Contents

0.1 Week 1

0.1.1 Describe the conceptual framework of statistical inference.

Statistical data analysis makes use of the construction of a statistical model formulated on the basis of certain assumptions. Laws are formulated using probabilistic theory and based on certain assumptions, and random variables represent how measured characteristics can vary from observation to observation. Statistical inference allows us to go back from the description of a sample to that of a larger whole and allows one to verify hypotheses formulated on the phenomena under study. In the study of collective phenomena it is necessary to define the event of interest, identify the collectivity in which the phenomenon occurs and choose the characteristics of the community considered of interest for understanding the phenomenon. The key components of this framework include:

- Population represents that set of units identified as equal in the problem under study. The population is the entire set of subjects/things of interest.
- A sample is the set of subjects from the population for which data are available. The ideal method of picking out a sample to study is called random selection.
- Probability is important since the main concepts in statistics are expressed in terms of variables and their related probability distributions.

0.1.2 Describe the inductive and deductive process involved in process of statistical science.

The process of statistical science uses both inductive and deductive reasoning.The inductive approach consists of three stages: Observation $\rightarrow$ Seeking patterns $\rightarrow$ Developing a theory or general (preliminary) conclusion. The deductive research approach consists of four stages: Start with an existing theory and create a problem statement $\rightarrow$ Formulate a falsifiable hypothesis, based on existing theory $\rightarrow$ Collect data to test the hypothesis $\rightarrow$ Analyze and test the data $\rightarrow$ Decide whether you can reject the null hypothesis.

In other words, inductive reasoning moves from specific observations to broad generalizations. Deductive reasoning works the other way around.

0.1.3 Provide an example of discrete quantitative variable which is time invariant. Provide an example of a categorical variable which is time-varying.

Provide an example of discrete quantitative variable which is time invariant. Provide an example of a categorical variable which is time-varying: Values of variables of interest may be detected according to a specific instant of time, some variables are always time invariant, such as the number of planets of our solar system. An example of a categorical variable that is time-varying would be a person's occupation.

0.1.4 State the main definitions of: sample unit, random sample, and population.

Sample unit: A sample is the set of subjects from the population for which data are available.

Random sample: Is a sample selected by using a method called random selection. This procedure uses a random method such as a computer-generated list of random numbers to select a unit so that each unit in the population has an equal chance of being selected. In this way the sample elements are selected randomly from the population, independently of any characteristics, and the statistical sampling theory is able to account for the expected variation.

Population: The population is the entire set of subjects/things of interest. The collective may be finite or potentially infinite when it consists of many units whose amount is not precisely detectable.

0.1.5 Provide a definitions for: random variable and sample space.

Random variable: A random variable is a mathematical abstraction that can serve a model for observable quantities. It is a function that assigns a numerical value to each point in the sample space.

Sample space: The sample space for a random phenomena is the set of all the possible outcomes.

0.1.6 Considering a continuous random variable define the cumulative distribution function and its properties.

Let X be a random variable, its probability distribution is also specified by F that is its cumulative distribution function. The probability $P(X \leq x)$ that a random variable takes values $\leq$ x is called cumulative probability. The cumulative distribution function is

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

for all real numbers x.

We recall the properties of cumulative distribution functions. Let X be a random variable with cumulative distribution function $F_X(t) = P(X \leq t)$. Then

- for every $t \in R$ we have $0 \leq F(t) \leq 1$,
- F is a non - decreasing function,
- $\lim_{t\to-\infty} F(t) = 0$, $\lim_{t\to+\infty} F(t) = 1$
- F is right-continuous.

The area under the function over an interval of values, which equals its integral over that interval, is the probability that the random variable falls in that interval.

0.1.7 Define percentiles and provide a definition for the Q-Q plot.

We can characterize a probability distribution by dividing points, which are called percentiles. The $(100p)$th percentile, $0 < p < 1$, is a point π_p such that

$$P(X \leq \pi_p) = p$$

and

$$P(X > \pi_p) = 1 - p$$

So, p is the solution of the equation $F(\pi_p) = p$. The Q-Q plot (quantile-quantile plot) is a graphical comparison of observed data with a theoretical distribution. It is the plot the k-th smallest observation against the expected value of the k-th smallest observation out of n in a standard normal distribution. The point is that in this way you would expect to obtain a straight line if data come from a normal distribution with any mean and standard deviation.

0.1.8 Define the empirical cumulative distribution function. Does it have any properties?

The empirical cumulative function is defined as the fraction of data smaller than or equal to x. Given a random sample of size n from a random variable X with cumulative distribution function $F(x)$ and given ordered sample values $x_i \leq x_i + 1$, $i = 1, ..., n-1$, is called empirical cumulative distribution function the function F^n with values in the interval $[0, 1]$ which assigns to each x its sample weight $1/n$:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x_i)$$

where $I(\cdot)$ is the indicator function,

$$\begin{cases} I(X_i \leq x_i) = 1 & \text{if } X_i \leq x_i \\ I(X_i \leq x_i) = 0 & \text{otherwise} \end{cases}$$

Therefore, $F^n(x)$ is the sample proportion of the n observations that fall at or below x.

In the situation of total absence of information about the law of probability F of a random variable from which the sample is drawn, the empirical cumulative distribution function can be assumed as an estimate of $F(x)$ due to some important properties.

0.2 Week 2

0.2.1 Describe the bivariate Gaussian distribution and list the characteristics of this distribution

Suppose we have two random variables, X and Y, with a joint probability density function given by $f(x, y)$. To calculate the probability that X lies between a and b and Y lies between c and d, we can integrate the joint density function over the region defined by $a < X < b$ and $c < Y < d$, as follows:

$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x, y), dx, dy$$

The joint distribution of X and Y is said to be a bivariate Gaussian (or normal) distribution if any linear combination of X and Y, denoted as $W = aX + bY$ for constants a and b, results in a random variable that has a univariate Gaussian distribution.
Parameters of this joint distribution are:

- The means μ_x, μ_y
- The variances σ_x^2, σ_y^2
- The correlation coefficient ρ_{xy}

The characteristics of this distribution are:

- If the variables X and Y are uncorrelated in a bivariate Gaussian distribution (i.e., (X,Y) $\sim$ N($\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy} = 0$)), then they are also independent (a result that does not generally apply to other distributions), and their marginal distributions are Gaussian distributions. Specifically, the joint probability density function of X and Y with zero correlation is given by:

  $$f(x, y; \rho_{x,y} = 0) = f(x)f(y)$$

  where $f(x)$ and $f(y)$ are the marginal probability density functions of X and Y, respectively.

