Statistical models

May 29, 2024

Contents

Chapter 1

Introduction

1.1 Statistical models: general definitions

Regression analysis consists in the investigation of the relationship that can be expressed as an equation connecting a *response/dependent variable* to one or more explanatory/predictor variables in the following steps:

1. behind any statistical model there are assumptions which are more or less reasonable for our data at hand; in the **specification step** we define the feature/assumptions we are

2. then there will be the **estimation step**: models are characterized by unknown quantities that have to be estimated; depending on the amount of assumptions we are willing to make we can use different estimating procedures (we will focus on ML methods, btw)

3. then some task tipically involved are **hypothesis testing on regression coefficient** and **model comparison** (to choose among candidate models)

1.1.0.1 Random samples

We are interested in a phenomenon `Y` (eg cholesterol level) but for practical reasons we cannot know the distribution of whole the population `P`; so we rely on a *observed sample* on n units $\mathbf{y}$ which is realization of the random mechanism called *random sample* $\mathbf{Y}$ (a collection of random variables). The observed sample is an element of the set of all possible samples $\boldsymbol{\mathcal{Y}}$ we can draw, called *sample space*.
Below some notation

`Y`	statistical phenomenon of interest in a given population `P`
$\mathbf{y} = (y_1, \ldots, y_n)^\top$	Observed sample *(numerical) values observed on n statistical units randomly drawn from the population* `P`
$\Downarrow$	
$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$	Random sample: *set of r.vs. that describe the possible value of* `Y` *in each random draw* $\Rightarrow$ $\mathbf{y} \in \boldsymbol{\mathcal{Y}} \subseteq \mathbb{R}^n$ *Sample space* $\Rightarrow$ $f_0(\mathbf{y})$ *Unknown "true" probability mass/density function of* $\boldsymbol{Y}$

1.1.0.2 Parametric statistical models

We want to have information about f_0 using our sample, we have two strategies:

- to introduce a parametric statistical model: we assume that f_0 is element of a broader set $\mathcal{F}$ of probability distribution having the same functional form and which differs by a set of k parameters $\boldsymbol{\theta}$ (which can be a scalar as well); the distribution of our interest is f with $\boldsymbol{\theta}_0$ unknown. So the problem is rephrased from extract information on f_0 to information of $\boldsymbol{\theta}_0$
- adopt a non parametric approach (not our focus here)

So regarding the parametric statistical model

$f_0 \in \mathcal{F} = \left\{ f(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\}$	parametric statistical model for $\boldsymbol{Y}_n$ *parametric family containing the "true" probability mass/density function of* $\boldsymbol{Y}$ $\Rightarrow$ $\boldsymbol{\Theta}$ *parameter space* $\Rightarrow$ $f_0(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta}_0)$ with $\boldsymbol{\theta}_0$ *unknown*

1.1.0.3 Parametric statistical model specification

Model specification is the process of choosing (*"specifying"*) a parametric statistical model $\mathcal{F}$ suitable for $\boldsymbol{Y}$.
Specifying it means introducing a set of assumptions that describe the statistical model; this is a crucial step since inferential procedures rely on the model to be correctly specified. There are tools to check wheter the model assumptions are adequate or not.
Model specification can be based on information about:

- the features of the statistical phenomenon `Y` of interest and of the population `P` (eg qualitative/quantitative, discrete/continuous, bounded or not). This course will be focused on this task.
- the sampling scheme: this will define the *dependence structure among the rvs in the random sample* $\boldsymbol{Y}$, eg sampling schemes with dependence vs independence among observations.
  We will mainly focus on independent observations.

Once specified the model we can make inference on parameters; errors in model specification will give error in inference.

1.1.0.4 Likelihood function of θ

Supposing we

- have chosen a parametric statistical model/family of distribution *for the random sample* $\boldsymbol{Y}$ $\mathcal{F} = \left\{ f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k \right\}$
- have our observed sample - realisation of the random sample $\boldsymbol{Y}$, $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)^\top$

The likelihood function is a way to combine these informations; $L(\boldsymbol{\theta})$ is *Likelihood function of* θ that is a function which treat observed data as fixed $L(\cdot) = L(\cdot; \mathbf{y})$ and of the type $L : \boldsymbol{\Theta} \to \mathbb{R}^+ \cup 0$, so going from the parameter space to the positive reals. Likelihood do depends on sample values so different sample will be characterized by different likelihoods.
Actually we have that for each possible $\boldsymbol{\theta}$ the likelihood function is $c(\mathbf{y}) f(\mathbf{y}; \boldsymbol{\theta})$, where $c(\mathbf{y})$ represents a multiplicative factor that does not depend on $\boldsymbol{\theta}$; so the likelihood function is proportional to the density/mass function evaluated on the observed sample.
The likelihood function:

- summarize all the information we have about f_0 (the "true" probability distribution of $\boldsymbol{Y}$):
  - on one hand the $f_0 \in \mathcal{F}$ parametric statistical model; the pre-experimental (a priori - before observing the actually drawn sample) information - theoretical assumptions
  - on the other hand the data/empirical evidence $\mathbf{y} = (y_1,\ y_2,\ \ldots,\ y_n)^\top$ in the observed sample
- literally shows how the probability/density of observing the actually drawn sample changes, as the value of the unknown parameter $\boldsymbol{\theta}$ changes
- from the practical pov, it can be interpreted asa way to measure the plausibility of each possible value of $\boldsymbol{\theta}$

Example 1.1.1. Assuming each Y_i has a gaussian distribution with common mean and variance and observation are independent, that is

$$Y_i \sim N\big(\mu, \sigma^2\big), \quad \text{IID} \quad i = 1, \ldots n, \quad \mu \in \mathbb{R}, \quad \sigma^2 \in \mathbb{R}^+$$

we have

- given that each random variable can take any value on the real line, the sample space is $\mathbb{R}^n$ ($\mathbf{y} \in \boldsymbol{\mathcal{Y}} = \mathbb{R}^n$)
- the parameter space is $R \times \mathbb{R}^+$, the first for mean the second for variance ($\boldsymbol{\theta} = \big(\mu, \sigma^2\big)^\top \in \boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$)
- the likelihood is the product (being observation independent) of gaussian density functions with data y_i replaced instead of x

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

Example 1.1.2. Using a more compact notation we can re-express/summarize the setup/distribution of the random vector $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ by introducing the multivariate normal, with

$$\boldsymbol{Y} \sim MVN_n \left(\underbrace{\boldsymbol{\mu} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}}_{n \times 1}, \underbrace{\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n}_{n \times n} \right)$$

