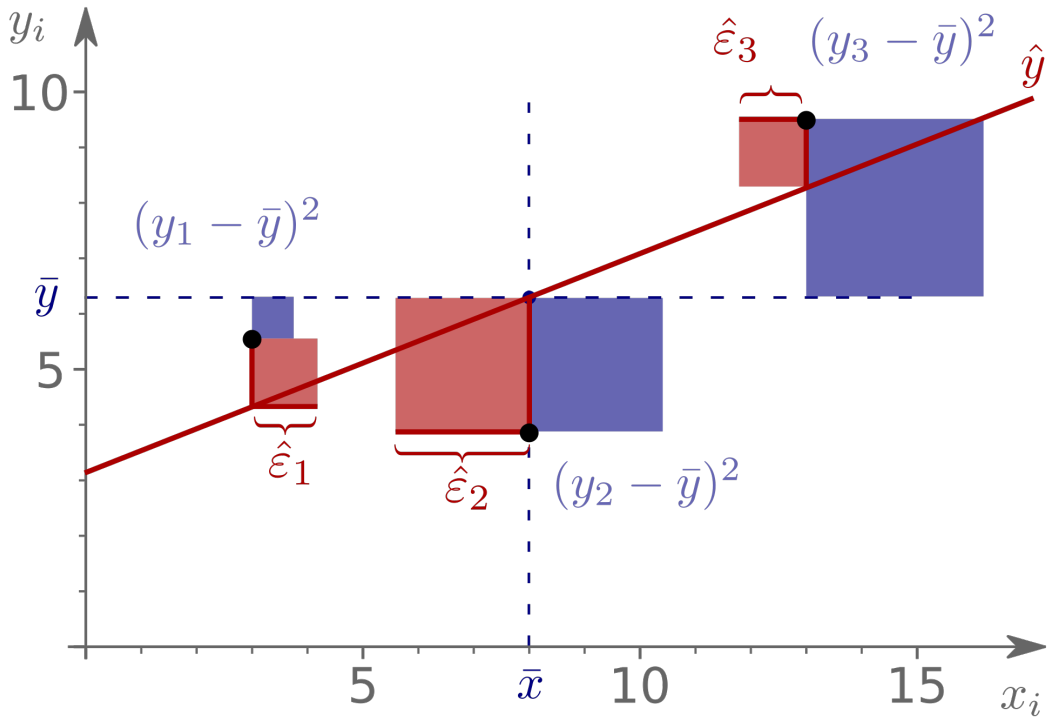
This means, seeing this as a function $f(\beta_0, \beta_1, ...\beta_{k-1})$ stating the sum of the squares of the differences between observed and estimated values, we are trying to minimize this function. With β_0 as a constant intercept, we can define $x_0 \equiv 1$.

$$\min_{\beta_0,...,\beta_{k-1}} f(\beta_0, \beta_1, ...\beta_{k-1}) = \min_{\beta_0,...,\beta_{k-1}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\beta_0,...,\beta_{k-1}} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k-1} \beta_j x_{ji} \right)^2$$

To find estimates of $\beta_0, \beta_1, ...\beta_{k-1}$ we find the derivative:

$$\frac{\partial f(\cdot)}{\partial \beta_j} \quad \text{for } j\varepsilon\{0,1,...,k-1\}$$

Figure III.1: Ordinary least squares



By Debenben at Wikipedia Commons. CC BY-SA 4.0.

$$\frac{\partial f(\cdot)}{\partial \beta_0} = \sum_{i=1}^{n} -2 \left(y_i - \sum_{j=0}^{k-1} \beta_j x_{ji} \right) \qquad (1)$$

$$\vdots$$

$$\frac{\partial f(\cdot)}{\partial \beta_j} = \sum_{i=1}^{n} -2 \left(y_i - \sum_{j=0}^{k-1} \beta_j x_{ji} \right) \qquad (j)$$

This means, k equations for k unknown β, so we can solve this. In matrix notation with X as the matrix of explanatory observations, the solution will be (according to the method shown in appendix **??**):

$$\hat{\beta} = (X'X)^{-1}X'Y$$

From equation (1), we have

$$\sum_{i=1}^{n} y_i - \sum_{j=0}^{k-1} \beta_j \sum_{i=1}^{n} x_{ji} = 0.$$

Dividing by n gives us

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{=\bar{y}} - \sum_{j=0}^{k-1} \beta_j \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{ji}}_{=\bar{x}} = 0$$

$$\implies \bar{y} = \sum_{j=0}^{k-1} \hat{\beta}_j \bar{x_j}.$$

In other words, the regression line goes through the sample mean.

III.2.B. Maximum Likelihood

The idea of this method is to *maximize the likelihood* or chance of getting our sample when observing a population. This means to puck a parametric family and find the parameters that maximize the chance of observing the sample. Recall, assumptions A1-A6 (III.1): $y_i \sim N\left((x)_i\beta, \sigma^2\right)$. Let $f(\cdot)$ designate the joint distribution of $\{y_1, y_2, ..., y_n\}$.

But we assumed that the error terms ϵ (and therefore $y_1, y_2, ..., y_n$) are from a random sample and thus i.i.d. In consequence, the joint density of $y_1, y_2, ..., y_n$ is equal to the products of the marginal densities of $y_1, y_2, ..., y_n$.

Example III.2: Density of a normal distribution

W is normally distributed: $W \sim N(\mu, \sigma^2)$. Then, the density of W is:

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

In consequence, the joint density of $y_1, y_2, ..., y_n$ is:

$$f(y, x, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left[-\frac{1}{2}\left(\frac{y_i - \sum_{j=0}^{k-1}\beta_j x_{ij}}{\sigma}\right)^2\right]$$

$f(\cdot)$ is a product of small values (all factors in the product are ≤ 1). This makes exact calculations hard. Also, taking the derivative for maximizing the function is less convenient with a product function than with a sum. So before maximizing the function, we perform a logarithmic transformation. The log-likelihood function then is as follows:

$$\begin{aligned}
\mathcal{L}(\beta, \sigma^2|y, x) &= ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left[-\frac{1}{2}\left(\frac{y_i - \sum_{j=0}^{k-1}\beta_j x_{ij}}{\sigma}\right)^2\right]\right) \\
&= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k-1}\beta_j x_{ij}\right)^2 \\
\frac{\partial\mathcal{L}}{\partial\sigma^2} &= 0 - \frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k-1}\beta_j x_{ij}\right)^2 \\
\Longrightarrow \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\underbrace{(y_i - \hat{y}_i)}_{=\hat{\epsilon}_i}{}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i - E(\hat{\epsilon}_i))^2
\end{aligned}$$

With $E(\hat{\epsilon}_i) = 0$. This expression is similar to calculating the variance as seen in definition

I.8. In this sense, we can see it as the variance of the error. Let us write $\mathcal{L}$ as follows:

$$\begin{aligned}
\mathcal{L} &= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\underbrace{(Y - X\beta)'}_{\text{row}}\underbrace{(Y - X\beta)}_{\text{column}} \\
&= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k-1}\beta_j x_{ij}\right)^2 \\
\frac{\partial\mathcal{L}}{\partial\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i, \qquad \text{Recall: } E(\epsilon_i) = 0 \\
\frac{\partial\mathcal{L}}{\partial\beta} &= 0 + 0 - \frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}\left[y'y - y'x\beta - \beta'x'y + \beta\right] \\
&= -\frac{1}{2\sigma^2}\left[0 - 2x'y + 2x'x\beta\right] \\
\text{Recall: } G(\beta) &= \beta'M\beta, \quad \frac{\partial G}{\partial\beta} = M'\beta + M\beta \\
(x'x)' &= x'x'' = x'x \\
\text{With symmetry } M &= M' \\
M'\beta + M\beta &= 2M\beta \\
\frac{\partial\mathcal{L}}{\partial\beta} &= 0 \Longrightarrow \left[x'y = (x'x)\beta\right] \\
\Longrightarrow \hat{\beta} &= (x'x)^{-1}x'y \\
\text{Provided } rank(x'x) &= k \quad (x'x \text{ has an inverse})
\end{aligned}$$

III.2.C. Properties of OLS & ML estimators

Expected value

$$E(\hat{\beta}) = E\left((x'x)^{-1}x'y\right) = (x'x)^{-1}x'E(y)$$

With x as a known an non-random variable $E(\hat{\beta}) = (x'x)^{-1}x'x\beta = \beta$

$$E(\hat{\beta}) = \beta$$

This means, $\hat{\beta}$ is an unbiased estimator.