- If (X,Y) follows a bivariate Gaussian distribution, then marginalizing with respect to X or Y yields univariate Gaussian distributions. Specifically, if (X,Y) $\sim$ N$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$, then the marginal distribution of X is given by:

$$X \sim N(\mu_x, \sigma_x^2)$$

and the marginal distribution of Y is given by:

$$Y \sim N(\mu_y, \sigma_y^2)$$

- If (X,Y) follows a bivariate Gaussian distribution, then the conditional distributions $(X|Y = y)$ and $(Y|X = x)$ are Gaussian distributions. In particular, the conditional expectation and variance of X given Y=y are given by:

$$E[X|Y = y] = \mu_x + \frac{\rho_{x,y}\sigma_x\sigma_y}{\sigma_y^2}(y - \mu_y),$$

$$Var(X|Y = y) = \sigma_x^2\left(1 - \rho_{x,y}^2\right)$$

Likewise, the conditional expectation and variance of Y given X=x are given by:

$$E[Y|X = x] = \mu_y + \frac{\rho_{x,y}\sigma_x\sigma_y}{\sigma_x^2}(x - \mu_x)$$

$$Var(Y|X = x) = \sigma_y^2\left(1 - \rho_{x,y}^2\right)$$

where $\rho_{x,y}$ is the correlation coefficient between X and Y.

- The contour lines of a bivariate Gaussian distribution are ellipses, where the direction of the major axis is defined by the correlation coefficient ρ_{xy}. A contour plot of a bivariate Gaussian density function is a two-dimensional plot that shows curves on which the density function $f(x, y)$ is constant. These curves are defined by N_c contour levels f_j, $j = 1, \ldots, N_c$, where N_c is the number of contours to be plotted.

0.2.2 Describe the variance-covariance and the correlation matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

where $\sigma_i^2 = \text{Var}(Y_i)$ and $\sigma_{ij} = \text{Cov}(Y_i, Y_j)$ for $i \neq j$. This is a squared symmetric matrix (since $\sigma_{ij} = \sigma_{ji}$) and is positive semidefinite. The correlation matrix of $Y_1, \ldots, Y_n$ is defined as

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{12} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \cdots & 1 \end{bmatrix}$$

where ρ_{ij} is the correlation between Y_i and Y_j ($\rho_{ii} = 1$). Also, this matrix is squared, symmetric, and positive semidefinite.

0.2.3 Explain the difference between the raw and partial correlation coefficient

The raw correlation coefficient measures the strength and direction of the linear relationship between two variables without taking into account the effect of other variables that may be related to them. It is calculated as the covariance between the two variables divided by the product of their standard deviations, and it ranges from -1 to 1. On the other hand, the partial correlation coefficient measures the strength and direction of the linear relationship between two variables after the influence of other variables has been removed. It is calculated as the correlation between the residuals of two variables that are obtained by regressing them on all other variables in the model. The partial correlation coefficient reflects the unique association between two variables that cannot be explained by other variables in the model.

0.2.4 What we mean when we speak of a spurious association?

A spurious association occurs when the apparent association between two variables is influenced by a third variable that, if not taken into account, would make the observed association between the other two different or absent. For instance, suppose that we find a high positive correlation between height and individual income. If income and height are positively correlated with gender, as in Italy, then the observed correlation between income and

height may be an indirect association phenomenon. In such cases, the association between two variables should be assessed while controlling for other potentially associated variables. Therefore, to avoid spurious associations, it is essential to consider the relevant confounding variables that may influence the relationship between the two variables of interest.

0.2.5 If two normal random variables have zero correlation they are also independent? Why?

If Y and X are statistically independent, then $Cov(Y, X) = 0$, and we say that they are uncorrelated. However, it is important to note that zero covariance does not necessarily imply independence, as there may be other types of associations or relationships between the variables, such as non-linear relationships or conditional dependence.

0.2.6 What is a contour plot? Describe why we observe ellipse.

A contour plot is a graphical representation of a two-dimensional surface, in which the plotted quantity, typically a density function $f(x, y)$, is displayed using a series of one-dimensional curves or lines of equal values, known as contours. The number of contours plotted, denoted by N_c, is arbitrary and can be chosen to best display the features of interest in the data. Each contour curve, defined by $f_j = f(x, y)$, represents a specific constant value of the density function. The contour plot is a useful tool for visualizing and analyzing patterns in data that vary across two dimensions. The ellipses' orientation provides a visual cue for the nature of the correlation between the two random variables whether it is positive or negative. Additionally, the center and length of the semi-axes can provide valuable insights into the data.

0.3 Week 3

0.3.1 Lists and comment on the properties of the estimators.

The primary characteristics of an estimator are as follows:

- An estimator T is considered unbiased if its expected value is equal to the parameter being estimated, i.e., $E_\theta(T) = \theta$. If the expected value is not equal to the parameter for some values of θ, T is considered biased, and the bias of T is defined as $B_\theta(T) = E_\theta(T) - \theta$. Ideally, an estimator $\hat{\theta}$ should approach the true value of θ as the sample size n increases. The mean squared error (MSE) of an estimator is a measure of its variability with respect to the parameter. The MSE is defined as $MSE(T) = E(T-\theta)^2 = V(T)+Dist(T)^2$. If the estimator is unbiased, then the MSE is equal to its variance, $V(T) = E[T - E(T)]^2$. Thus, the MSE has two components: one representing the variability of the estimator, and the other representing its bias.

- An estimator is said to be consistent if it converges to the true value of the parameter in probability as the sample size increases. In other words, the sample distribution of the estimator becomes more and more concentrated around the true value of the parameter as n increases. Consistency is a property that holds in the limit as n approaches infinity. If an estimator is not consistent, it is generally not considered suitable for use.

- An efficient estimator is one that tends to be closer to the true value of the parameter, on average, than other estimators.

0.3.2 Describe the likelihood function and the maximum likelihood estimation. Provide an example.

The method known as maximum likelihood estimation (MLE) is used to find the parameter value for which the observed data is most likely to occur. Assuming a particular family of probability distributions, MLE estimators are consistent, asymptotically unbiased, and asymptotically efficient. The likelihood function, denoted as L, or the logarithm of the likelihood (log-likelihood), denoted as ℓ, is not a probability distribution but describes the support that the observed data give to the possible parameter values. Due to its properties, the maximum likelihood method is the most widely used technique for deriving estimators since it allows us to choose the most plausible parameter values based on the observed data and chosen model.