Remark 1. Multivariate gaussian are fundamental in inference so a review follows

1.2 Multivariate Gaussian distributions review

1.2.0.1 Joint probability density function

Multivariate gaussian distribution is used when we have a random vector Y which can take values in $\mathbb{R}^n$ and being a vector we will have a vector of means (it contains all the expected values of the elements in the random sample, μ_1 will be the expected value of Y_1) and a variance covariance matrix which will contains variances of the random variables contained in the random vector as long as the relationship/covariances between each pair of them. $\boldsymbol{\mu}$ can be any real valued vector, $\boldsymbol{\Sigma}$ must be square symmetric and positive definite (variances must be striclty positive and covariances are bounded between the product of the square root of the variances

$\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$	n-dimensional random variable
$\mathbf{y} = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$	set of possible values of $\boldsymbol{Y}$ (joint realisations of the n random variables)
$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^\top$	n-dimensional real-valued vector
$\boldsymbol{\Sigma}$	$n \times n$ real-valued, symmetric matrix (positive definite - invertible)

From a functional pov the joint density expression is as follows. We have an expression involving trhe inverse of varcov matrix, its determinant, the difference between vectors y and mu and lots of quantity that relies on matrix algebra

$$\boldsymbol{Y} \sim MVN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff f(y_1, \dots, y_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right]}{(2\pi)^{\frac{n}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}}$$

The point is that this function assign each vector Y a non negative real value (its density).
With random vector we can apply expected value and variance operators obtaining respectively the vector containing the expected value of each element in the vector, while when applying the vcariance operator one gets the matrix

containing variances (main diagonal) and covariances (off diagonal)

$$\begin{aligned}\mathrm{E}[\boldsymbol{Y}] &= \boldsymbol{\mu} \\ \mathrm{Var}[\boldsymbol{Y}] &= \mathrm{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \mathrm{E}[\boldsymbol{Y}]\,\mathrm{E}[\boldsymbol{Y}]^\top = \boldsymbol{\Sigma}\end{aligned}$$

1.2.0.2 Standardised multivariate Gaussian distribution

As long as univariate case we have the special case of standard gaussian random vector which is obtained choosing null vector as means and identity matrix as variance-covariance

$\boldsymbol{\mu} = \mathbf{0}_n = (0, 0, \ldots, 0)^\top$ n-dimensional null vector
$\boldsymbol{\Sigma} = \mathbf{I}_n$ $n \times n$ identity matrix

In this case the functional form of the density become simpler because it's the only case in which uncorrelation implies independence and so we can write the joint density as product of marginal ones

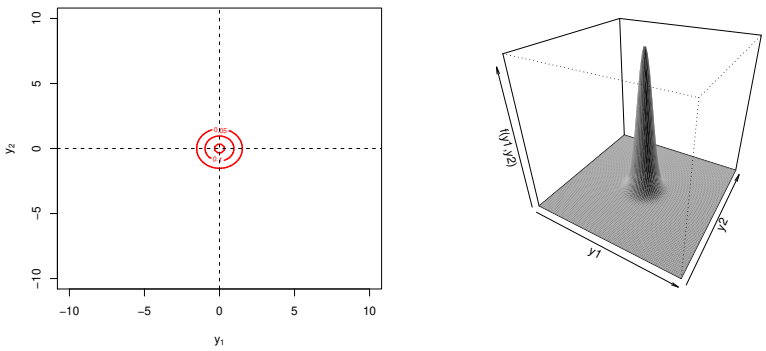
$$\begin{aligned}f(y_1, \ldots, y_n; \mathbf{0}_n, \mathbf{I}_n) &= \frac{\exp\left[-\frac{1}{2}\sum_{i=1}^n y_i^2\right]}{(2\pi)^{\frac{n}{2}}} \\ &= \prod_{i=1}^n \frac{\exp\left[-\frac{y_i^2}{2}\right]}{\sqrt{2\pi}}\end{aligned}$$

Example 1.2.1 (Examples with $n = 2$)**.** In two dimension we can plot the density; if

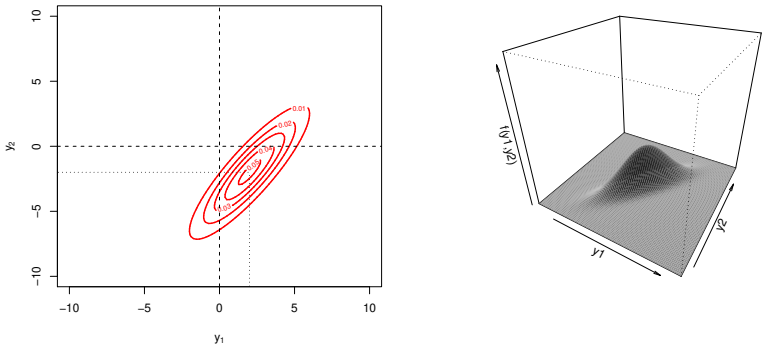
- $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ distribution plotted in figure 1.1 a and b. with identity varcov matrix shape is a circle
- $\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$ $\boldsymbol{\Sigma} = \begin{bmatrix} 4.8 & 5.4 \\ 5.4 & 7.95 \end{bmatrix}$ distribution plotted in figure 1.1 c and d. By introducing correlation we move toward elliptical distribution. The orientation of the ellipses depend on the correlation, here positive; the stretch of the ellypse depends on the difference between variances

1.2.0.3 Some properties

- each marginal distribution of order $q < n$ (eg any subvector) is a q-dimensional multivariate Gaussian distribution: so if a vector is multivariate gaussian, even the single Y_i composing are as well;
- each conditional distribution of a subvector of Y, $Y_{1a}, Y_{2a}, \ldots, Y_{ha}$, given another portion of the subvector $Y_{1b}, Y_{2b}, \ldots, Y_{lb}$, then this is an h-dimesional multivariate Gaussian distribution as well. So we can say that mvn we can say is closed both to marginalization (first point) and to conditioning: whenever we extract a marginal or a conditional distribution from a gaussian rv, we obtaina gaussian as well



(a) (b)



(c) (d)

Figure 1.1: example1.