Variance

$$\begin{aligned}\hat{\beta} &= (x'x)^{-1}x'y = (x'x)^{-1}x'(X\beta+\epsilon)\\ &= (x'x)^{-1}x'x\beta + (x'x)^{-1}x'\epsilon\\ &= \beta + (x'x)^{-1}x'\epsilon\end{aligned}$$

Recall: Let W be a random variance. Then:

$$\begin{aligned}var(W) &= E\left((W-E(W))^2\right)\\ var(\hat{\beta}) &= E\left((\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right)\\ &= E\left((x'x)^{-1}x'\epsilon\epsilon'x(x'x)^{-1}\right)\\ &= (x'x)^{-1}x'E(\epsilon\epsilon')x(x'x)^{-1}\\ &= (x'x)^{-1}x'\sigma^2 I_n x(x'x)^{-1}\\ &= \sigma^2(x'x)^{-1}x'x(x'x)^{-}1 = \sigma^2(x'x)^{-1},\ (x'x)^{-1}x'x = I_k\\ var(\hat{\beta}) &= \sigma^2(x'x)^{-1}\\ \widehat{var(\hat{\beta})} &= \hat{\sigma^2}(x'x)^{-1}\end{aligned}$$

$$\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_{k-1}\end{pmatrix}$$

The standard error is the square root of the variance.

$$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma^2}(x'x)^{-1}_{jj}}$$

$$var(\hat{\beta}) = \begin{pmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0,\hat{\beta}_1) & \cdots & cov(\hat{\beta}_0,\hat{\beta}_{k-1})\\ cov(\hat{\beta}_1,\hat{\beta}_0) & var(\hat{\beta}_1) & \cdots & cov(\hat{\beta}_1,\hat{\beta}_{k-1})\\ \vdots & \vdots & \ddots & \vdots\\ cov(\hat{\beta}_{k-1},\hat{\beta}_0) & cov(\hat{\beta}_{k-1},\hat{\beta}_1) & \cdots & var(\hat{\beta}_{k-1})\end{pmatrix}$$

Gauss-Markov theorem
Assume that A1-A5 (III.1.A) hold. Then, the ML-OLS estimators achieve minimum variance in the class of linear, unbiased estimators. So, they are most efficient.

Proof: Let $\tilde{\beta}$ be another linear, unbiased estimator. Without loss of generality, we write

$$\begin{aligned}
\tilde{\beta} &= \hat{\beta} + Dy = \left[(x'x)^{-1} + D\right] y \\
E(\tilde{\beta}) &= \beta \quad (\tilde{\beta} \text{ is unbiased }) \\
E(\tilde{\beta}) &= E\left(\left[(x'x)^{-1}x' + D\right][x\beta + \epsilon]\right) \\
&= \underbrace{(x'x)^{-1}x'x}_{I_k}\beta + Dx\beta + \left[(x'x)^{-1}x' + D\right] \underbrace{E(\epsilon)}_{=0\,(A2)} \\
&= \beta + Dx\beta = (I + Dx)\beta = \beta \Longrightarrow Dx = 0 \\
var(\tilde{\beta}) &= var(Cy) = Cvar(y)C' = C\sigma^2 I_m C' \\
&= \sigma^2 CC' \\
&= \sigma^2 \left[(x'x)^{-1}x' + D\right]\left[(x'x)^{-1}x' + D'\right] \\
&= \sigma^2 \left[(x'x)^{-1}\underbrace{x'x(x'x)^{-1}}_{I_k} + \underbrace{(x'x)^{-1}x'D'}_{=0} + DD' + \underbrace{Dx(x'x)^{-1}}_{=0}\right] \\
&= \sigma^2(x'x)^{-1} + \sigma^2 DD' \\
&= var(\hat{\beta}) + \sigma^2 DD'
\end{aligned}$$

DD' is positive semi-definite, i.e. $\forall v$, $v'DD'v = (D'v)'(D'v) \geq 0$. Recall, this means that $norm(D'v) \geq 0$. So, $var(\tilde{\beta})$ is larger than the variance $var(\hat{\beta})$. As a consequence, $\hat{\beta}$ is BLUE (III.2.A). Linearity here means, in y, $\hat{\beta} = (x'x)^{-1}xy$) with x known.

Famous Statisticians iv: Carl Friedrich Gauß

Johann Carl Friedrich Gauß (30 April 1777 – 23 February 1855) in Brunswick, in the Duchy of Brunswick-Wolfenbüttel (now part of Lower Saxony, Germany, as the son of poor working-class parents. *Gauß was a child prodigy.* A contested story relates that, when he was eight, he figured out how to add up all the numbers from 1 to 100.

Gauß' intellectual abilities attracted the attention of the Duke of Brunswick, who sent him to the Collegium Carolinum (1792 to 1795, and to the University of Göttingen (1795 to 1798). While there, Gauß independently rediscovered several important theorems. In 1796, he became the first to prove the quadratic reciprocity law, which allows mathematicians to *determine the solvability of any quadratic equation.* He also conjectured the prime number theorem, which gives a good understanding of how prime numbers are distributed among integers. In his 1799 doctorate, Gauß proved *that every nonconstant single-variable polynomial with complex coefficients has at least one complex root.* Gauß also made important contributions to number theory in his 1801 book Disquisitiones Arithmeticae.

In 1831 Gauß developed a fruitful collaboration with physicist Wilhelm Weber, leading to new knowledge in magnetism and the *discovery of Kirchhoff's electric circuit laws.* They constructed the first electromechanical telegraph in 1833, which connected the observatory with the institute for physics in Göttingen. In 1840, Gauß published his influential "Dioptrische Untersuchungen", in which he gave the first systematic analysis on the formation of images under a paraxial approximation (*Gaußian optics*).

Gauß' personal life was overshadowed by the early death of his first wife, Johanna Osthoff, in 1809, soon followed by the death of one child, which caused him to become depressed. When his second wife died in 1831 after a long illness, one of his daughters took over the household and cared for Gauß for the rest of his life. Gauß had six children, 3 with each wife.

Gauß made major contributions to various areas of mathematics (including geometry and number theory) and physics (including magnetism and optics). The importance of his contributions is often compared to those of Newton. He also made some key contributions to statistics. The most important one is the development of *least squares estimation recursive methods*, which he discusses in a book on planetary orbits. He also proposed some, which are used for time series analysis and were used to help calculate the trajectory of the Apollo spacecraft. He is also *credited with developing the normal distribution* (also called the Gaussian distribution or bell curve), which is extremely useful in probability and statistics.

Sources: https://en.wikipedia.org/wiki/Carl_Gauss, http://www.sciencedirect.com/science/article/pii/0315086078900496

III.2.D. Inference

Let us now also assume that A6 (III.1.F, normality of ϵ) holds. Since $\hat{\beta}$ is a linear estimator (in y), and since $y \sim N(x\beta, \sigma^2 I_n)$ with i.i.d., then $\hat{\beta}$ as a linear combination of a independent normals is also normal distributed.

Hypothesis Testing
From the fact that $\hat{\beta} \sim N$,

$$z^* = \frac{\hat{\beta}_l - \beta_{l0}}{\sigma(\hat{\beta}_l)} \sim N(0,1)$$

with $l\varepsilon\{0, 1, ..., k-1\}$, $\sigma(\hat{\beta}_l)$ as the standard error of $\hat{\beta}$ which is the square root of its variance.

In practice, we do not know $\sigma(\hat{\beta}_l)$. Because of that, we use

$T^* = \frac{\hat{\beta}_l - \beta_{l0}}{s(\hat{\beta}_l)}$. If H_0 is true, $t* \sim t(n-k)$ with $k = \#$ of β in the model.

Confidence Interval
We have, given $\alpha\varepsilon(0,1)$,

$$CI(1-\alpha) = \hat{\beta}_l \pm t_{\frac{\alpha}{2}}(n-k) \cdot s(\hat{\beta}_l)$$

Joint test (f-test)
Step 1:

$$H_0 : \beta_l = 0 \wedge \beta_{l+1} = 0 \wedge \cdots \wedge \beta_{k-1} = 0$$
$$H_1 : \text{not } H_0$$

Step 2:

$$F^* = \frac{(SSE_R - SSE_F)/\#restrictions}{SSE_F/(n-k)}$$

with $\#restrictions = k - l$. If H_0 is true, $F^* \sim F(k-l, n-k)$. SSE is the sum of the square errors $SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2$. R are parameters for the restricted model (obtained by imposing H_0) and F of the full model with $k = \#$ of β.

Example III.3: Restricted linear model

y_i = demand for bikesharing at location i.

$$F: \ y_i = \beta_0 + \beta_1 Inc_i + \beta_2 GenZ_i + \beta_3 GenY_i + \beta_4 GenX_i + \beta_5 Edu_i + \beta_6 Carpp_i + \epsilon_i$$
$$R: \ y_i = \beta_0 + \beta_1 Inc_i + \beta_2 GenZ_i + \beta_3 GenY_i + \mu_i$$

Joint test: $H_0 : \beta_4 = 0 \wedge \beta_5 = 0 \wedge \beta_6 = 0$

Step 3:

a. Pick $\alpha = Pr(\text{Type I error})$,

b. calculate F^*_{calc},

c. rejection region and locate F^*_{calc}, and

d. conclude.