Given a sample $x = (x_1, \ldots, x_n)$ of n independent and identically distributed observations drawn from a population with an unknown parameter

$\theta \in \Theta \subseteq R$, where $f(x;\theta)$ denotes the joint probability function defined according to the family of distributions, the likelihood function is defined as

$$L(\theta) = f(x_1;\theta)\dots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

The plot of the likelihood function portrays the probability of the observed data for all possible values of the parameter, but it typically does not integrate to 1. In practice, $L(\theta)$ provides the evidence in favor of any single value of θ in Θ. If $L(\theta^{(1)}) > L(\theta^{(2)})$, where $\theta^{(1)}$ and $\theta^{(2)}$ are two values of θ, then the probability of the observed sample is larger under $\theta^{(1)}$, and so more evidence exists in favor of this value of θ. The likelihood function allows for the ordering of different values of θ according to the "degree of likelihood" they get from the data in the sample. The maximum likelihood estimation proposes an estimate of θ (if it exists) as a "most likely" value, which is the value of θ that maximizes the likelihood function $L(\theta; x_1, \dots, x_n)$. It is natural to estimate θ as the value of the parameter that maximizes $L(\theta)$. This leads to the maximum likelihood estimate, which is formally defined as the value $\hat{\theta} = \hat{\theta}(x)$ such that

$$L(\hat{\theta}) = \sup_{\theta\in\Theta} L(\theta). \tag{1}$$

Note that the symbol $\hat{}$ over a parameter is called caret and read as hat. Consequently, the maximum likelihood estimator (MLE) is $\hat{\theta}$. In practice, the MLE is the parameter value for which the observed sample is the most likely. In many situations, finding the maximum likelihood estimator $\hat{\theta}$ requires solving an optimization problem using differential calculus. It is usually easier to maximize the log-likelihood instead of the likelihood $L(\theta)$:

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta),$$

since the logarithm is a monotonically increasing transformation, and the two problems are equivalent. The natural logarithm of the likelihood function $\log L(\theta) = \ell(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$ is also a monotonically increasing transformation of the function and is easier to maximize to derive the point estimate and interval estimate of the parameter. In some cases, the maximization problem for a vector of parameters $\theta = (\theta_1, \dots, \theta_k)$ with $k > 1$ can be solved by solving the system of likelihood equations:

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0, \qquad i = 1, \dots, k,$$

which can find one or more candidates for the MLE since the first derivative being zero is only a necessary condition for a maximum. Then, to confirm that a maximum has been found, we check that the matrix of the second derivatives, with elements

$$\left.\frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}} < 0, \qquad i = 1, \ldots, k, \quad \text{and} \quad j = 1, \ldots, k,$$

is negative definite.

One example can be to consider a random variable X that represents a success event, such as the fact that the return of bonds in a certain portfolio is greater than 3%. Suppose we want to find the proportion of bonds with a return higher than 3%, given a random sample where we observe $x_1 = 1$, $x_2 = 0$, and $x_3 = 1$. These three random variables X_1, X_2, and X_3 have a Bernoulli distribution with probability $P(X > 3) = p$. We want to find the value of $\hat{p}$, which is the maximum likelihood estimator of the parameter p, that maximizes the following likelihood:

$$L(x_1 = 2, x_2 = 0, x_3 = 1; p) = P(X_1 = 1)P(X_2 = 0)P(X_3 = 1) = p^2(1 - p)$$

Note that the above expression derives from the fact that $P(X_i = 1) = p^{1-1}(1 - p)^1$. We want to find $\hat{p}_{ML}$, which maximizes $L(p)$ for $p \in [0, 1]$. To do this, we first find the log-likelihood:

$$l(p) = \log L(p) = 2 \log p + \log(1 - p)$$

We then take the first derivative of the log-likelihood with respect to p:

$$\frac{\partial \ell(p)}{\partial p} = \frac{2}{p} - \frac{1}{1 - p} = 0$$

which results in $\hat{p}_{ML} = 2/3$. The second derivative is always negative, indicating that the maximum likelihood estimate of the probability is 0.667. In general, the maximum likelihood estimator of the parameter p, given a random sample $X_1, \ldots, X_n$ drawn from a Bernoulli distribution, is:

$$\hat{p}_{ML} = \frac{\sum_{i=1}^{n} X_i}{n}$$

0.3.3 What are the properties of the estimator obtained through the maximum likeli- hood method?

The properties are:

- MLEs can be biased (although this does not apply to the multiple linear regression model). However, bias decreases as sample size increases: MLEs are asymptotically unbiased, meaning that as n increases, any bias they have decreases to 0;
- The exact sample distribution can sometimes be difficult to obtain, but, as we shall see, MLEs have a truly remarkable property: their asymptotic sample distribution is generally always a Gaussian distribution;
- MLEs are consistent: as n increases, the estimator converges towards the parameter value;
- MLEs are asymptotically efficient: other estimators do not have smaller standard errors and do not tend to fall closer to the parameter.

0.3.4 What we refer to when we mention the Fisher Information. For which purpose do we employ these quantities?

Fisher Information is a statistical tool that quantifies the amount of information contained in a sample regarding an unknown parameter. It provides a measure of how much the probability distribution of the sample changes with variations in the parameter value. The Fisher information is defined as

$$\bar{I}(\theta) = E_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right] = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]$$

where usual regularity conditions are assumed.
The Fisher information, in the context of a model with assumed regularity conditions and a sample of n independent observations, provides a measure of the amount of information on the parameter θ provided by the sample. Specifically, $\overline{I}(\theta)$ is the expected value of the square of the first derivative of the log-likelihood function with respect to θ, while $I(\theta)$ is simply n times $\overline{I}(\theta)$. This information can be visualized as a $p \times p$ matrix of the second derivatives

of the log-likelihood function with respect to each pair of parameters, with the sign changed. A high value of the Fisher information indicates that there is a region in the parametric space with a high likelihood. Overall, $\overline{I}(\theta)$ provides a measure of the amount of information provided by a single observation, while $I(\theta)$ measures the amount of information provided by a sample of size n.