- $Y_1, \dots, Y_n$ are independent *if and only if* $\mathbf{\Sigma}$ is diagonal (if and only if they are uncorrelated) and in the case the joint distribution is the product of the n univariate gaussian distribution

$$f(\mathbf{y}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma_i^2}\right]}{(2\pi)^{\frac{n}{2}}\left[\prod_{i=1}^{n}\sigma_i^2\right]^{\frac{1}{2}}} = \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{1}{2\sigma_i^2}(y_i-\mu_i)^2\right]\right\}$$

- linear combinations ov MVN random variables (important property): let
  - $\boldsymbol{Y}$ be a n-dimensional Gaussian vector characterized by parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$
  - $\mathbf{A}$ a $q \times n$ real-valued matrix (fixed not random)
  - $\mathbf{b}$ an n-dimensional real-valued vector,

  then the linear combination using $\mathbf{A}$ and $\mathbf{b}$, $\mathbf{Z} = \mathbf{A}(\boldsymbol{Y} + \mathbf{b})$ (so we add a constant to the vector Y and premultiply by A), is a q-dimensional Gaussian vector with parameters trasformed in the following way: mean $\mathbf{A}(\boldsymbol{\mu} + \mathbf{b})$ and varcov $\mathbf{A\Sigma A}^\top$.
  This reminds the univariate case where if $Z = a(Y + b)$ then $\mathbb{E}[Z] = a(\mathbb{E}[Y] + b)$ and $\operatorname{Var}[Z] = a^2 \operatorname{Var}[Y]$

Important remark 1. Nella notazione del prof quando una lettera maiuscola è ingrassetto, se ha anche italic è un vettore ($\boldsymbol{Y}$), altrimenti solo grassetto per matrici $\mathbf{A}$

Example 1.2.2 (Standardization as linear combination)**.** We can use the last property to standardize any gaussian random vector. If $\mathbf{\Sigma}$ is positive definite, there's a way to obtain $\mathbf{\Sigma}^{\frac{1}{2}}$, inverse of which can be thought as square root since $\mathbf{\Sigma} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}$. This matrix will be invertible with inverse $\mathbf{\Sigma}^{\frac{1}{2}}$. if $\mathbf{\Sigma}$ is invertible. So in this case, by setting the $\mathbf{A}$ and $\mathbf{b}$ as follows:

NB: we dont' delve in how to obtain the square root here

- $\mathbf{A} = \mathbf{\Sigma}^{-\frac{1}{2}}$ such that $\mathbf{\Sigma} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}$ and $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{I}_n$
- $\mathbf{b} = -\boldsymbol{\mu}$

then $\boldsymbol{Z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu})$ will be an n-dimensional *standardised* Gaussian vector. Again this is similar to what occurs in the univariate where to standardize a variable we have to subtract the mean and divide by the standard deviation

Chapter 2

Gaussian linear model

2.1 An introductory example

2.1.1 Simple linear regression

2.1.1.1 Setup

It is known a glucose level in blood between 100 and 125 mg/dL represents a risk factor for developing diabetes:

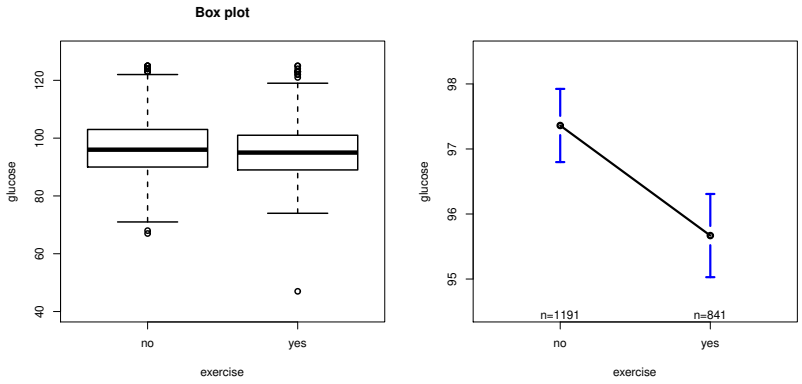
- the **Aim:** Does physical activity (a modifiable factor related to life style) contribute to the reduction of the glucose level, thus preventing a severe disease?
- **available information:** data from an observational study for glucose level and physical activity (yes-no) on a sample of 2032 women not affected by diabetes after menopause

The aim is to look at if there are difference in glucose level between active and non active women; we look at conditional distributions (via boxplot and group means) (fig 2.1):

- he two boxplot are quite similar in terms of variability; a little shift in location of the boxplot (yes slightly lower median) but there's a lot of overlapping.
- by zooming to the means with confidence interval there seems to be a difference here in mean glucose level (rather small 1.5-1.7) and there's no overlap between the 2 ci; probablu the diff between two sample means is statistically significant

2.1.1.2 Models specification

We then state a more formal description of the relation between glucose level and physical activity using a parametric statistical model with parameters related to difference between groups (our main interest); once formalized we will be able to perform ML estimation and hypothesis test using tools exploiting maximum likelihood



(a) (b)

Figure 2.1: glucose and exercise

- the starting point is the joint distribution, which can be splitted in the product of marginal (of exercise) times the conditional of glucose level given exercise which is our main interest

  $$f(\texttt{glucose}_i, \texttt{exercise}_i) = f(\texttt{glucose}_i|\texttt{exercise}_i) f(\texttt{exercise}_i) \quad i = 1, \ldots, 2032$$

  however rather than focusing on joint distribution we focus our attention on conditional distribution

- regarding the conditional distribution we make the following assumptions (which are somewhat reasonable looking at the graphs before):

  A) conditional expected values, that is the expected values of the conditional distribution are summarized by

  $$\mathrm{E}\left[\texttt{glucose}_i|\texttt{exercise}_i\right] = \beta_0 + \beta_1 \mathbf{1}\left\{\texttt{exercise}_i = \texttt{yes}\right\}, \forall i$$

  where

  $$\mathbf{1}\left\{\texttt{exercise}_i = \texttt{yes}\right\} = \begin{cases} 1 & \text{if } \texttt{exercise}_i = \texttt{yes} \\ 0 & \text{otherwise} \end{cases}$$

  B) the conditional variances are supposed to be constant, the two conditional distribution have the same variability (it's independent from the regressors)

  $$\mathrm{Var}\left[\texttt{glucose}_i|\texttt{exercise}_i\right] = \sigma^2, \forall i$$