Example III.4: Testing for different values than zero

Compare to the test for parameters to be zero in example III.3.

$$y_i = \beta_0 + \cdots + (1 + \beta_6) Carpp_i + \epsilon_i$$
$$R: \ y_i = \beta_0 + \cdots + \beta_3 GenY_i + Carpp_i + \mu_i$$

Joint test: $H_0 : \beta_4 = 0 \wedge \beta_5 = 0 \wedge \beta_6 = 0$

III.2.E. Goodness of Fit

Square Errors

R^2 is the fraction of the variation of the dependent variable explained by the model. Consider $\{y_1, y_2, ..., y_n\}$ as our dependent variable. One way to measure the variations of y is

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

with SST as the total sum of squares. One way to characterize how much a linear model *explains* the variations of y is:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$R^2 \equiv \frac{SSR}{SST}$$

We also can define

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

with SSE as the square sum of errors and $SST = SSR + SSE$.

$$\begin{aligned}
SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i\hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i) + 2\sum_{i=1}^{n}(y_i - \hat{y})(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
\sum_{i=1}^{n}\hat{\epsilon}_i(\hat{y}_i - \bar{y})^2 &= \hat{\epsilon}' \cdot (x\hat{\beta} - \bar{y}1) = \hat{\epsilon}'x \cdot \hat{\beta} - \bar{y}\hat{\epsilon}'1 \\
\hat{\epsilon}'x &= (y - \hat{y})'x = (y - x(x'x)^{-1}x'y)'x \\
&= (y' - y'x(x'x)^{-1}x')x \\
&= y'(x - x(x'x)^{-1}x'x) = y'(x - x) = 0 \\
\hat{\epsilon}'1 &= \sum_{i=1}^{n} = \hat{\epsilon}_i = 0
\end{aligned}$$

because $E(\epsilon) = 0$ which comes from $\frac{\partial \mathcal{L}}{\partial \beta_0} = 0$. In consequence,

$$\begin{aligned}
SST &= SSR + SSE \\
R^2 &= \frac{SSR}{SST} \\
SSR, SST, SSE &\geq 1 \\
\implies 0 \leq R^2 &\leq 1
\end{aligned}$$

Limitation of R^2 is that it will increase with the numbers of explanatory variables. It reflects the benefits of adding more variables, but not the cost of including them: the precision with which the β are certain, their standard error. As a result, statisticians have introduced

$$\bar{R^2} = 1 - \frac{SSE/(n-k)}{SSR(n-1)} = 1 - \frac{n-1}{n-k}\frac{SSE}{SST} = 1 - \frac{n-1}{n-k}(1 - R^2)$$

because

$$\frac{SSE}{SST} = \frac{SSE + SSR - SSR}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST} = 1 - R^2.$$

We still have that $\bar{R^2} \leq 1$ but now $\bar{R^2}$ can be ≤ 0. Higher values of R^2, close to 1, indicate a better fit.

Model F-test

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i\,k-1} + \epsilon_i$$

Most statistical softwares perform *model F-tests* by default. They test:

$$H_0 : \ \beta_1 = 0 \wedge \beta_1 = 0 \wedge \cdots \wedge \beta_{k-1} = 0 \quad H_1 : \ not\, H_0$$

If H_0 is true, $F^* \sim F(k-1, n-k)$.

This means, H_1 is at least one $\beta \neq 0$. This F-test tells us about the usefulness of the model as whole.

$$F^* = \frac{(SSE_R - SSE_F)/(k-1)}{SSE_F/(n-k)}$$

In general, necessarily $SSE_R \geq SSE_F$ because the explanatory variables of the restricted model are in the full model as well. Let us relate F^* to R^2.

$$F^* = \frac{\frac{SSE_R - SSE_F}{SST}/(k-1)}{\frac{SSE_F}{SST}/(n-k)} = \frac{\left(1 - R_R^2 - (1 - R_F^2)\right)/(k-1)}{(1-R_F^2)/(n-k)} = \frac{(R_F^2 - R_R^2)/(k-1)}{(1-R_F^2)/(n-k)}$$

What is R_R^2?

$$\begin{aligned} SSR_R &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ \text{here:} \quad &= \sum_{i=1}^{n} (\bar{y} - \bar{y})^2 \\ R_R^2 &= 0 \\ \Rightarrow F^* &= \frac{R_F^2/(k-1)}{(1-R_F^2)/(n-k)} \end{aligned}$$

Information criteria

There are measures of relative fit of a model that take parsimony into account. They provide a trade-off between complexity and simplicity.

1. Akaike Information Criterion (AIC)

   $$AIC = -2ln(\mathcal{L}^*) + 2k$$

   with $ln(\mathcal{L}^*)$ as the value of log-likelihood function at $\hat{\beta}$ and k as the number of β. Smaller AIC values indicate a better model because unnecessary complexity results in a higher AIC. A significant difference in the quality of models usually results in a AIC difference greater than 3.

2. Bayesian Information Criterion (BIC) Also known as *Schwarz information criterion*, the BIC is similar to AIC, but favors more parsimonious models because sample size is taken into account.

$$BIC = -2ln(\mathcal{L}^*) + k\,ln(n)$$

We only can compare models estimated on the same dataset, but one does not need to be a subset of the other.

Interpretation

Example III.5: Gas mileage of a car

$mpg_i = \beta_0 + \beta_1 EngSze_i + \beta_2 Weight_i + \beta_3 cw + \beta_4 MY16 + \beta_5 MY17 + \beta_6 MY18 + \beta_7 MY19$

Where $EngSze, Weight, DragCoeff$ are continuous and $MY16, MY17, MY18, MY19$ are categorical and therefore binary variables: $= 1$ if yes, $= 0$ otherwise.

Multicollinearity: when one explanatory x_1 variable is almost a linear correlation of the other explanatory variables $x_2, ...x_{k-1}$:

$$x_1 = \alpha_0 + \alpha_2 x_2 + \cdot\alpha_{k-1} x_{k-1} + \mu \longrightarrow R^2$$

Then we can calculate the *Variance Inflation Factor*:

$$VIF_1 = \frac{1}{1 - R_1^2}$$

If $VIF \geq 10$, the choice of variables should be improved.

III.3. Verification of assumptions

It is essential to check the assumptions (III.1) we made when we built our linear model.

Linearity

Before estimating our model, plot y vs each continuous variable . If the relationship is *far* from being linear, transform the explanatory variable (ln, power,...). Afterwards, plot $\hat{\epsilon}$ versus $\hat{y}$. There should be no trends, so the graph should be random distributed after the transformation.

Unbiasedness
$E(\epsilon) = 0$. No need to check this assumptions, it is ensured by the estimation.

Homoscedasticity
We can detect heteroscedasticity (the opposite) if we have patterns when plotting $\hat{\epsilon}$ versus $\hat{y}$, they should be randomly distributed. Besides the graphical evaluation, it can be assessed with statistical tests such as *Park*, *Goldfield-Quandt* or *White*. The model can also be estimated with robust standard errors so it can handle moderate deviations from homoscedasticity. If there is heteroscedasticity, it is not per se an issue for the model, but it becomes a problem for testing β.

Uncorrelated errors
This assumption is typically violated for time series or spatial data.

Example III.6: Time series

Consider $\begin{cases} y_k = \beta_0 + \beta_1 x_t + \epsilon_t \\ \epsilon_t = \rho\epsilon_{t-1} + \mu_t \end{cases}$ with $\mu_t \sim N(0, \sigma_\mu^2)$ as an i.i.d. and $|\rho| < 1$.

We have:

$$\begin{aligned} \epsilon_t &= \rho\epsilon_{t-1} + \mu_t \\ \epsilon_{t-1} &= \rho\epsilon_{t-2} + \mu_{t-1} \\ &\vdots \\ \epsilon_{t-n+1} &= \rho\epsilon_{t-n} + \mu_{t-n+1} \end{aligned}$$

By substitution:

$$\begin{aligned} \epsilon_t &= \mu_t + \rho\mu_{t-1} + \cdots + \rho^{n+1}\epsilon_{t-n+1} \\ \epsilon_t &= \mu_t + \rho\mu_{t-1} + \cdots \\ var(\epsilon_t) &= \sigma^2 + \rho^2\sigma^2 + \rho^2\sigma^2 + \cdots && = \sigma^2(1 + \rho^2 + \rho^4 + \rho^6 + \cdots) \\ &= \frac{\sigma^2}{1 - \rho^2} \end{aligned}$$

We know for X as a random variable:

$$\begin{aligned} var(X) = \sigma^2,\ E(X) = \mu \\ a > 0,\ var(aX) &= E\left((ax - E(aX))^2\right) \\ &= E\left((ax - E(aX))^2\right) \\ &= a^2 E\left(E(X - E(X))^2\right) \\ &= a^2\sigma^2 \end{aligned}$$

Suppose it holds for $n-1$:

$$\begin{aligned} &- \alpha \overbrace{(1 + \alpha + \cdots + \alpha^{n-1})}^{=S} \\ &= (1-\alpha)S = 1 - \alpha^n \\ S &= \frac{1-\alpha^n}{1-\alpha} \end{aligned}$$

Here, $\rho^2 = \alpha$ since $|\rho| < 1$.

$$\lim_{n \to +\infty} S_n = \frac{1}{1-\alpha}$$

Example III.6 tells us that under serial conditions, the errors are *off* so OLS (or ML) is no longer efficient (but still unbiased and consistent). To detect a serial correlation, we can use the *Durbin-Watson statistic* (DW)

$$d = \frac{\sum_{t=2}^{T} (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T} \hat{\epsilon}_t^2}$$

with T as the number of observations. It can be shown that $d \approxeq 2(1-p)$ where p is the sample autocorrelation of errors. Moreover, $0 \leq d \leq 4$:

$$\text{If } \begin{cases} d \ll 2, \text{ then we have positive serial correlation.} \\ d \gg 2, \text{ then we have negative serial correlation.} \end{cases}$$

The difficulty is that the distribution of d depends on the explanatory variables within the model, so the DW test is not always conclusive. To test for t positive series correlation at

level α which is the probability of a type I error:

$$\text{If } \begin{cases} d < d_{L,\alpha}, \text{ then reject } H_0 \text{ of no series correlation} \\ d > d_{U,\alpha}, \text{ then fail to reject } H_0 \\ d_{L,\alpha} \leq d \leq d_{U,\alpha}, \text{ then cannot conlcude} \end{cases}$$

With $d_{L,\alpha}$ as the lower and $d_{U,\alpha}$ as the upper bound. Similar for testing negative serial correlation:

$$\text{If } \begin{cases} (4-d) < d_{L,\alpha}, \text{ then reject } H_0 \text{ of no series correlation} \\ (4-d) > d_{U,\alpha}, \text{ then fail to reject } H_0 \\ d_{L,\alpha} \leq (4-d) \leq d_{U,\alpha}, \text{ then cannot conlcude} \end{cases}$$

In transportation, we often analyze data with spatial correlation. If we ignore spatial correlation, our estimators may be biased and inconsistent.