0.3.5 Can you provide a brief description of what we mean when we refer to the Bayesian Inference?

Bayesian inferences using the posterior distribution, where it plays a role similar to that of the classical confidence interval in frequentist statistics. Specifically, a credible interval is an interval that contains a specified percentage of the posterior density $g(\theta|y)$, analogous to a confidence interval in classical statistics. For example, a 95% credible interval for θ is the region between its 2.5th and 97.5th percentiles of the posterior distribution. This simple method of constructing a posterior interval based on percentiles of the posterior distribution is widely used in practice.

0.3.6 What is it the nonparameteric bootstrap?

The bootstrap is a powerful tool for statistical inference that allows the assessment of the accuracy of any estimator or algorithm, regardless of its complexity. With modern computational resources, an infinite sequence of future trials can be numerically implemented, making the bootstrap a remarkable feature of computer-age statistical inference. One of the advantages of the bootstrap over maximum likelihood is its ability to compute standard errors and other quantities in settings where mathematical formulas are not available. In the non-parametric context, where only sample information is available, resampling methods allow the data to "speak as much as possible" by extrapolating all available information about the parameter θ. Resampling methods involve iterative reuse of the sample and encompass all modern methods aimed at evaluating and improving the accuracy of complex estimators, for which analytical forms of standard errors are often not available. Accuracy evaluation is a fundamental phase of the inference process, and the bootstrap is a computational resampling method that treats the sample distribution as if it were the population distribution. This method's simplicity, combined with the availability of increasingly inexpensive computing power,

has led to its rapid and wide-ranging development and application in various fields of statistics, including inference, regression models, time series analysis, and sampling theory, among others. The standard error of $\bar{X}$ describes the variation of the sample mean $\bar{x}$ from sample to sample of the same size n, while the term standard deviation σ refers to the population.

0.3.7 Which is the main use of this procedure?

In non-parametric contexts, the available information is limited to the sample data. Therefore, the aim is to extract as much information as possible from the data, allowing them to "speak" for themselves. Resampling methods are employed based on this principle, whereby the sample is iteratively reused to extract the maximum amount of information possible in the absence of any prior information that would allow hypotheses to be formulated about the underlying distribution $F(x)$.

0.3.8 How we can applied it and under which assumptions?

Resampling methods, in the absence of prior information that allows hypotheses to be formulated about F(x), the data are extensively exploited through the iterative reuse of the sample.

0.3.9 Describe the bootstrap distribution of the estimator. Why we use it to construct confidence intervals?

Bootstrap is introduced as a basic tool for inference since it may assess estimation accuracy of any estimator (and algorithm) no matter how complicated. In statistics a measure of accuracy of an estimate is provided by the associated standard error, without the standard error we cannot use the confidence interval formula of a point estimate. There are many estimators (and algorithms) not having a direct mathematical formulas to calculate standard errors. The advantage of the bootstrap over the maximum likelihood is that it allows us to compute maximum likelihood estimates of standard errors and other quantities in settings where no mathematical formulas are available.

Bootstrap represents a computational resampling method that treats the sample distribution as if it is the population distribution. The resampling procedure provided by Bootstrap is identical to that of the original sample, or mimics the original sampling.

0.3.10 Describe the bootstrap percentile method to obtain a confidence interval for the parameter.

Confidence interval estimation aims to give us an idea of where an unknown parameter, like θ, might lie based on the information we get from a sample. Instead of just giving us a single value for θ, a confidence interval gives us a range of possible values that θ might fall within, along with a probability that the true value of θ is contained within that range. For example, if we have a 95% confidence interval for θ, it means that there is a 95% chance that the true value of θ is within that interval. We usually choose the confidence level, like 95%, before we collect any data. Once we have our data, we can calculate the lower and upper limits of the interval, called TL and TU, based on the data we collected. It's important to note that the probability we choose, like 95%, applies to the interval itself, which is random, and not to the true value of θ, which is fixed but unknown.

0.4 Week 4

0.4.1 Write the extended formula of the multiple linear regression model and comment on each quantity.

In the multiple linear regression model the expected value of the response variable is specified conditional on the observed value of the explanatory variables. The response is continuous.

The extended notation of the multiple linear regression model when two explanatory variables are considered, is the following:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

when we have more parameter we can express it like:

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i$$

The coefficients are the model parameter to be estimated. Where:

- β_0 is the intercept
- β_j for $j > 0$ are the regression coefficients referred to the explanatory variables
- ϵ_i is a term referred to as the random component or unit-specific error, is a random variable that enter in the model in adaptive way ($E[\epsilon_i] = 0$ and $var(\epsilon_i) = \sigma^2$).

An alternative formula is possible:

$$E[Y] = \mu_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_j$$

0.4.2 Write the multiple linear regression model in matrix notation and specify the dimension of each quantity.

The multiple linear model can be represented by the following set of equations: In terms of matrix algebra, $\mathbf{Y}$ is the column vector representing the

$$\begin{aligned} Y_1 &= \beta_0 + \textstyle\sum_{j=1}^{m} \beta_j x_{1j} + \epsilon_1 \\ Y_2 &= \beta_0 + \textstyle\sum_{j=1}^{m} \beta_j x_{2j} + \epsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \textstyle\sum_{j=1}^{m} \beta_j x_{nj} + \epsilon_n \end{aligned}$$

response and the model is expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{Y}_{n\times 1}$ is the vector collecting the response variable
- $\mathbf{X}_n \times (p+1)$ is the matrix of explanatory variables, a non-stochastic matrix of full rank
- $\boldsymbol{\beta}_{(p+1)\times 1}$ is the vector of the model parameters

- $\boldsymbol{\epsilon}_{n\times 1}$ is the vector of the error terms. We assume that:
  - it is a random variable $E[\epsilon_i] = 0$ and constant variance $var(\epsilon_i) = \sigma^2$ for all i
  - error are assumed independent: $cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

0.4.3 Specify the assumptions required for the multiple linear regression model.

Non trovata su teaching notes

The multiple linear regression model makes the following assumptions:

- Linearity: The relationship between the independent variables and the dependent variable is linear. This means that the effect of each independent variable on the dependent variable is constant over the range of values of that variable.
- Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for any other observation.
- Homoscedasticity: The variance of the residuals (the difference between the predicted value and the actual value of the dependent variable) is constant across all values of the independent variables. This means that the spread of the residuals is the same for all values of the independent variables.
- Normality: The residuals follow a normal distribution. This means that the distribution of the residuals is bell-shaped.
- No multicollinearity: The independent variables are not highly correlated with each other. This means that there is no linear relationship between any two independent variables. If there is high correlation between any two independent variables, it may be difficult to determine which variable is causing the effect on the dependent variable.