  C) regarding dependence between observation a reasonable assumption is to assume that the conditional distribvution of two units in the sample are uncorrelated

  $$\mathrm{Corr}\left(\texttt{glucose}_i|\texttt{exercise}_i, \texttt{glucose}_j|\texttt{exercise}_j\right) = 0, \forall i \neq j$$

D) finally having seen quite symmetry of conditional distributions, we assume that conditional distribution are gaussian (with mean and variance as stated before)

$$\texttt{glucose}_i|\texttt{exercise}_i \sim \mathrm{N}\left(\beta_0 + \beta_1 \mathbf{1}\left\{\texttt{exercise}_i = \texttt{yes}\right\}, \sigma^2\right), \forall i$$

2.1.1.3 Estimation

```
> summary(modello1)
Call:
lm(formula = glucose ~ exercise, data = hers.nod)

Residuals:
     Min       1Q   Median       3Q      Max
 -48.668   -6.668   -0.668    5.639   29.332
 Coefficients:
                Estimate  Std.  Error  t value  Pr(>|t|)
   (Intercept)   97.3610        0.2815   345.85    0.0000
   exerciseyes   -1.6928        0.4376    -3.87    0.0001

Residual standard error:  9.715 on 2030 degrees of freedom
Multiple R-squared:  0.007318, Adjusted R-squared:  0.006829
F-statistic:  14.97 on 1 and 2030 DF, p-value:  0.000113
```

Here the test of interest regards exerciseyes; there is a significant difference (which does not need to be clinically relevant) while `Residual standard error` is an estimate of $\sqrt{\sigma^2}$: there is a lot of variability even within the same conditional distribution and this reflect the fact that R square is low.

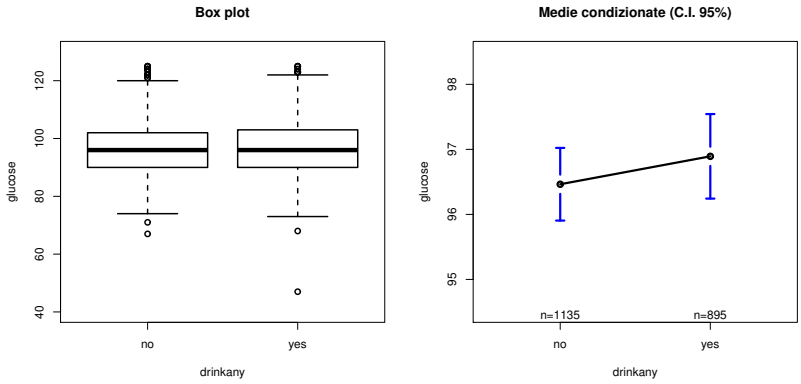
2.1.2 Multiple linear regression

2.1.2.1 Introducing other regressors

Women that are physically active may completely differ from women that are not, due to a number of other characteristics (socio-economical status, life style, health conditions Some of these characteristics could be associated with both the glucose level and physical activity (eg women that are physically active could be younger, haltier and have different habits related to alcohol consumption).
Since data were collected through an observational study, these characteristics could act as confounders, thus preventing a correct evaluation of the effect of physical activity on glucose level.
Some plotting regarding drinking age and bmi is done in figures 2.2 (not great differences graphically), 2.3 (slight tendency of decreasing trend), 2.4 (increasing).

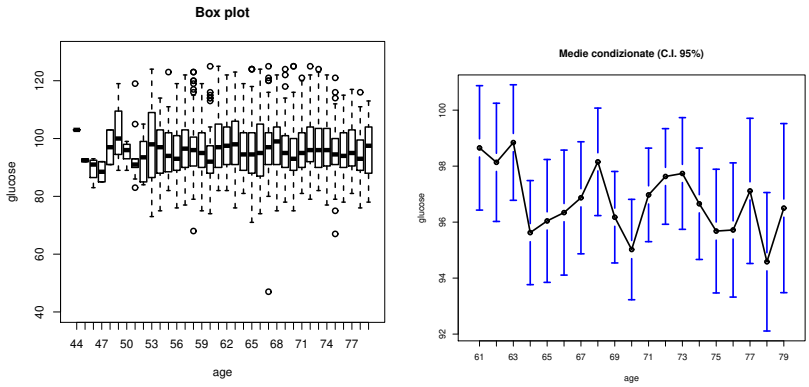
2.1.2.2 Model definition/specification

As done before to put together all



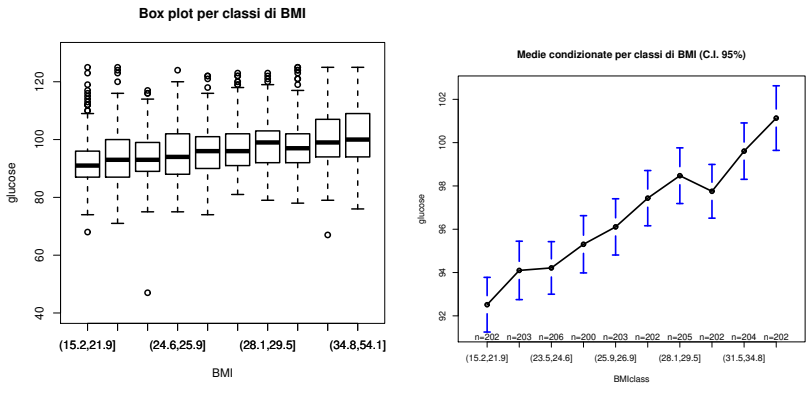
(a) (b)

Figure 2.2: glucose and drink



(a) (b)

Figure 2.3: glucose and age



(a) (b)

Figure 2.4: glucose and bmi

- the joint density is as follow

$$
\begin{aligned}
&f(\texttt{glucose}_i, \texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i) \\
&= f(\texttt{glucose}_i | \texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i) \cdot f(\texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i) \\
&= f(y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}) f(x_{1i}, x_{2i}, x_{3i}, x_{4i}), i = 1, \ldots, 2032
\end{aligned}
$$

  but our main focus given are the conditional distribution

- to focus on the conditional distribution assumptions

  A) the main extension relative to the univariate case is on the first assumption

$$
\begin{aligned}
&\mathrm{E}\left[Y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}\right] \\
&= \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} \\
&= \beta_0 + \beta_1 \mathbf{1}\left\{\texttt{exercise}_i = \texttt{yes}\right\} + \beta_2 \mathbf{1}\left\{\texttt{drinkany}_i = \texttt{yes}\right\} + \beta_3 \texttt{age}_i + \beta_4 \texttt{BMI}_i, \forall i
\end{aligned}
$$