Exogeneity of regressors
When some explanatory variables are not exogenous, so determined outside of the model, the ML or OLS estimators are no longer consistent. Some reasons for this inconsistency are:

1. Omitted variable is substantially correlated with the model's explanatory variable.
2. y and one of the explanatory variable are jointly determined.

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\ x_i = \gamma_0 + \gamma_1 y_i + \gamma_2 \zeta_i + \mu_i \end{cases}$$

3. Some variables are measured with substantial errors, so systematically and large.
4. The relationship between the dependent variable and the regressors is changing over time (non-stationarity).

Normality of errors
All β are still unbiased and consistent but the testing will be off. There are some methods to detect non-normality of errors:

1. Draw a histogram of $\hat{\epsilon} = y - \hat{y}$. It should be the bell-shaped *Gaussian curve*.
2. Generate a Q–Q (quantile-quantile) plot. The Q-Q plot shows the quantiles of $\hat{\epsilon}$ versus the quantiles of a $N(0,1)$. The plot should be a straight line.
3. Formal tests of normality, e.g. the *Lilliefors test* for small samples ($n \leq 2000$) or *Jarque-Bera* for samples of $n \geq 2000$.

Example III.7: Jarque-Bera test

$$JB^* = \frac{n}{6}\left(g^2 + \frac{(k-3)^2}{4}\right)$$

With n as the size, g as the skewness, and k as the kurtosis of the sample. Under H_0, $JB^* \sim \chi^2(2)$.

III.4. Regression diagnostics

III.4.A. Unusual and influential observations

The idea is, that if a single (or a few) observations are different from the rest of the sample they can have a big impact on the β. The goal is to detect these observations and to analyze them.

An *outlier* is an observation with a large estimated error $|\hat{\epsilon}|$. An observation is *influential* if removing it substantially changes one or more β. How to detect outliers?

- Graph $\hat{\epsilon}_i$ versus y_i
- Generate and examine studentized residuals

The studentized residuals for observation i is $\frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ where h_{ii} is H_{ii} with $\underbrace{H}_{n\times n} = x(x'x)^{-1}x'$.

H is a symmetric matrix:

$$H' = x''(x'x)^{-1}x' = x(x'x)^{-1}x' = H$$
$$H^2 = x(x'x)^{-1}\underbrace{x'x(x'x)^{-1}}_{=I_k}x' = H$$

Because $H^i = H \;\forall\, i$, H is idempotent.

Experience shows that we need to examine observations for which $|\hat{\epsilon}| \geq 2$. Influential means the observation has *leverage*, thus we look for measures of leverage.

- One measure of leverage is h_{ii}.
  Rule of thumb: if $|h_{ii}| > \frac{2k}{n}$ with k number of observations and sample size n, examine observation i.

- Another common measure of leverage is the *Cook's distance*.

$$D_i = \frac{\hat{\epsilon}_i^2}{k \cdot MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \geq 0$$

  Rule of thumb: if $D_i \geq \frac{4}{n}$, examine observation i.

- Another measure of influence:

$$dffits_i = \frac{\hat{y}_i - \hat{y}_{i-(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

  with $\hat{y}_{(i)}$ as the estimated y_i after removing observation i and $\hat{\sigma}_{(i)}$ as the the standard deviation of ϵ after removing observation i.

  Rule of thumb: if $|dffits_i| \geq 2\sqrt{\frac{k}{n}}$, examine observation i.

- Alternatively, we can examine how observation i impact β_j using

$$dfbeta_{j,i} \equiv \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)}\sqrt{(x'x)^{-1}_{jj}}}$$

  with $i = 1, ..., n$ and $j = 0, 1, ..., k-1$. Rule of thumb: if $|dfbeta_{j,i}| \geq \frac{2}{\sqrt{n}}$, examine observation i. A limitation of this measure is that we should look at $(k-1) \times n$ *dfbeta* which grows to a large amount of measure values. Also, if we have more than one influential observation and they are clustered the estimate will not work well.

III.4.B. Multicollinearity

We have *multicollinearity* if one explanatory variables can be approximated by a linear combination of other explanatory variables. Without loss of generality, we can then say i.e.

$$x_i \cong \alpha_0 + \alpha_2 x_2 + \cdots \alpha_{k-1} x_{k-1}$$

In case the approximate is exact, there is perfect collinearity which means $(x'x)^{-1}$ would not exist. This often happens if you forgot to exclude a categorical variable as a baseline (for k categories we need to have exact $k-1$ binary, categorical, "dummy" variables). If the relationship is only approximate, you have a numerical problem: $(x'x)^{-1}$ is not calculated precisely and the standard errors of β are going to be inflated, it will look like the β are not statistically different from 0.

To detect multicollinearity, calculate *before* estimating your model $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the "R^2" from regressing x_j on the other explanatory variables.

Rule of thumb: $\begin{cases} \text{if } VIF_j \geq 10, \text{ there is a clear multicollinearity problem.} \\ \text{if } 5 \leq VIF_j \leq 10, \text{ there could be multicollinearity problem (gray area).} \\ \text{if } VIF_j \leq 5, \text{ there is no multicollinearity problem.} \end{cases}$

Remedies:

1. If the problem is mild ($MAXVIF \varepsilon [5, 10]$), do nothing but document the problem.
2. If the problem is more severe, drop some x_j.
3. Collect more data for the model.
4. *Mean center & standardize* explanatory variables: $x_j \rightarrow \zeta_j = \frac{x_j - \bar{x_j}}{s_j}$.
5. Use specialized estimators (e.g. ridge regression instead of OLS).

With the Cook's Distance and knowing $MSE = \hat{\sigma}^2$:

$$D_i = \frac{\hat{\epsilon}_i^2}{k\hat{\sigma}^2} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

Use percentiles of $F(k, n-k)$. Find the percentile corresponding to D_i.

$$\begin{cases} \text{If this percentile is } \leq 20\% \, i \text{ is not influential.} \\ \text{If this percentile is } \geq 50\% \, i \text{ is influential.} \end{cases}$$

In between, we cannot make a statement.

III.5. Box-Cox transformation

The *Box-Cox transformation* (B-C) can change non-normally distributed data to more normally distributed data. It applies to data which is > 0 because it is a power transformation. It can be used both for continuous explanatory and dependent variables or only for some of those.

$$y_i = \beta_0 + \sum_{j=1}^{k-i} \beta_j x_{ij} + \epsilon_i$$

$$\longrightarrow y_i^\lambda = \tilde{\beta}_0 + \sum_{j=1}^{k-i} \tilde{\beta}_j x_{ij}^\theta + \epsilon_i$$

But only in case x_{ij} is continuous, otherwise leave the variable unchanged.

$$\begin{cases} \text{If } \lambda = 0, \text{ replace } y_i^\lambda \text{ with } ln(y_i) \\ \text{If } \theta = 0, \text{ replace } x_{ij}^\theta \text{ with } ln(x_{ij}) \end{cases}$$

Definition III.1: Box-Cox transformation

$$x_{ij}^{*} = \begin{cases} \frac{x_{ij}^{\theta}-1}{\theta} \text{ if } \theta \neq 0 \\ ln(x_{ij}) \text{ if } \theta = 0 \end{cases}$$

λ and θ can be obtained by maximum likelihood, so the product of marginal likelihoods for observing our sample is maximized.

See also the density of a normal distribution as shown in example III.2.

$$\mathcal{L}(\beta, \sigma^2, \lambda, \theta | y, x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2} \left(\frac{y_i^{\lambda} - (\beta_0 + \beta_1 x_{i1}^{\theta} + \cdots + \beta_{k-1} x_{ik-1}^{\theta})}{\sigma} \right)^2 \right]$$

Example III.8: Continuity of Box-Cox

Let $x > 0$:

$$\lim_{\theta \to 0} \frac{x^{\theta} - 1}{\theta} \approxeq \frac{1 + \theta ln(x) - 1}{\theta} = ln(x)$$

$$x^{\theta} = e^{\theta ln(x)}$$

If θ is small, $e^{\theta ln(x)} \approxeq 1 + \theta ln(x)$. With the expansion:

$$e^{\epsilon} \approxeq 1 + \epsilon$$

$$e^{\epsilon} = \sum_{n=0}^{+\infty} \frac{\epsilon^n}{n!}$$

III.6. Tobit models

We frequently encounter datasets where observations are clustered at the lower and/or at the upper bound. This could result for example from observations going beyond what is measured or correspond to a different behavior. We cannot ignore this feature and simply estimate a linear regression model because in this case its estimator would be biased and inconsistent. One solution to deal with this problem is the *Tobit model*, named after James Tobin (1958).