0.4.4 Describe the estimation method employed in the classical linear regression.

The method employed in the classical linear regression is the least square method. This method (like the maximum likelihood) allow to deriving parameter estimates by minimizing the sum of squared deviation between observed and fitted values. Thus the regression plane is maximally consistent with the observed points with respect to the chosen Euclidean distance.

The system of equation become:

$$D(\beta_0, \beta_1, ..., \beta_n) = arg\min \sum_{i=1}^{n} [y_i - [\beta_0 + \beta_1 xi1 + ... + \beta_n x_{in})]^2$$

which is a system of $p + 1$ partial derivatives. Form this we get

- $\hat{\beta}_0$ is the estimator of intercept
- $\hat{\beta}_j$ for $j > 0$ is the estimator of the conditional regression coefficient referred to x_j.

The least squared estimates provide the prediction equations closest to the data, minimizing the sum of squared residual.

0.4.5 In which way can we decompose the variability of the response?

Thanks to some properties of the least square estimators we can derive the proportion of the total variance that is explained by the multiple linear regression model measuring the goodness of fit. We start form the total sum of square (TSS) that measure the empirical variance of the observe responses.

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

TSS can be decomposed. The decomposition work by adding and removing the fitted values, and then thanks to the residual properties some therms are zero. This lead to the empirical variance of the observed response is derived as an additive decomposition into the empirical covariance of the residual and the fitted values.

Total sum of square is the sum of squared residual plus the sum of squares due to the regression model.

$$TSS = SSE + SSR$$

So the viability is divided in the variability that the model can explain and the residual one. This decomposition is used in the index R^2.

0.4.6 How is the least squares estimator for the intercept obtained?

The parameter $\hat{\beta}_0$ for the intercept is obtained by difference as it follow:

$$\hat{\beta}_0 = \bar{y} - (\sum_{j=1}^{m} \hat{\beta}_j \bar{x}_j)$$

0.4.7 How do we interpret the values of the estimated regression coefficients?

The interpretation for the estimated regression coefficients is the following:

- The estimator of intercept ($\hat{\beta}_0$): provide the conditional expected value of the response variable when all the explanatory variables are equal to zero
- Estimator of the conditional regression coefficient referred to x_j ($\hat{\beta}_j$ for $j > 0$): measures the influence of variable x_j on the response holding fixed all the other variables. Basically it indicate the expected change in the response when x_j increases by one unit, holding fixed the other variables. It is a conditional effect adjusted for all the other explanatory variables.

0.4.8 What is the multiple R-squared and the adjusted R-squared?

The multiple R-squared is a relative index that measure the goodness of fit.

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

I takes values in $[0, 1]$, a value of 1 means that the residual variance is null, so the observed and the fitted values are the same; while a value of 0 means the t the enplaned variability of the model is null, so the explanatory variables do not contribute to explain the variability of the response. A value of 0.7 means that the model explain the 70% of the total variability of the response.

The adjusted R-squared is a version of the R-squared index that account the fact that the index increases when explanatory variables are added into the model.

$$R^2_{adj} = 1 - \frac{(n-1)}{n-(p+1)}(1-R^2)$$

where $n-p-1$ are the degree of freedom when there are $(p-1)$ parameters.

0.4.9 Why R-squared is adjusted?

R^2 is used to evaluate if the model is correctly specified. An high value of this index unfortunately doesn't directly mean that the model is correctly specified, because R-squared increases each time we add a new explanatory variable the model. For this reason we need the adjusted version, because R^2_{adj} doesn't monotonically increase when more explanatory variable are added to a model. However the penalty is to small, so other indices are used to evaluate the goodness of fit.

0.4.10 What is the residual standard error, and how is it interpreted? How are defined the predicted (fitted values)? And the residuals?

The residual standard error (RSE) is defined as:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p+1)}}$$

A lower value of S indicate a better model, because observed and fitted values are closer together. Commonly index used for evaluating the goodness of fit of a model is based on this index. It square form is an unbiased estimator for the error terms.

The RSE has at denominator the sum of squared residuals. The residual referred to the i observation is defined as de difference between the observed

value and the fitted value $y_i - \hat{y}_i$, where $\hat{y}_i$ is the fitted value. The fitted value is defined by

$$\hat{\mu}_i = \hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{ij}$$

and it basically an estimates of the conditional expected value of the response. $\hat{\mu}_i = \hat{y}_i$ is the fitted value and is determined at fixed values of the covariates x_{ij}.

0.4.11 Specify the properties of the residuals and explain how we expect to be the scatterplot of the residuals by id number in the case the assumptions of the model are satisfied.

Residuals are inspected to check empirically the tenability of the model assumptions. The properties of residuals are the following:

- I propriety: the sum of residuals is zero when the model includes the intercept.
- II property: residuals are orthogonal with respect to each explanatory variable, so the underlying random variable Z is uncorrelated with each covariate x_j.
- III property: Z and $\hat{Y}$ are uncorrelated, so their covariance is zero, and the residual are orthogonal to the fitted values.

In plots of residuals in case of assumption of the model are satisfied:

- Positive residual and negative residuals are balanced
- Residuals fluctuate randomly
- The variability of residuals is constant

0.4.12 What particular points (leverage, anomalous and influential) can be determined by the graphical inspection of the residuals and by which graph?

The Leverage points can be identified by the plot of residuals against each explanatory variable. If one ore more values are distant from the others they

are considered leverage points. Leverage points highly influence the model fit and the estimated parameter value is not valid.

The plot of the fitted values against the residuals, or the standardized/studentized residual can help to identify outliers. Outliers are observation that do not seem to follow the same data-generating process, usually this values have high value of the standardised or studentized residual.

0.4.13 Which are the measures we employ to detect these particular points numerically?

Leverage is a measure of an observation's potential influence of the fit. The diagnostic for mesuring leverage points is the following:

$$h_{ii} = \frac{1}{n} + \frac{(z_i - \bar{z})^2}{SSE}$$

this take value in $[0, 1]$ and large is the value unusual is the relative covariate value. Leverage points not always cause problems.

Point corresponding to a particularly high value of the standardised or studentized residual is an outlier, it is measure as it follow:

$$m_i = \frac{z_i}{s(1 - \mathbf{x}_i^T(\mathbf{X}_i\mathbf{X}_i)^{-1}\mathbf{x}_i)^{1/2}}$$

where s is an estimate of the conditional standard deviation σ. Outliers are observation that do not seem to follow the same data-generating process, it is not always appropriate to eliminate them. Some models can handle them.