  B) we assume constant conditional variance

$$\mathrm{Var}\left[Y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}\right] = \sigma^2, \forall i$$

  C) we keep the uncorrelation

$$\mathrm{Corr}\left(Y_i | x_{1i}, x_{2i}, x_{3i}, x_{4i}, Y_h | x_{1h}, x_{2h}, x_{3h}, x_{4h}\right) = 0, \forall i \neq h$$

  D) again the normality assumption

$$\texttt{glucose}_i | \texttt{exercise}_i, \texttt{drinkany}_i, \texttt{age}_i, \texttt{BMI}_i \sim N\left(\mu_i, \sigma^2\right), \forall i$$

2.1.2.3 Estimation

```
> summary(modello2)
Call:
lm(formula = glucose ~ exercise + drinkany + age + BMI, data = hers.nod)
...
 Coefficients:
               Estimate  Std.  Error  t value  Pr(>|t|)
   (Intercept)   78.9624       2.5928    30.45    0.0000
   exerciseyes   -0.9504       0.4287    -2.22    0.0267
   drinkanyyes    0.6803       0.4220     1.61    0.1071
           age    0.0635       0.0314     2.02    0.0431
           BMI    0.4892       0.0416    11.77    0.0000
```

Effect of physical activity changes (it was -1.69) so part of the effect was due to covariates, but is still significant. BMI is another important factor

2.2 General definition

2.2.0.1 Basic assumptions

In general considering

- Y_i: random variable that describes the value for the dependent variable observed on the i-th sample unit ($i = 1, \ldots, n$);
- $x_{1i}, x_{2i}, \ldots, x_{pi}$ values of the regressors for the i-th sample unit (*covariate pattern*), where p is the number of regressors observed on all the units.

A gaussian parametric model relates Y_i and $x_{1i}, x_{2i}, \ldots, x_{pi}$ assuming:

A) the conditional expected value will be assumed to be a linear combination (*linearity assumption* of the expected value)

$$\mathrm{E}\left[Y_i | x_{1i}, \ldots, x_{pi}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \forall i$$

B) the conditional variance will be assumed constant (*homoskedasticity assumption*)

$$\mathrm{Var}\left[Y_i | x_{1i}, \ldots, x_{pi}\right] = \sigma^2, \forall i$$

C) the *incorrelation assumption*

$$\mathrm{Cor}\left[Y_i | x_{1i}, \ldots, x_{pi}, Y_h | x_{1h}, \ldots, x_{ph}\right] = 0, \forall i \neq h$$

D) the name of the model comes from the last assumption, the *gaussianity assumption* regarding conditional distributions

$$Y_i | x_{1i}, \ldots, x_{pi} \sim N\big(\mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \sigma^2\big), \forall i$$

Important remark 2 (Linearity). Linearity in this contest mean two different things:

- the conditional expectedvalue is linear *in the regressors* since its a linear combination of regressors given a set of regression coefficient
- it is also linear *in the parameter*: it's a linear combination in the parameters given a certain value for the regressors

It's important to keep in mind the duality of this concept: at some point we'll make gaussian model more flexible by removing one of this two linearity. We'll define model still linear in the parameters but nonlinear in the regressors.

2.2.0.2 Parameter space and sample space

Definition 2.2.1 (Parameter space)**.** Model parameters to be estimated:

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{(p+1)}$: $(p+1)$-dimensional real-valued vector
- $\sigma^2 \in \mathbb{R}^+$: positive scalar value

So overall the parameter to be estimated can be collected in a vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma^2\right)^\top \in \boldsymbol{\Theta} = \mathbb{R}^{(p+1)} \times \mathbb{R}^+$ where $\boldsymbol{\Theta}$ is the parameter space that is the set of possible values and the set of parameter to be estimated will be $p + 2$

Definition 2.2.2 (Conditional sample space)**.** Its the set of possible observation and is given by

$$\mathbb{R} \times \left\{ (x_{1i}, \ldots, x_{pi})^\top, i = 1, \ldots, n \right\}$$

where the first $\mathbb{R}$ is due to the dependent variable, which can take any real value, and $\left\{ (x_{1i}, \ldots, x_{pi})^\top, i = 1, \ldots, n \right\}$ is a discrete sets of points of observed data (covariate patterns), which is treated as they were constants/not random (we're ignoring the distribution of the regressors): in other words here we're conditioning on the observed values of the regressors

Example 2.2.1. In the simple case where we have only a dummy variable it is $(\mathbb{R} \times \{0\}) \cup (\mathbb{R} \times \{1\})$

2.2.0.3 Probability density function (1)

Given the assumption provided before we have that the joint density for all the r.vs. $Y_1, \ldots, Y_n$ conditional to the regressor values is

$$\begin{aligned} f(y_1, \ldots, y_n | x_{11}, \ldots, x_{p1}, x_{1n}, \ldots, x_{pn}) &\overset{(1)}{=} \prod_{i=1}^{n} f(y_i | x_{1i}, \ldots, x_{pi}) \\ &\overset{(2)}{=} \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2 \right] \right\} \\ &= \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \right\}^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2 \right] \end{aligned}$$

where

- (1) the joint conditional distribution is the product of univariate conditional distribution due to independence assumption (uncorrelation + normality = independence)
- (2) due to gaussian distribution: here we substitute with normal density and exploiting the first assumption regarding expected value and the homoskedasticity one

Now we see this latter is just a multivariate gaussian density function where we have a diagonal variance/covariance matrix: any marginal distribution of a multivariate gaussian is still gaussian as we have by assuming incorrelation we are implicitly saying we have independence

2.2.0.4 Matrix representation

It's useful to formalize our model using matrix representation; if:

- $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ is n-dimensional random variable that describes the values for the dependent variable jointly observed on n sample units
- $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ are the observed sample values;

- $\mathbf{x}_i = (x_{0i}, x_{1i} \ldots, x_{pi})^\top$ contains the value of the regressors for the l'i-th sample unit plus an additional element, which is $x_{0i} = 1, \forall i$, constant/"fake" regressor associated with the intercept.
  Therefore $\mathbf{x}_i$ will have $p+1$ elements, where p is the number of regressors;

- on the other hand we define $\mathbf{x}_{[j]} = (x_{j1}, x_{j2} \ldots, x_{jn})^\top$ as the value for a single regressor ($j = 0, \ldots, p$) observed values for all the units (eg for the intercept $x_{[0]} = (1, 1, \ldots, 1)^\top$)

Thus the regressor matrix (matrix containing all the values of the regressors observed on all the units) is an $n \times (p+1)$ matrix and can be seen alternatively as column vector of rows/units or as row vector of columns/regressors:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{p1} \\ 1 & x_{12} & \ldots & x_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \ldots & x_{pn} \end{bmatrix} = \left[\begin{array}{c} \mathbf{x}_1^\top \\ \hline \mathbf{x}_2^\top \\ \hline \vdots \\ \hline \mathbf{x}_n^\top \end{array}\right] = \left[\mathbf{x}_{[0]} \middle| \mathbf{x}_{[1]} \middle| \cdots \mathbf{x}_{[p]}\middle|\right]$$

Once we've defined this stuff representation we come up with compact notation for all the remaining stuff we've introduced before.