Definition III.2: Tobit model

The version where the clustering of values is at lower bound 0:

$$y_i = \begin{cases} y_i^* \text{ if } y_i^* \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

where

$$y_i^* = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + u_i,\ u_i \overset{i.i.d.}{\sim} N(\sigma, \sigma^2).$$

y_i^* is a latent variable that is observable (unobserved) only when $y_i \geq 0$. Model parameters can be obtained by maximum likelihood.

$$\mathcal{L} = \prod_{i=1}^{n} \left[\frac{1}{\sigma}\varphi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right]^{d_i} \left[1\Phi\left(\frac{x_i\beta}{\sigma}\right)\right]^{1-d_i}$$

where

$$\begin{aligned} \varphi &\equiv \text{ the density of a standard normal distribution} \\ \Phi &\equiv \text{ the cumulative for a standard normal distribution} \\ \sigma &\equiv \text{ the standard deviation of } u_i \\ d_i &= \begin{cases} 1 \text{ if } y_i > 0 \\ 0 \text{ otherwise} \end{cases} \end{aligned}$$

For this model,

$$E(Y) = Pr(Y^* \geq 0) \cdot E\left[Y | Y \geq 0\right] + Pr(Y^* < 0) \cdot E\left[Y | Y < 0\right]$$

$$E(Y) = \Phi\left(\frac{x_i\beta}{\sigma}\right)\left[x_i\beta + \sigma\frac{\varphi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)}\right]$$

where $\frac{\varphi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)}$ is named *Mill's ratio*.

This model can be generalized to lower bounds different from 0, upper boundaries, and both.

$$y_i = \begin{cases} y_i^* \text{ if } y_L \leq y_i^* \leq y_U \\ y_L \text{ if } y_i^* < y_L \\ y_U \text{ if } y_i^* < y_U \end{cases}$$

For more details see Wooldridge (2016).

IV. Models for Discrete Dependent Variables

In transportation, there are many cases where the dependent variable of interest is *not continuous*.

Examples:

1. Binary: $\begin{cases} 1 \text{ if you travel to a national park this weekend} \\ 0 \text{ otherwise} \end{cases}$

2. Categorical: mode choice to LAX $\begin{cases} \text{rideshare} \\ \text{train} \\ \text{shuttle} \\ \text{car} \end{cases}$

3. Ranked data

4. Count data

IV.1. Models for binary outcomes

Typically, the values are 0 or 1. The models shown here are non-linear. They explain the probability of observing either 0 or 1 as an outcome.

IV.1.A. Statistical model

Three ways to introduce these models:

1. Latent variable

2. Random utility

3. Probability model

Latent variable approach

The latent variable approach (with latent in the meaning of *unobserved*) gives us the probability to observe one outcome for the dependent variable y.. Let y be a (0,1) binary variable. Let $x_1, x_2, ..., x_{k-1}$ be independent variables we consider using to explain y. The key idea is to assume there exists a latent continuous variable y^* which is related to the x as follows:

$$y_i^* = x_i\beta + \epsilon_i$$
$$y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ otherwise} \end{cases}$$

y_i^* can be interpreted as the potential that the event of interest will occur. From our model we know

$$\begin{aligned} Pr(y_i = 1|x_i) &= Pr(y_i^* > 0|x_i) \\ &= Pr(x_i\beta > 0|x_i) = Pr(-\epsilon_i < x_i\beta|x_i) \qquad = F(x_i\beta) \longrightarrow \end{aligned}$$

with $F(\cdot)$ as a cumulative distribution statement. Depending on the distributional assumption for ϵ_i, we get different models.

- If $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$, we obtain the binary probit model

$$Pr(y_i = 1|x_i) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

- If $\epsilon_i \overset{i.i.d.}{\sim} L(0, \frac{\pi^2}{3})$, we obtain the binary logit model

$$Pr(y_i = 1|x_i) = \frac{exp(x_i\beta)}{1 + exp(x_i\beta)}$$

  With $L(\cdot)$ as a logistic distribution. This term represents a *proper* probability because it returns values in the interval $(0, 1)$.

Definition IV.1: Odds function

The *odds function* is given by:

$$\Omega(x_i) = \frac{Pr(y_i = 1|x_i)}{Pr(y_i = 0|x_i)} = \frac{Pr(y_i = 1|x_i)}{1 - Pr(y_i = 1|x_i)} \quad \varepsilon(0, +\infty)$$

The odds function represents the ratio of the probability of the event happening

(outcome 1) over the event not happening (outcome 0).

For the logit model, the odds function becomes:

$$\begin{aligned}\Omega(x_i) &= \frac{\frac{exp(x_i\beta)}{1+exp(x_i\beta)}}{1-\frac{exp(x_i\beta)}{1+exp(x_i\beta)}} \\ &= exp(x_i\beta) = e^{x_i\beta} \\ \Rightarrow ln\left(\Omega(x_i)\right) &= x_i\beta\end{aligned}$$

IV.1.B. Interpretation

Logit model

It is easier to interpret because we have an explicit expression for the model probabilities. As an example, consider the case with 3 explanatory variables:

$$\begin{aligned}\Omega(x_1, x_2, x_3) &= exp(x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) \\ \Omega(x_1 + 1, x_2, x_3) &= exp(x_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \beta_3 x_3) \\ &= exp(x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) \cdot exp(\beta_1) \\ \Rightarrow exp(\beta_1) &= \frac{\Omega(x_1 + 1, x_2, x_3)}{\Omega(x_1, x_2, x_3)}\end{aligned}$$

So, for a unit change in x_i, the odds change by a factor e^{β_i} holding all other variables constant. Note that with $p = Pr(y = 1, x)$:

$$\begin{aligned}\Omega &= \frac{p}{1-p} \\ \Rightarrow \Omega(1-p) &= p \\ \Rightarrow \Omega &= p(1+\Omega) \\ \Rightarrow p &= \frac{\Omega}{1+\Omega}\end{aligned}$$

For the logit, another way to present results is to look at the percentage change in odds:

$$\frac{\Omega(x_1 + 1, x_2, x_3) - \Omega(x_1, x_2, x_3)}{\Omega(x_1, x_2, x_3)} = (e^{\beta_1} - 1)$$

Example IV.1: Probability of a event in a logit model

With the latent variable approach:

$$y_i = \begin{cases} 1 \text{ if } y_i^* \equiv x_i\beta + \epsilon_i \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

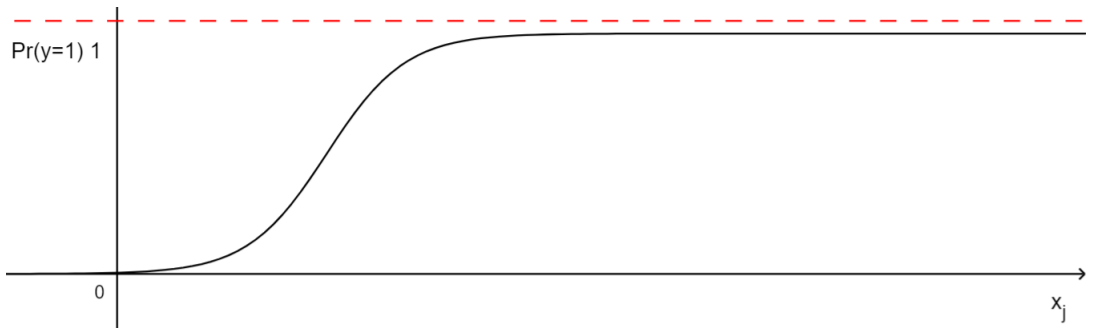
$$\epsilon_i \overset{i.i.d.}{\sim} L(0, \frac{\pi^2}{3}) \text{ logit model}$$

$$\longrightarrow Pr(y_i = 1|x_i) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

More general interpretation

Plot predicted provabilities versus explanatory variables as seen in Figure IV.1. Fix the other explanatory variables and only plot the change over one x_j (only works for the case of x_j being continuous). In case x_j is discrete, generate table of probabilities instead.

Figure IV.1: Logistic model



Marginal Changes

The third possibility is to report *marginal changes* in probabilities across the sample

$$\frac{\partial Pr(y_i = 1|x_i)}{\partial x_{ij}}$$

with x_{ij} as a continuous explanatory variable.

IV.1.C. Measures of fit

A number of different measures of fit have been proposed.

- $Count\, R^2 = \frac{\#\, of\, correct\, model\, predictions}{samplesize}$
- $McFadden's\, R^2 :\ \rho^2 = 1 - \frac{ln(\mathcal{L}^*_{full})}{ln(\mathcal{L}^*_{intercept})}$ where $\mathcal{L}^*_{intercept})$ is the model with no explanatory variables. A larger value indicates a better model.
- $McFadden's\, adjusted\, \hat{R}^2 :\ \hat{\rho}^2 = 1 - \frac{ln(\mathcal{L}^*_{full})-k}{ln(\mathcal{L}^*_{intercept})}$ with k as the number of β in the model (compare to $\bar{R}^2$ for linear regression).

The idea of the measures is that if a model is not explaining much, $ln(\mathcal{L}^*_{full})$ is going to be close to $ln(\mathcal{L}^*_{intercept})$ and so McFadden's (adjusted) R^2 will be close to 0.