An influent value contributes highly to determine the fitted values. It is determined using the Cook's distance. This distance measure the changes on the estimated coefficient when a single observation is removed from data.

$$d_i = \frac{z_i^2 h_{ii}}{(p+1)s^2(1 - h_{ii})^2}$$

It is a non negative measure, and usually values over 1 identify an influential value, that should be exterminated.

0.4.14 Which are the properties of the least squared estimators estimators?

The least squares estimates are equal to those obtained applying the maximum likelihood estimates. Least square hold the following properties:

- least squares estimator is unbiased for β so $E[\beta] = 0$. For Gauss-Markov theorem the least squares estimator has minimal variance and so is the best linear unbiased estimator (BLUE), is the most efficient.
- variance-covariance matrix of the least squared estimators is given by $Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X}) - 1$.
- $\hat{\beta}$ is distributed according to a multivariate Gaussian distribution. $\hat{\beta} \sim N(\beta_k, \sigma^2_{kk})$ where σ^2_{kk} is the k-th element in the diagonal $(\mathbf{X}^T\mathbf{X})^{-1}$.

0.4.15 Which is the distributional assumption for the error terms? And what does it implies for the model.

In the basic formulation of the multiple linear regression model it is assumed that the error term is a continuous random variable with the following moments : $E[\epsilon_i] = 0$, $var(\epsilon_1)$. The variance $\sigma^2 > 0$ is constant , this means that it does not vary with the statistical units. The errors are assumed to be independent of each other and , therefore, uncorrelated, so the covariance is zero. So we assume that the errors are normally distributed. If we have unequal variance in the errors terms this means that there is heteroskedastic and may be caused by the presence of outliers or an incorrectly specified model. If we run regression with heteroskedasticity the model would have unreliable predictions.

0.4.16 What does it mean the assumption of homoschedastity? How the variance of the error terms is estimated?

The assumption of homoscedasticity of the errors is important because, if it is fulfilled, means that the variance of the errors is constant so the model is able to give reliable predictions. The variance of the error terms, which is assumed as constant and denoted with σ^2 , is an unknown parameter. An unbiased estimator of the error variance in a linear model having intercept and p explanatory variables is represented by:

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - (p+1)}$$

That is $E[S^2] = \sigma^2$.

0.4.17 Which is the distribution of the response variables when we assume a multivariate distribution for the error terms?

When we assume a multivariate distribution for the error terms also the response variable follow a multivariate normal distribution because the errors are also normally distributed.

0.4.18 In which way we calculate confidence intervals for the regression coefficients?

A confidence interval for a regression coefficient is determined by considering the following statistics having a Student-t distribution with n-p-1 degrees and the following formula:

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{S^2(X'X)^{-1}_{kk}}}$$

Where the denominator of this formula denotes the standard error of $\hat{\beta}_j$. When the sample size is high, the ratio has an approximate standard Norma Distribution. A confidence interval for $\hat{\beta}_k$ at confidence level $(1 - \alpha)$ is:

$$\pm t^*_{1-\alpha/2,n-p-1}$$

where $t^*_{1-\alpha/2,n-p-1}$ is the quantile of the Student-t distribution.

0.4.19 In a model where we try to explain diastolic blood pressure according to age of the patients and body mass index, and we estimate a confidence interval at 95% at a confidence level of 95% of (0.387; 0.634), how do we interpret these values?

The diastolic blood pressure value, according to age and mass index, with a confidence level of 95% range between 0.387 and 0.634. The 95% is not the probability that the point estimate fall in the interval but means that the interval includes the point estimate 95% of the times.

0.4.20 Why we perform the F-test?

The F-test is used to verify the global null hypothesis of the null model. The null hypothesis is $\beta_k = 0$ for all k, $k = 1, ..., p$, and if it should not be rejected it means that the response for each i, $i = 1, ..., n$, can be expressed as just a signal, the intercept β_0 plus an error term ϵ_i and therefore the response is independent of all the covariates (null model). F test statistic values evidences against H_0.

0.4.21 How it is performed the F-test?

First of all we need to consider the full model with p explanatory variables that is: $Y = X\beta + \epsilon_i$ and i is compared under H_0 with the following null model: $Y = I\beta_0 + \epsilon$, where none of the explanatory variables in the model have an effect on the response. Under the null hypothesis: $Y \sim N_n(\beta_0 I, \sigma^2 I)$. The observed sums of square residuals (SSE): $SSE = \sum_i (y_i \hat{y}_i)^2$ and for the null model: $TSS = \sum_i (y_i - \overline{y}_i)^2$. Larger values of $(TSS - SSE)$ give stronger evidence against H_0. Under the previous assumption of the model with p explanatory variables the distribution of residual deviance SSE/σ^2 is proportional to a chi-squared distribution with $n - p - 1$ degrees of freedom. Under the null hypothesis TSS/σ^2 has a distribution proportional to a chi-squared distribution with $df = n - 1$. For testing H_0 the distribution of the test statistics is provided by the ratio of two independent chi-squared distributions divided by their degrees of freedom and the resulting distribution of the test statistic under the null hypothesis is a Fisher-Snedecor F with $df_1 = p$, $df_2 = n - p - 1$ degrees of freedom:

$$F = \frac{(TSS - SSE)/p}{SSE/[n - (p + 1)]} \sim F_{p,n-p-1}$$

Let α be the significant level, we reject the null hypothesis if the test statistic is larger than the $(1 - \alpha)$ quantile of the corresponding F distribution. When the value of the observed test statistic f_{oss} is higher than the critical value identified on the basis of the reference distribution $F_{p,n-p-1}$ considering the significance level, there is no empirical evidence in favor of the null hypothesis and it is reject.

0.4.22 In a model where we try to explain diastolic blood pressure according to age of the patients and body mass index the result of the F-test is the following F-statistic 82.59 on 2 and 726 degrees of freedom p-value $<$ 0.00002. Which are the observed quantities and how we interprete these results?

The value of the f_{oss} statistic is equal to 82.59 points and the p-value associated is very small, < 0.00002, so according to these results we can reject the null hypothesis. H_0 states that the null model is the best one where none of the explanatory variables in the model have an effect on the response, but in this case the diastolic blood pressure (response) is influenced by the patient age and body mass index.

0.4.23 With reference to the previous example, which is the test we use to validate the hypothesis that $H_0 : \beta_{age} = 0$ vs $H_0 : \beta_{age} \neq 0$?