- first, regarding the **conditional expected value** for a single unit, using the matrix notation, its done by vector multiplication of its covariate times the betas

  $$\mathbb{E}\left[Y_i | x_{1i}, \ldots, x_{pi}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad \forall i$$

  This for a single unitrandom variable, but we can express the vector of conditional expected value for all the sample as

  $$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{x}_1^\top \boldsymbol{\beta} \\ \mathbf{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ \mathbf{x}_n^\top \boldsymbol{\beta} \end{bmatrix} = \mathbb{E}\left[\mathbf{Y}\right]$$

- for what concerns the **probability density function** we can express the equation found before more compactly as

  $$\begin{aligned} f\left(\mathbf{y} | \mathbf{X}; \boldsymbol{\beta}, \sigma^2\right) &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2\right] \\ &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \end{aligned}$$

  where we have rewritten the sum of squares in square brackets as dot product of the vector containing the elements of the sum $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ times itself (we have the sum of squares of differences between observed values $\mathbf{y}$ and expected values $\mathbf{X}\boldsymbol{\beta}$)

- according to assumptions A) to E), thus $\mathbf{Y}$, given the regressor values is distributed as multivariate Gaussian being composed by single gaussian, with expected value as derived before, and diagonal varcov matrix (common variance and no covariance between variables)

$$\mathbf{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$$

  This is a more compact notation we can use when referring to a Gaussian linear regression model

2.2.0.5 An alternative definition

An equivalent alternative definition for the family of gaussian model; the previous definition was focused on assumption of the conditional distribution of Y given the regressor. There's a completely equivalent way to express gaussian models which start from a different starting point. Considering:

- Y_i a random variable that describes the value for the dependent variable observed on the i-th sample unit ($i = 1, \ldots, n$)
- $x_{1i}, x_{2i}, \ldots, x_{pi}$ the values of the regressors for the i-th sample unit (*covariate pattern*)

rather than focusing on conditional distribution we start assuming that each random variable Y_i in the sample can be decomposed in the sum of two quantities: the deterministic component and the random error.

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}}_{\text{deterministic component}} + \underbrace{\varepsilon_i}_{\text{random error}}$$

This formulation with random error latter emcompass the fact that there's no deterministic relation between X and Y or, in other terms, there's something that cannot be explained in Y_i using only X (so unit with the same X are allowed to have different Y_i).
In this setting the assumptions are focused on the error and especially on its conditional distribution given the regressors:

A) the conditional expected value is null for all the units (some will be positive, some negative but on average it cancels out):

$$\mathrm{E}\left[\varepsilon_i | x_{1i}, \ldots, x_{pi}\right] = 0, \forall i$$

B) the conditional variance is constant/independent

$$\mathrm{Var}\left[\varepsilon_i | x_{1i}, \ldots, x_{pi}\right] = \sigma^2, \forall i$$

C) there's no conditional correlation between error of different unit

$$\mathrm{Cor}\left[\varepsilon_i | x_{1i}, \ldots, x_{pi}, \varepsilon_h | x_{1h}, \ldots, x_{ph}\right] = 0, \forall i \neq h$$

D) the conditional distribution of the random error is gaussian

$$\varepsilon_i | x_{1i}, \ldots, x_{pi} \sim N\left(0, \sigma^2\right), \forall i$$

Putting all things together considering the sample we have that:

- the vector $\boldsymbol{\varepsilon}$ containing all the unit error $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ ...
- has a conditional distribution which is a multivariate gaussian distribution with an expected values vector of 0 and diagonal variance covariance matrix with constant diagonal: $\boldsymbol{\varepsilon}|\mathbf{X} \sim MVN_n\big(\mathbf{0}_n, \sigma^2\mathbf{I}_n\big)$
- now since $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ given the fact that $\mathbf{Y}$ is a linear transformation of $\boldsymbol{\varepsilon}$ ($\mathbf{Y} = \mathbf{A}(\boldsymbol{\varepsilon}+\mathbf{b})$ with $\mathbf{A} = \mathbf{I}_n$ and $\mathbf{b} = \mathbf{X}\boldsymbol{\beta}$), thanks to the properties of multivariate Gaussian distributions, we have that the conditional distribution of $\mathbf{Y}$ is multivariate gaussian as well with the following distribution

$$\mathbf{Y}|\mathbf{X} \sim MVN_n\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\big)$$

2.3 Maximum likelihood estimation

2.3.1 Likelihood and related quantities

2.3.1.1 Likelihood function

The unknown parameters in a Gaussian linear regression models are

- $\boldsymbol{\beta}$ regression coefficients (including the intercept)
- σ^2 conditional variance

The betas are of major interest in the estimation process while σ^2 is a parameter which is estimated and informative but typically of less interest.
Given the regressor values in matrix $\mathbf{X}$ and the observed values for the dependent variable on the sample units in vector $\mathbf{y}$, the likelihood function is obtained using the joint conditional density function which is the product of the single conditional density functions for each unit

$$\begin{aligned} L\big(\boldsymbol{\beta}, \sigma^2\big) &= L\big(\boldsymbol{\beta}, \sigma^2; \mathbf{y}\,|\mathbf{X}\big) \\ &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2\right] \\ &= \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\right\}^n \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \end{aligned}$$

We look at this function considering $\mathbf{y}$ and $\mathbf{X}$ fixed and we want to maximize it by choosing the unknown parameters.
Several function can be obtained starting from the likelihood function, developed in what follows.