Model parameters are obtained by maximum likelihood. The likelihood function for a random sample is given by

$$\mathcal{L} = \prod_{i=1}^{n} Pr(y_i = 1|x_i)^{y_i} \cdot (Pr(y_i = 1|x_i))^{1-y_i}$$

$$y_i = \begin{cases} 1 \\ 0 \end{cases}$$

$$\text{For the logit: } Pr(y_i = 1|x_i) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

$\mathcal{L}$ is a product of probabilities $\mathcal{L} < 1$ so $ln(\mathcal{L}) < 0$. Hence, $0 > ln(\mathcal{L}^*_{full}) > ln(\mathcal{L}^*_{intercept})$. In consequence,

$$\frac{ln(\mathcal{L}^*_{full})}{ln(\mathcal{L}^*_{intercept})} \leq 1 \text{ so, } 1 - \frac{ln(\mathcal{L}^*_{full})}{ln(\mathcal{L}^*_{intercept})} \geq 0.$$

IV.1.D. Hosmer-Lemeshow statistic

The *Hosmer-Lemeshow statistic* is used to asses if the model is well-specified. The idea is to compare predicted probabilities with observed data. The steps are:

1. Fit the model.
2. Compute predicted probabilities ($\hat{\pi}_i \equiv Pr(y_i = 1|x_i)$).
3. Sort the data by $\hat{\pi}_i$ from smallest to largest.
4. Divide observations in G groups ($G \approx 10$), so each group has $\approx \frac{10}{G}$ observations.

5. Within each group, calculate the mean prediction $\bar{\pi}_g = \left(\sum_{i \varepsilon g} \hat{\pi}_i \right) \frac{1}{n_g}$ with n_g as the number of observations in group g.

$$\bar{y}_g = \frac{1}{n_g} \sum_{i \varepsilon g} y_i$$

6. Calculate the Hosmer-Lemeshow statistic HL^*:

$$HL^* = \sum_{g=1}^{G} \frac{(n_g \bar{y}_g - n_g \bar{\pi}_g)^2}{n_g \bar{\pi}_g (1 - \bar{\pi}_g)}$$

If the model is well-specified, $HL^* \sim \chi^2(G-2)$.

IV.2. Models for nominal outcomes

An example for a nominal outcome is mode choice in transportation. This includes unordered variables with more than two outcomes.

IV.2.A. General model

Models are described in more detail in *Discrete choice methods with simulation* by Kennedy (2008).

Consider a decision maker who has different alternatives for a cause of action. We call this alternatives a *choice set*. We assume that

- these alternatives are mutually exclusive,
- the choice set is exhaustive (it contains all alternatives of interest), and
- the number of alternatives (elements in the choice set) is finite.

In addition, it is assumed that the decision maker is going to pick the alternative that maximizes his utility, denoted by U_{nj} where n is an index for the decision maker and j an index for the alternative.

Hence, alternative i is preferred if and only if $\forall\, j \varepsilon \{1, ..., J\}$, $U_{ni} \geq U_{nj}$ with J number of alternatives.

A researcher does not know the preferences of the decision makers. She observes some of their attributes (stored in vector s_n for decision maker n) and some characteristics of each

alternative (stored in vector x_{nj}). A function is specified that relates s_n ad x_{nj} to decision maker's n utility:

$$V_{nj} \equiv V(x_{nj}, s_n) \text{ for } j\varepsilon\{1, ..., J\},\ n\varepsilon\{1, ..., N\}$$

The representative utility V_{nj} is not equal to the actual utility U_{nj} because of incomplete information, expressed by an error term:

$$U_{nj} = V_{nj} + \epsilon_{nj}$$

Then, we can express the probability that person n picks alternative i as the probability its utility is higher:

$$\begin{aligned} P_{ni} &= Pr(U_{ni} \geq U_{nj}, \forall j\varepsilon\{1, ..., J\}) \\ &= Pr(V_{ni} + \epsilon_{ni} \geq V_{nj} + \epsilon_{nj}, \forall j\varepsilon\{1, ..., J\}) \\ &= Pr(\epsilon_{nj} - \epsilon_{ni} \leq V_{nj} - V_{ni}, \forall j\varepsilon\{1, ..., J\}) \end{aligned}$$

This is a cumulative distribution statement. Depending on the choice of distribution for ϵ_{nj}, we obtain a different model. In general, P_{ni} is a $J-1$ dimensional integral (because it expresses a difference). Such integrals are usually computationally challenging, especially for cases with large $J > 4$. The following assumptions simplify the calculation:

$$\epsilon_{nj} \overset{i.i.d.}{\sim} N(0, \sigma_{nj}^2) \tag{1}$$

In that case:

$$\epsilon_{nj} - \epsilon_{ni} \sim N(0, \sigma_{nji}^2) \tag{2}$$

Then,

$$\sigma_{nji}^2 = var(\epsilon_{nj}) - 2cov(\epsilon_{nj}, \epsilon_{ni}) + var(\epsilon_{ni}) \tag{3}$$

which leads to the multinomial probit model.

Instead, we assume $\epsilon_{nj} \overset{i.i.d.}{\sim} EV(1)$, the Gumbel distribution.
The CDF of a Gumbel (ω, η) is:

$$F(\zeta) = e^{-e^{\eta(\zeta-\omega)}}$$

The corresponding PDF is:

$$f(\zeta) = -\eta e^{-\eta(\zeta-\omega)} e^{-e^{\eta(\zeta-\omega)}}$$

Typically, we assume $\omega = 0$ and $\eta = 1$. For that assumptions, the variance of the Gumbel distribution becomes $\frac{\pi^2}{6}$. In that case, it can be shown that:

$$P_{ni} = \frac{e^{+V_{ni}}}{\sum_{j=1}^{J} e^{+V_{nj}}}$$

This result comes from the fact that the difference of two i.i.d. Gumbels is a logistic distribution. It is easy to check that $0 < P_{ni} < 1$ and $\sum_{j=1}^{J} P_{ni} = 1$.

For convenience, V_{nj} is assumed to be linear function of s_n and x_{nj}:

$$V_{nj} = \underbrace{y_{nj}}_{1xk} \underbrace{\beta_j}_{kx1}$$

Only the difference in utility is of matter. So, we can set the value of one alternative specific constant. The scale of utility is arbitrary, so we need to set the variance of the errors in our model.

IV.2.B. Measures of fit

The measures of fit are similar to the ones used for binary models (IV.1.C). We need to know the expression of the likelihood function:

$$\mathcal{L}(\beta|y,x) = \prod_{i=1}^{n} \prod_{j=1}^{J} P_{nj}^{\mathcal{G}(y_n=j)}$$

$$\mathcal{G}(y_n = j) = \begin{cases} 1 \text{ if } y_n = j \\ 0 \text{ otherwise} \end{cases}$$

IV.2.C. Interpretation

In general, we can use the same approaches as for binary models (IV.1.B).

Calculating predicted probabilities

$$P_{nm} = \hat{Pr}(y_n = m|x) = \frac{exp(x\hat{\beta}_{m|J})}{\sum_{j=1}^{J} exp(x\hat{\beta}_{j|J})}$$

This equation is for the logit model with alternative J as the baseline.

Special case of the multinomial logit model

$$Pr(y_i = m|x_i) = \frac{exp(x_i\beta_{m|J})}{\sum_{j=1}^{J} exp(x_i\beta_{j|J})}$$

Let

$$\begin{aligned}\Omega_{m|J}(x_i) &= \frac{Pr(y_i = m|x_i)}{Pr(y_i = J|x_i)} \\ &= x^{x_i\beta_{m|J}}\end{aligned}$$

Let $\Omega_{m|J}(x_i)$ be the odds that observations i selects alternative m given J as the baseline. Let then $\Omega_{m|J}(x_i, x_{il} + 1)$ be the odds obtained by adding 1 to the explanatory variable l. Then:

$$\frac{\Omega_{m|J}(x_i, x_{il} + 1)}{\Omega_{m|J}(x_i)} = e^{\beta_{m|J,l}}$$

Back to the general case

Select a *profile*, so pick the explanatory variables to correspond to a *typical decision-maker* with *typical choices* and for continuous explanatory variables, plot the change of probability P_{ni} over the explanatory variable $x_i l$. For discrete explanatory variables, create a table to store the different values for probabilities for the J alternatives when the discrete variables change by one unit.

Also, the change in predicted probabilities can be calculated for predicted probabilities:

$$\frac{\partial Pr(y_i = m|x_i)}{\partial x_{il}}$$

IV.2.D. Specification problems

Omitted variables

If a *relevant* explanatory variable is omitted, multinomial parameter estimates are inconsistent if

1. the omitted variable is correlated with included explanatory variable, or
2. the omitted variable is correlated across alternative outcomes or has a different variance for different outcomes.

The errors are not i.i.d.

In this case, the parameter estimators will be inconsistent.

Parameters are not quite stationary
In this case, the parameter estimators will be inconsistent as well (similar to the issues when estimating a linear model for measures which alter over time instead of observations constant for the model).

Only for multinomial logit: Independence of irrelevant alternatives (IIA)
This is specific problems for multinomial logit models (MNL) and does not apply to other models.

Recall that for $(m,n)\varepsilon\{1,...,J\}^2$:

$$Pr(y=m) = \frac{e^{x\beta_{m|J}}}{\sum_{j=1}^{J} e^{x\beta_{m|J}}}$$

$$Pr(y=n) = \frac{e^{x\beta_{n|J}}}{\sum_{j=1}^{J} e^{x\beta_{n|J}}}$$

$$\frac{Pr(y=m)}{Pr(y=n)} = \frac{e^{x\beta_{m|J}}}{e^{x\beta_{n|J}}}$$

This ratio does *not depend* on the characteristics of alternatives other than m and n. Adding or deleting alternatives should not affect such ratios.