If we consider the significance test of $H_0 : \beta_{age} = \beta_{H_0}$ with respect to $H_1 : age \neq H_0$ the test statistic is a t-Student distribution:

$$t_{\beta_k} = \frac{\hat{\beta}_k - \beta_{H_0}}{SE_{\hat{\beta}_k}}$$

which is the number of standard errors that $\hat{\beta}_{age}$ falls from the H_0 value of 0 if $H_0 : \beta_{age} = \beta_{H_0} = 0$. It's null t distribution has $df = n - p - 1$. The global test F provide strong evidence that at least one β_k is different from 0 but the t inferences reveals a statistically significant individual effect.

0.4.24 If the results of the previous test are the following $\hat{\beta}_{age} = 0.336$ and t-value = 0.362 and p-value = 0.791. Describe the reported quantities and explain the test result.

The t-value is the result reported by the t-Student statistic and is quite low (0.362) , but in order to understand if we can reject the null hypothesis or

not we need to control the p-value. The p-value is the probability, presuming that H_0 is true, that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_α Smaller p-value reflect stronger evidence against H_0, in the following case p-value is 0.791, is quite high for these reason we accept the null hypothesis, there is not enough evidence against H_0.

0.4.25 Which is multicollinearity? Can you pride an example with an illustration?

When there is collinearity there is linear associations between two explanatory variables. Multicollinearity is present if we are in a situation where more than only two explanatory variables are highly linearly related. If we have perfect multicollinearity the matrix of covariates $X^T X$ is non-invertible and this implies that there are not unique least squares estimates.

0.4.26 Which measure is employed to assess the presence of excessive collinearity? What we can do to solve the problem?

One measure to check if there's collinearity is the Variance Inflation Factor (VIF):

$$VIF(\hat{\beta}_k) = \frac{1}{1 - R^2}$$

VIF represent the multiplicative increase in variance of the variance of $\hat{\beta}_k$ due to their linear dependence of x_k to the other covariates. The VIF measure how much is the variance of $\hat{\beta}_k$ due to multicollinearity. In general high value of VIF indicates that explanatory variables are linearly associated and my be necessary to:

- construct a single combined variable from the variables which are higly correlated.
- omit some variables because the information they provides is redundant (feature selection)
- use another approach to estimate the model parameters named ridge regression.

0.4.27 Variable selection is performed with information criteria. Which are the most popular and which is the principle behind their usage?

The information criteria allow for selecting the most useful covariates from among those available. The information criteria consider the measure of distance between two density (entropy), among these the most important are the Bayesian Information Criterion (BIC) and the Akaike Information criterion (AIC).

- The BIC index is written with respect to the log-likelihood function and is defined as follow:

$$BIC = -2\hat{l}(\theta) + \#par \log(n)$$

  where $\hat{l}$ is the value of the log-likelihood of the model at the point of maximum with a number of p explanatory variables, which is weighted by the number of parameters ($\#par$) with respect to the logarithmic number of observations n available. Smaller values of BIC indicates better models.

- The Akaike information criterion is written with respect to the model deviance as follow:

$$AIC = n \log(z'z) + 2\#par$$

  where $Dev(z)$ is the residual deviance of the model. The best model is the one with lowest value of this index.

0.4.28 Automatic procedures are used jointly with the information criteria. Which are they and how they differ?

Step-wise method use a restricted search through the space of potential models using information criteria for choosing between models. These indices are calculated for each model by adding or removing one covariate at a time and using stepwise testing approaches that compare successive models. The three procedures , mainly automatic search strategies, are:

- The procedure backward starts with all the covariates in the model and removes an eligible variable at each iteration if the increment it makes to the AIC index is negligible. The procedure stops when all variables have been evaluated and the model has been reduced to a model with fewer parameters. This algorithm can identify variables that are predictive in combination but not individually.

- The procedure forward reverses the background method, it start with no covariates in the model and then for all predictors not in the model, the AIC value is considered when they are added in the model. The procedure stop when no new prediction can be added.

- The procedure stepwise is addresses to to the situation where variables are added or removed early in the process. The criteria may lead to choosing models with different numbers and types of variables, so application requirements must also guide the choice of the model.

0.5 Week 5

0.5.1 How prediction is performed in the multiple linear regression model?

In the multiple linear regression model, prediction is performed by using the estimated regression coefficients to calculate the predicted value of the dependent variable for a given set of values of the independent variables. Specifically, if we have a multiple linear regression model with p independent variables $(x_1, x_2, ..., x_p)$ and a dependent variable y, the model can be expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the estimated regression coefficients, and ϵ is the random error term. To make a prediction for a new observation with values of the independent variables $x_1, x_2, ..., x_p$, we plug these values into the model equation and solve for the predicted value of y:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

where $\hat{y}$ is the predicted value of y for the given values of the independent variables.

0.5.2 Describe a prediction confidence interval

The prediction interval is the range that likely contains the value of the dependent variable for a single new observation given specific values of the independent variables. The confidence interval gives an idea of the precision of the prediction, and it helps to assess how reliable the model is in making predictions for new observations.

0.5.3 In which way categorical variables are added into the model?

There are categorical covariates, also called factors, that may be included in the linear regression model. Categorical variables must be coded, for example, a binary variable can be coded into two categories $j = 0$ for not having the feature and $j = 1$ otherwise, but other codings for j are possible. To obtain identifiability in a model including a covariate with J categories, we exclude the parameter of one category (also defined as a baseline or reference category). For example, assuming a variable with $j = 1, ..., J$ categories, with $J = 4$, the multiple linear regression model is written according to the following equation

$$Y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

for $i = 1, ..., n$. we have variable x_i called dummy where $x_{ij} = 1$ if the i-th unit has been assigned to the factor level j and is zero otherwise.

0.5.4 To what we refer when we compare two parallel regression lines?

When we compare two parallel regression lines we probably are comparing the two observations generated from a linear regression model having a binary categorical variable, this model in fact is also defined as model with parallel lines.

0.5.5 Describe the Binomial distribution and report and example on real data which can be considered as realized observations from this model

The Binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent and identical trials, where each trial has two possible outcomes (success or failure) and the probability of success is constant across all trials.