2.3.1.2 Log-likelihood function

There are practical (with the log we have sum and dealing with maximization of sum is easier than dealing with maximization of product) and technical/theorical

reasons for using the log likelihood, which is:

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2) &= \ln L(\boldsymbol{\beta}, \sigma^2) \\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2 \\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Note that the first term $-\frac{n}{2}\ln 2\pi$ is an additive constant independent from the unknown parameters and it can be ignored in the maximization.

2.3.1.3 Score function for β

Development Starting from the likelihood function there are other functions that can be derived and are needed in the maximization:

- the *score function* $U(\boldsymbol{\beta}) = \dfrac{\partial}{\partial \boldsymbol{\beta}} \ln L(\boldsymbol{\beta}, \sigma^2)$ is the gradient of $l(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ (it's a *vector with* $p+1$ *elements*); in other words ...

- each of its element $U_j(\boldsymbol{\beta}) = \dfrac{\partial}{\partial \beta_j} \ln L(\boldsymbol{\beta}, \sigma^2)$, $j = 0, \ldots, p$ is the first partial derivative of $l(\boldsymbol{\beta}, \sigma^2)$ with respect to β_j $(j = 0, \ldots, p)$

$$\begin{aligned} U_j(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_j}\left\{-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2\right\} \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) \cdot 2 \cdot (-1) \cdot x_{ji} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) x_{ji} \end{aligned}$$

  so it ends multiplying the residual time the x of the considered beta over σ^2

Remark 2. per l'esame dice di non chiedere la derivazione ma ci potrebbero essere domande riguardo l'espressione finale

Matrix representation for $U(\beta)$ Exploiting the dot product we can express the score function in matrix form:

- the single element of the vector will be

$$U_j(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) x_{ji}}{\sigma^2} = \frac{\mathbf{x}_{[j]}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}$$

- the full vector can be expressed as

$$U(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\mathbf{x}_{[0]}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \frac{\mathbf{x}_{[1]}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ \vdots \\ \frac{\mathbf{x}_{[p]}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \end{bmatrix} = \frac{\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}$$

  So calculating the score function is easy for a computer by applying this last equation

An alternative derivation of $U(\beta)$ An equivalent way to express the score function exploits the differentiation rules for functions with vector arguments (we get the same results we can obtain it in the last way if we dont know these tools)

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2) = \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \right\}$$
$$-\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \mathbf{y}^\top\mathbf{y} - \underbrace{\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}}_{2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \right\}$$

Then recalling that $\frac{\partial}{\partial \boldsymbol{\delta}}\boldsymbol{\delta}^\top\mathbf{A} = \mathbf{A}$ and $\frac{\partial}{\partial \boldsymbol{\delta}}\boldsymbol{\delta}^\top\mathbf{A}\boldsymbol{\delta} = 2\mathbf{A}\boldsymbol{\delta}$

$$U(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\left\{\mathbf{0}_{p+1} - 2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}\right\} = \frac{\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma^2}$$

Remark 3. non ci ha speso molto, forse tornerà utile in futuro

2.3.1.4 Observed Fisher information for β

Derivation We have that

- the observed Fisher information is the negative of the Hessian matrix of the loglike function $l(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$, so it's $(p+1) \times (p+1)$ *matrix*:

$$i(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} \ln L(\boldsymbol{\beta}, \sigma^2)$$

- its generic element $i_{jl}(\boldsymbol{\beta})$ is the second partial derivative of $l(\boldsymbol{\beta}, \sigma^2)$ with respect to β_j and β_l $(j, l = 0, \ldots, p)$

$$i_{jl}(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \beta_j \partial \beta_l} \ln L(\boldsymbol{\beta}, \sigma^2)$$

son on a practical pov we have

$$\begin{aligned} i_{jl}(\boldsymbol{\beta}) &= -\frac{\partial}{\partial \beta_l} U_j(\boldsymbol{\beta}) \\ &= -\frac{\partial}{\partial \beta_l} \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi}) x_{ji} \\ &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ji} x_{li} \cdot (-1) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ji} x_{li} \end{aligned}$$

we end up with a rather simple expression involving the sum of cross product of the j-th and the l-th regressors

Remark 4. Called information because we're measuring the curvature of the loglikelihood function so the more loglikelihood function is curved the more information we have regarding our estimate to be the best

Matrix representation Again exploiting the dot product we can have a compact representation

- starting from the single element we have that

$$i_{jl}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} x_{ji} x_{li}}{\sigma^2} = \frac{\mathbf{x}_{[j]}^\top \mathbf{x}_{[l]}}{\sigma^2}$$

- then for the full matrix

$$i(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\mathbf{x}_{[0]}^\top \mathbf{x}_{[0]}}{\sigma^2} & \frac{\mathbf{x}_{[0]}^\top \mathbf{x}_{[1]}}{\sigma^2} & \ldots & \frac{\mathbf{x}_{[0]}^\top \mathbf{x}_{[p]}}{\sigma^2} \\ \frac{\mathbf{x}_{[1]}^\top \mathbf{x}_{[0]}}{\sigma^2} & \frac{\mathbf{x}_{[1]}^\top \mathbf{x}_{[1]}}{\sigma^2} & \ldots & \frac{\mathbf{x}_{[1]}^\top \mathbf{x}_{[p]}}{\sigma^2} \\ \vdots & & & \\ \frac{\mathbf{x}_{[p]}^\top \mathbf{x}_{[0]}}{\sigma^2} & \frac{\mathbf{x}_{[p]}^\top \mathbf{x}_{[1]}}{\sigma^2} & \ldots & \frac{\mathbf{x}_{[p]}^\top \mathbf{x}_{[p]}}{\sigma^2} \end{bmatrix} = \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2}$$

An alternative derivation Again, exploiting the differentiation rules for functions with vector arguments, we end with exactly the same results

$$i(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l(\boldsymbol{\beta}, \sigma^2) = -\frac{\partial}{\partial \boldsymbol{\beta}^\top} U(\boldsymbol{\beta}) = -\frac{1}{\sigma^2} \frac{\partial}{\partial \boldsymbol{\beta}^\top} \left(\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \right)$$

and recalling that $\frac{\partial}{\partial \boldsymbol{\delta}^\top} \mathbf{A}\boldsymbol{\delta} = \mathbf{A}$

$$i(\boldsymbol{\beta}) = -\frac{1}{\sigma^2} \left[\mathbf{0}_{(p+1)\times(p+1)} - \mathbf{X}^\top \mathbf{X} \right] = \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2}$$