This property (which applies only to MN logit models) is called *independence of irrelevant alternatives (IIA)* and should be tested.

For the test, estimated coefficients from a full model are compared to those of a restricted model (where restricted means that at least one alternative has been removed).

There are two common tests:

1. *Hausmann-McFadden (1984)*
   Step 1: Fit the full model with J alternatives $\hat{\beta}_F$
   Step 2: Fit a restricted model by eliminating one or more alternatives $\hat{\beta}_R$
   Step 3: Let $\hat{\beta}_F^*$ be the subset of $\hat{\beta}_F$ corresponding to $\hat{\beta}_R$
   Step 4: Calculate test statistic $HM = (\hat{\beta}_R - \hat{\beta}_F^*)' \left[\widehat{var}(\hat{\beta}_R) - \widehat{var}(\hat{\beta}_F^*)\right]^{-1} (\hat{\beta}_R - \hat{\beta}_F^*)$.
   If IIA holds, $HM \sim \chi^2(dim(\hat{\beta}_R))$.

2. *Small-Hsiao (1985)*
   The test randomizes the sample, so repeating the test will lead to different results.

Neither of the tests is very reliable.

IV.2.E. Extensions and testing of coefficients

- One coeffcient: t-test
- x-cofficients: F-test or likelihood test

Same as before, a number of extensions have been proposed for models for nominal outcomes, built around the logistic distribution but do not require the IIA to hold, for example *nested logit* or *cross-nested logit* (see Train (2009)).

IV.3. Models for ordered outcomes

In many transportation applications, we obtain ordered data. For example:

1. Quantitative ratings or rankings (scale 1 to 5)
2. Ordered opinions or reviews
3. Categorical frequency data (e.g. crash w/ fatalities, crash w/ injuries, w/o injuries)

If ordering of the is not taken into account (so it is taken as categorical), the estimator will lose efficiency. Linear regression would be inconsistent because the distance is not meaningful. However, just the fact that the values of a variable can be ordered *does not imply that an ordinal model is appropriate*. For example, answering "don't know" does not equate "neutral", which means the outcomes are not fully ordered.

IV.3.A. Statistical model

We rely on a latent-variable approach (McKelvey & Zavoina, 1975). Consider a dependent variable y that can take J different values $\{1, 2, ..., J\}$.

The model is then

$$\begin{aligned} y_i = m \varepsilon \{1, ..., J\} \text{ if and only if } \tau_{m-1} \leq y_i^* < \tau_m \text{ (Measurement model)} \\ \text{with } \tau_0 = -\infty \wedge \tau_J = +\infty \\ \text{where } y_i^* = x_i \beta + \epsilon_i,\ i \varepsilon \{1, ..., n\} \text{ (Structural model)} \end{aligned}$$

with the sample size n and τ as the $J - 1$ cutpoints or thresholds to estimate.

Example IV.2: Model for questionnaire results

We are analyzing a question with 4 outcomes: *Strongly disagree (SD), disagree (D),*

agree (A), strongly agree (SA)

$$y_i = \begin{cases} 1 = SD \text{ if } \tau_0 \equiv -\infty \leq y_i^* < \tau_1 \\ 2 = D \text{ if } \tau_1 \leq y_i^* < \tau_2 \\ 3 = D \text{ if } \tau_2 \leq y_i^* < \tau_3 \\ 4 = D \text{ if } \tau_3 \leq y_i^* < \tau_4 \end{cases}$$

$$y_i^* = x_i\beta + \epsilon_i$$

$$\begin{aligned} Pr(y_i = m|x) &= Pr(\tau_{m-1} \leq y_i^* < \tau_m|x_i) \\ &= Pr(\tau_{m-1} \leq x_i\beta + \epsilon_i < \tau_m|x_i) \\ &= Pr(\tau_{m-1} - x_i\beta \leq \epsilon_i < \tau_m - x_i\beta|x_i) \\ &= F(\tau_m - x_i\beta) - F(\tau_{m-1} - x_i\beta) \end{aligned}$$

Where F is the CDF of ϵ_i. Depending on the choice for distribution we of ϵ_i, we obtain a different model. We can estimate β and $\tau = (\tau_1, ...\tau_{J-1})$ using maximum likelihood. The likelihood function is:

$$\mathcal{L}(\beta, \tau|y, x) = \prod_{i=1}^{n}\prod_{j=1}^{J}\left[F(\tau_j - x_i\beta) - F(T_{j-1} - x_i\beta)\right]^{\delta_{ij}}$$

$$\text{where } \delta_{ij} = \begin{cases} 1 \text{ if } y_i = 1 \\ 0 \text{ otherwise} \end{cases}$$

As usual, we need to take the log before maximizing this expression.

For the ordered logit model, we have:

$$Pr(y = m|x_i) = \frac{e^{\tau_m - x_i\beta}}{1 + e^{\tau_m - x_i\beta}} - \frac{e^{\tau_{m-1} - x_i\beta}}{1 + e^{\tau_{m-1} - x_i\beta}}$$

IV.3.B. Interpretation

Because ordered models are not linear, their interpretation is more complex and requires more than only one approach.

Predict probabilities for different profiles
Calculate $\widehat{Pr}(y_i = j|x_i)$, $j\varepsilon\{1, ..., J\}$ for selected values of x.

Visualize predicted probabilities
The predicted probabilities of continuous explanatory variables are graphed as a function.

Create table of predicted probabilities
The predicted probabilities of binary explanatory variables are can be showed in tables.

Marginal change in probabilities
For continuous variables, you can calculate marginal changes

$$\frac{\partial Pr(y_i = m|x_i)}{\partial x_{il}}$$

where $m\varepsilon\{1, ..., J\}$ and x_{il} as the l-th explanatory variable (assumed to be continuous) for person i.

Odds ratio for ordered logit model (OLM)

$$\begin{aligned}
Pr(y_i = m|x_i) &= \frac{e^{\tau_m - x_i\beta}}{1 + e^{\tau_m - x_i\beta}} - \frac{e^{\tau_{m-1} - x_i\beta}}{1 + e^{\tau_{m-1} - x_i\beta}} \\
Pr(y_i \leq m|x_i) &= \frac{e^{\tau_m - x_i\beta}}{1 + e^{\tau_m - x_i\beta}} - \frac{e^{\tau_{m-1} - x_i\beta}}{1 + e^{\tau_{m-1} - x_i\beta}} \\
&\quad + \frac{e^{\tau_{m-1} - x_i\beta}}{1 + e^{\tau_{m-1} - x_i\beta}} - \frac{e^{\tau_{m-2} - x_i\beta}}{1 + e^{\tau_{m-2} - x_i\beta}} \\
&\quad \vdots \\
&\quad + \frac{e^{\tau_1 - x_i\beta}}{1 + e^{\tau_1 - x_i\beta}} - \frac{e^{\tau_0 - x_i\beta}}{1 + e^{\tau_0 - x_i\beta}} \\
\Longrightarrow Pr(y_i \leq m|x_i) &= Pr(y_i = m|x_i) + Pr(y_i = m-1|x_i) + \cdots + Pr(y_i = 1|x_i) \\
\Longrightarrow Pr(y_i \leq m|x_i) &= \frac{e^{\tau_m - x_i\beta}}{1 + e^{\tau_m - x_i\beta}} \\
Pr(y_i > m|x_i) &= 1 - \frac{e^{\tau_m - x_i\beta}}{1 + e^{\tau_m - x_i\beta}} = \frac{1}{1 + e^{\tau_m - x_i\beta}} \\
\frac{Pr(y_i \leq m|x_i)}{Pr(y_i > m|x_i)} &= e^{\tau_m - x_i\beta} \equiv \Omega_{\leq m/>m}(x_i) \\
\frac{\Omega_{\leq m/>m}(x_i, x_{il} + 1)}{\Omega_{\leq m/>m}(x_i)} &= \frac{e^{\tau_m - (\beta_l x_{i1} + \cdots \beta_1 (x_{il}+1)) + \cdots + \beta_k x_{ik}}}{e^{\tau_m - (\beta_l x_{i1} + \cdots \beta_1 x_{il}) + \cdots + \beta_k x_{ik}}} = e^{-\beta_l}
\end{aligned}$$

This states that for a unit increase in x_{il}, the odds that an outcome is $\leq m\varepsilon\{1, .., J\}$ change by a factor $e^{-\beta_l}$ holding all other variables constant.

IV.3.C. Measures of fit

The measures of fit for models with ordered outcomes are similar to the measures of fit for binary and multinominal logit models (IV.1.C and IV.2.B).

- *Count* R^2 as the ratio of the number of correctly predicted outcomes by the number of observed units.
- $'s\, R^2$
- $McFadden's\, adjusted\, R^2$

IV.3.D. Hypothesis testing

Individual coefficients
Three steps for hypothesis testing:

1. $H_0 :\ \beta_l = \beta_{l0}\ vs.\ H_1 : <\neq, >$
2. Use $z^* = \frac{\hat{\beta}_l - \beta_{l0}}{\hat{\sigma}_{\hat{\beta}_l}}$
   If the model is well specified, $z^* \sim N(0,1)$ (standard normal).
3. Conclude the results.

Multiple coefficients

IV.3.E. Specification Problems

The problems of multinominal models (IV.2.D) apply here as well, but for models with ordered outcomes there is one additional issue.