We call X a binomial random variable and n and p are the parameters of X: n is called the size parameter, and p is the probability of success. We usually write:

$$X \sim \text{Bimomial}(n, p) \text{ or } X \sim B(n, p)$$

If $X \sim B(n, p)$, then its support is $\{0, 1, 2, ..., n\}$and its probability mass function is

$$f(k) = P(X = k) = \begin{cases} \frac{n!}{k!(n-k)!} p^k q^{n-k} & k \in \{0, 1, 2, ..., n\} \\ 0 & \text{oherwise} \end{cases}$$

An example of real data that can be modeled by the Binomial distribution is the number of successful free throws made by a basketball player in a fixed number of attempts, where each attempt has two possible outcomes (success or miss) and the probability of success is assumed to be constant for each attempt. For instance, if a basketball player attempts 20 free throws and makes 15 of them, we can model the number of successful free throws as a Binomial random variable with $n = 20$ and $p = 0.75$ (assuming a constant probability of making a free throw for each attempt). In this case, the PMF of the Binomial distribution can be used to calculate the probability of making exactly k free throws in 20 attempts for different values of k.

0.5.6 Describe the logistic regression model

The logistic regression model is used when the dependent variable is binary, and this is explained using p explanatory variables that can be either quantitative or qualitative. Sample observations are assumed to be realizations from a binomial distribution. The logit of the probability of interest (success) is expressed as a linear function of the explanatory variables, which may be either continuous or categorical; the continuous variables may be expressed

as deviations from the mean, or they can be squared; binary of categorical variables may be included, as well as interaction terms.

Let $p_i = P(Y_i = 1)$, let $logit(p_i)$ denote $\frac{log(p_i)}{(1-p_i)}$. The logistic regression model for binary data is formulated as

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

where x_i denotes the column vector of p explanatory variables $(x_{i1}, ..., x_{ip})$ observed for unit i, $i = 1, ..., n$.

0.5.7 Define the deviance used to compare models

A comparison between nested models can be performed with the likelihood ratio test which uses the deviance. We denote $l(\hat{\theta})$ the maximized log-likelihood and let $l(y)$ denote the maximum achievable log-likelihood that is the value which could be obtained with a saturated model that is a model with 0 degrees of freedom. The problem of this model is that it does not smooth the data and it is not parsimonious. We just employ the saturated model as a benchmark for constructing a measure called as likelihood ratio statistic that compare the saturated model with the chosen one.

$$D(y; \hat{\theta}) = 2log[\frac{\text{maximum likelihood for saturated model}}{\text{maximum likelihood for chosen model}}$$

$$D(y; \hat{\theta}) = 2[l(y) - l(\hat{\theta})]$$

0.5.8 For which type of data is employed the multinomial logit model?

The multinomial distribution is an extension of the binomial where the response can take more than two values. This distribution refers to the number of times that k-th modality (or level or category) of a categorical covariate is observed. We refer to n independent trials and to $X_1, X_2, ..., X_k$ random variables representing the number of time a collection of k events is observed and $p_1, p_2, ..., p_k$ are the probabilities of each event. The distribution is specified under the following constraints:

$$p_1 + p_2 + ... + p_k = 1$$

0.5.9 In the multinomial logit model the probability that the random variable assume a certain category j can be expressed as...

In the multinomial logit model, the probability that the dependent variable takes on a specific category j can be expressed as the exponential function of a linear combination of the independent variables, divided by the sum of the exponential functions for all categories.

$$P(Y_i = j) = p_{ij} = \frac{e^{x_i' \beta_j}}{1 + \sum_{j=2}^{J} x_i' \beta_j}$$

The linear combination of the independent variables is estimated using maximum likelihood estimation, and the resulting coefficients are used to predict the probability of each category for new observations.

0.6 Week 6

0.6.1 Describe the finite mixture models.

The finite mixture models allow us to explore the structure of the data in inferential terms and determine groups of units, called clusters, by assuming the reference population to be heterogeneous. The units are classified in distinct groups that are homogeneous with respect to a characteristic of interest. These models are useful when dealing with unobserved heterogeneity.

The mixture components, on which the cluster are based, can be seen as the densities of the subpopulations, and the mixing weights are the proportions of this subpopulations to the overall population. In this type of model the choice of components requires particular care.

In Mixtures of Gaussian distribution we have $D = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ where $x_i \in X$ are assumed to be iid and with the following density function that is the sum of the weights and multivariate Gaussian density function with a mean vector and a variance covariance matrix. Cluster correspond to subpopulations distributed as Gaussian density function with different parameter values, in order to estimate this parameter a two step procedure is applied. First we get the maximum likelihood estimation of parameters, then a maximum a posteriori procedure (MAP) is applied in order to assign each unit to the mixture component with the largest posteriori conditional probability.

0.6.2 How the number of components is selected?

The issue of the choice of the number of components in the mixture and the appropriate variance-covariance matrix can be addressed using the BIC.

0.6.3 Which are the structure of the variance covariance matrix that you know and which are the feature of each one?

The variance-covariance matrix of the least squared estimators is given by

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

X is a non stochastic matrix of full rank.

Under the Gaussian distribution for the error terms $\epsilon \sim N_n(0, \Sigma)$ where 0 is a vector of x errors and Σ is a diagonal matrix with elements σ^2 on the main diagonal. Therefore $Y \sim N_n(\mu, \sigma^2 I)$ provided that $\hat{\beta} = (X'X)^{-1}X'y$ and $\hat{\beta}$ is distributed according to a multivariate Gaussian distribution.

So the distribution of the least squares estimator is

$$\hat{\beta} =\sim N_{p+1}(\beta, \sigma^2(X'X)^{-1})$$

0.6.4 When and for which scope finite mixture models can be used?

Finite mixture models are applied to data with two main purpose:

- provide an appealing semi parametric framework in which to model unknown distributional shapes.
- provide a probabilistic clustering of the data into a finite number of clusters.

These models allow to explore the structure of the data in inferential terms and determine groups of units (clusters) by assuming the reference population to be heterogeneous.

0.6.5 In which way units are classified into clusters?

The assumption of an heterogeneous reference population make possible to classify the units into distinct groups that are homogeneous with respect to the characteristic of interest, similar to cluster analysis, which allow a set of units to be grouped in such a way that the units in the same group are more similar to each other than those in other groups.

0.6.6 This model also provides estimation of the density?

Yes, mixture models also provide estimation of the density. The mixture components can be seen as the densities of the subpopulations and the mixing weights are the proportions of each subpopulation in the overall population. Mixture models with components having a Gaussian distribution are widely used because they are mathematically tractable and it is quite easy to derive maximum likelihood estimates of the parameters.