Parallel regression assumption
Recall that for $m\varepsilon\{1, ..., J\}$:

$$Pr(y = m|x) = F(\tau_m - x\beta) - F(\tau_{m-1} - x\beta)$$
$$\tau_0 = -\infty\ \wedge\ \tau_J = +\infty$$

As a result, $Pr(y \leq m|x) = F(\tau_m - x\beta)$. For example, in the case of $J = 4$ we have the probabilities as shown in GRAPHIC

$$Pr(y \leq m|x) = F\,(\,\tau_m \underbrace{\qquad x\beta}_{-\text{independent of } m}\,)$$

We have the same β for each probability. This allows models for ordered outcomes to be more efficient than normal, categorical models with different parameters β for different categories, which are the more general form:

$$Pr(y \leq m|x) = F(\tau_m - \underbrace{\qquad x\beta_m}_{\text{dependent of } m}\qquad)$$

It is however a good idea to test if the assumption is correct that the β in the statements $Pr(y \leq m|x) = F(\tau_m - x\beta_m)$ are equal, so a joint test along the more general model to asses whether it can be simplified to a model for ordered outcomes. There are different test available for this, most common is the *Brant test* (1990).

Usually, it worths checking this assumption if then a model for ordered data can be used. These are attractive for ordered data because they are much more concise than just categorical models (by factor J).

IV.4. Models for count data

Count data is very common in transportation, some example include

- the number of (#) vehicles waiting at a toll plaza,
- # of trucks waiting at a warehouse,
- # of accidents on a given road segment during a given time period, or
- # of cars owned by households in a census tract.

Linear regression is not appropriate for this kind of question. Instead, some common count models are used. As for other discrete data models, count data models predict the probability of observing different outcomes.

Poisson model

The *Poisson model* is one of the most common discrete count models. It is named after French mathematician Siméon Denis Poisson (1781-1840).

IV.4.A. Statistical model

Let N_i be a random variable that generates count data. Then,

$$Pr(N_i = n) = \frac{e^{-\lambda_i}\lambda_i^n}{n!}$$

with $\lambda_i = e^{x_i\beta}$ as the arrival rate and $n\varepsilon\mathbb{N}$. This relation comes from the exponential expansion.

$$e^{-\lambda} = 1 + \frac{(-\lambda)^1}{1!} + \frac{(-\lambda)^2}{2!} + \cdots \Rightarrow e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$$

Clearly, $Pr(N_i = n) \geq 0$. Let us show that $\sum\limits_{n=0}^{+\infty} Pr(N_i = n) = 1$.

$$\begin{aligned}\sum_{n=0}^{+\infty} Pr(N_i = n) &= \sum_{n=0}^{+\infty} e^{-\lambda_i}\frac{\lambda_i^n}{n!} = e^{-\lambda_i}\sum_{n=0}^{+\infty}\frac{\lambda_i^n}{n!} \\ &= e^{-\lambda_i}e^{+\lambda_i} = e^0 = 1\end{aligned}$$

So, $Pr(N_i = n)$ is a proper probability.

$$\begin{aligned}E(N) &= \sum_{n=0}^{+\infty} nPr(N = n) = \sum_{n=0}^{+\infty} n\frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}\sum_{n=1}^{+\infty}\frac{\lambda^n}{(n-1)!} \\ &= e^{-\lambda}\sum_{n=1}^{+\infty}\frac{\lambda\cdot\lambda^{n-1}}{(n-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{+\infty}\frac{\lambda^k}{(k)!} \text{ with } k = n-1 \\ &= \lambda e^{-\lambda}e^{+\lambda} = \lambda = E(N)\end{aligned}$$

Using the knowledge of $E(N) = \lambda$, the variance can be calculated.

$$
\begin{aligned}
var(N) &= \sum_{n=0}^{+\infty}(n-\lambda)^2 e^{-\lambda}\frac{\lambda^n}{n!} \\
&= \sum_{n=0}^{+\infty}(n^2 - 2n\lambda + \lambda^2)e^{-\lambda}\frac{\lambda^n}{n!} \\
&= \underbrace{\sum_{n=0}^{+\infty} n^2 e^{-\lambda}\frac{\lambda^n}{n!}}_{=S_1} - \underbrace{2\sum_{n=0}^{+\infty} e^{-\lambda}\frac{\lambda^n}{n!}}_{=S_2} + \underbrace{\lambda^2 \sum_{n=0}^{+\infty} e^{-\lambda}\frac{\lambda^n}{n!}}_{=S_3} \\
S_1 &= \sum_{n=0}^{+\infty}\left[\underbrace{n(n-1)}_{\to S_1^a} + \underbrace{n}_{\to S_1^b}\right] e^{-\lambda}\frac{\lambda^n}{n!} \\
S_1^a &= \sum_{n=1}^{+\infty}(n-1)e^{-\lambda}\frac{\lambda^n}{(n-1)!} = \sum_{n=1}^{+\infty} e^{-\lambda}\lambda^2\frac{\lambda^{n-2}}{(n-2)!} = e^{-\lambda}\lambda^2\sum_{n=2}^{+\infty}\frac{\lambda^{n-2}}{(n-2)!} \\
&= \lambda^2 e^{-\lambda}\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!} = \lambda^2 e^{-\lambda}e^{+\lambda} = \lambda^2 \qquad \text{with } k = n-2 \\
S_1^b &= \sum_{n=1}^{+\infty} e^{-\lambda}\frac{\lambda\lambda^{n-1}}{(n-1)!} = \lambda \\
S_2 &= \lambda\lambda = \lambda^2 \\
S_3 &= e^{-\lambda}\sum_{n=0}^{+\infty}\frac{\lambda^n}{n!} = 1 \\
\Rightarrow var(N) &= S_1^a + S_1^b - S_2 + S_3 = \lambda^2 + \lambda - 2\lambda^2 + \lambda^2 = \lambda = E(N)
\end{aligned}
$$

The equality of variance and expected value is called equidispersion.

IV.4.B. Estimation

The β_0 in $\lambda_i = e^{x_i\beta}$ with x_i as a row vector and β as a column vector can be estimated by maximum likelihood.

$$\mathcal{L}(\beta|x) = \prod_{i=1}^{n} e^{-\lambda_i(\beta)}\frac{\lambda_i^{n_i}(\beta)}{n_i!}$$

with n_i as the observed counts and

$$e^{-\lambda_i(\beta)}\frac{\lambda_i^{n_i}(\beta)}{n_i!} = Pr(N_i = n_i)$$

as the probability of n_i to be observed. The log of this function is numerically easy to estimate. The Poisson ML estimation is consistent, asymptotically normal and asymptotically efficient.

IV.4.C. Interpretation

Consider a discrete change δ in continuous variable x_{il}.

$$\frac{E\left(N_i|x_i, x_{il}+\delta\right)}{E\left(N_i|x_i\right)} = \frac{e^{\beta_0+\beta_1 x_{1l}+\cdots+\beta_l(x_{il}+\delta)\cdots\beta_{k-1}x_{ik-1}}}{e^{\beta_0+\beta_1 x_{1l}+\cdots+\beta_l(x_{il})\cdots\beta_{k-1}x_{ik-1}}}$$
$$E\left(N_i|x\right) = \lambda_i = e^{x_i\beta}$$

So if $\delta = 1$, we get e^{β_l}. This means, β_l is the log of the ratio of expected counts when x_l is increased by 1 unit.

If x_l is continuous, the elasticity of λ_i with respect to x_l is

$$\frac{\partial\lambda_i}{\partial x_l}\frac{x_l}{\lambda_i} = \beta_l\lambda_i\frac{x_l}{\lambda_i} = \beta_l x_l,\ \lambda_i = e^{x_i\beta}.$$

The marginal change in λ_i with respect to x_l is

$$\frac{\partial\lambda_i}{\partial x_l} = \beta_l\lambda_i.$$

IV.5. Measures of fit

The measures of fit are similar to the methods of pseudo-squared deviance estimators used for other discrete models (IV.1.C, IV.2.B, and IV.3.C).

- $McFadden's\, R^2:\ \rho^2 = 1 - \frac{max\, logLikelihoodFullModel}{max\, logLikelihoodRestricteModel} = \frac{LL_f(\beta^*)}{LL_r(0)}$
- $McFadden's\, adjusted\, R^2$
- $Count\, R^2$

IV.6. Testing

Negative binomial regression model (NBR)
In many cases, $E(N) \neq var(N)$. If $E(N) > var(N)$, we name it underdispersion, in case $E(N) < var(N)$ it is called overdispersion.

One possibility to deal with this problem is to add an error term to the equation of λ_i.

$$\lambda_i = e^{x_i\beta+\epsilon_i}$$

If $e^{\epsilon_i} \overset{i.i.d.}{\sim} \Gamma(1,\alpha)$, we obtain the NBR. For the NBR, $var(N_i) = E(N_i) + \alpha E^2(N_i)$.

$$Pr(N_i = m) = \frac{\Gamma(n+\alpha^{-1})}{\Gamma(\alpha^{-1})n!}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)\left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^n$$

where

$$\Gamma(\zeta) = \int_0^{+\infty} e^{-t}t^{\zeta-1}dt$$
$$\Gamma(n) = (n-1)!$$

The estimation of NBR models is similar to the Poisson models:

$$\mathcal{L} = \prod_{i=1}^{n} Pr(N_i = n_i)$$

Also, the interpretation, measuring of fit and testing follow the same framework. An additional test for the equidispersion of α can be necessary.