Remark 5. again non ci ha speso molto

2.3.1.5 Expected Fisher information for β

The observed information is a quantity that is specific to a given sample. Along with the observed there's the expected Fisher information as well: it's the expected value of the obserserved Fisher information across the possible samples

- it's a $(p+1) \times (p+1)$ matrix defined as

$$I(\boldsymbol{\beta}) = \mathrm{E}\left[i(\boldsymbol{\beta})\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} \ln L\big(\boldsymbol{\beta}, \sigma^2\big)\right]$$

- the generic element is the expected value of the generic element of $i(\boldsymbol{\beta})$, that is

$$I_{jl}(\boldsymbol{\beta}) = \mathrm{E}\left[i_{jl}(\boldsymbol{\beta})\right] = \mathrm{E}\left[\underbrace{\frac{\sum_{i=1}^n x_{ji}x_{li}}{\sigma^2}}_{\text{independent of } \mathbf{Y}}\right] = \frac{\sum_{i=1}^n x_{ji}x_{li}}{\sigma^2}$$

  and it turns out to depends only on the value of the j-th and l-th regressor. Note that the expected values are computed considering the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ (thus holding fixed the values of the regressors so in our sample the only random quantity is Y which does not figure under expected value which have just a constant)

- finally

$$\mathrm{E}\left[i(\boldsymbol{\beta})\right] = \frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2}$$

Important remark 3. So when dealing with gaussian linear regression models the observed and expected fisher information coincides: this sample in the sample space is as informative as any sample in the sample space with respect to the unknown parameters.
This is something that does not happen all the time; for other more complicated models this does not hold. We will appreciate the benefit of this equivalence when it comes to GLM

2.3.1.6 Properties of the score function

Differently from the fisher information matrixes, the score function $U(\boldsymbol{\beta}) = \frac{\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}$ is a function/quantity which depends on

- $\boldsymbol{\beta}, \sigma^2$ the *unknown* model parameters
- $\mathbf{X}$ the regressor values
- $\mathbf{y}$ the observed values of the dependent variable (*realisations of the r. v.* $\mathbf{Y}$)

So each sample will be characterized by a different log-likelihood function and score function: the first partial derivative can be different from sample to sample but the second partial derivatives (and information matrix) will be constant (if we condition on X)

We may think as the score function as a random variable itself, being a linear transformation of vector $\mathbf{Y}$; will have it's expected variable, varcov matrix etc. In gaussian linear regression models we can come up with the distribution of the score function:

- conditionally on the regressors values, $U(\boldsymbol{\beta})$ is the realisation of a random vector, that can be expressed as a linear transformation of $\mathbf{Y}$:

$$\mathbf{A} = \frac{\mathbf{X}^\top}{\sigma^2},\ \mathbf{b} = -\mathbf{X}\boldsymbol{\beta} \implies U(\boldsymbol{\beta}) = \mathbf{A}(\mathbf{Y} + \mathbf{b})$$

- assuming the Gaussian linear model assumptions, thanks to the properties of MVN gaussian, if we do the math we end with the fact that the conditional distribution of the score function given $\mathbf{X}$ is MVN as well (with $p+1$ elements) with 0 mean (it's independent of beta, whichever value they have) and variance covariance matrix coinciding with expected fisher information matrix

$$\mathbf{Y}|\mathbf{X} \sim MVN_n\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\big) \implies U(\boldsymbol{\beta})|\mathbf{X} \sim MVN_{p+1}\left(\underbrace{\frac{\mathbf{X}^\top}{\sigma^2}\left[\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}\right]}_{\mathbf{0}_{p+1}}, \underbrace{\frac{\mathbf{X}^\top}{\sigma^2}\sigma^2\mathbf{I}_n\frac{\mathbf{X}}{\sigma^2}}_{\frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2} = I(\boldsymbol{\beta})}\right)$$

the fact that score function has MVN distribution is very simple in gaussian models, but this results can be extended to other kind of models as well.

2.3.1.7 Standardising the score function

We can standardise the score function which make the MVN to have a varcov equal to the identity matrix.
In principle we define the square root of fisher expected information matrix that is $I(\boldsymbol{\beta})^{-\frac{1}{2}}$ such that:

$$I(\boldsymbol{\beta}) = I(\boldsymbol{\beta})^{\frac{1}{2}} I(\boldsymbol{\beta})^{\frac{1}{2}}$$
$$I(\boldsymbol{\beta})^{\frac{1}{2}} I(\boldsymbol{\beta})^{-\frac{1}{2}} = I(\boldsymbol{\beta})^{-\frac{1}{2}} I(\boldsymbol{\beta})^{\frac{1}{2}} = \mathbf{I}_{p+1}$$

We have that $I(\boldsymbol{\beta})^{-\frac{1}{2}} = \sigma\big(\mathbf{X}^\top\mathbf{X}\big)^{-\frac{1}{2}}$ exists if and only if the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, that is, if and only if $\mathbf{X}$ has full column rank.
If the square root exists we can transform the score function such that we have a resulting standardized MVN as follows:

$$I(\boldsymbol{\beta})^{-\frac{1}{2}} U(\boldsymbol{\beta}) \sim MVN_{p+1}(\mathbf{0}_{p+1}, \mathbf{I}_{p+1})$$

Furthermore we have that the following quadratic form is distributed as chi-square

$$U(\boldsymbol{\beta})^\top I(\boldsymbol{\beta})^{-1} U(\boldsymbol{\beta}) = \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{p+1}$$

this expression is nothing but the sum of the squared elements of the standardized score function; if the standardized score function has elements that are all standard gaussian random variables (and those RV are also independent) then they squared sum will be chi-square.

This idea of working with standardized stuff can work theoretically: we will be never be able to compute the actual value of the score function observed in

our sample (because it depends on unknown quantity betas and σ^2). However the behaviour of the standardized score function is crucial in order to study the properties of the Maximum likelihood estimators, especially in context different from the standard gaussian, where we aren't able to come up with a closed formula expression to compute the maximum likelihood estimate and we have to use numerical maximization procedures for loglikelihood (while this stuff is unused in the more simple linear regression).

2.3.1.8 Some general properties of the score function

Aside a moment from gaussian model, in general, let

- $L(\boldsymbol{\theta})$ be likelihood function associated with *any* given parametric statistical model ($\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$)
- we are able to compute the score function $U(\boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta})$

Then under general regularity conditions it is possible to prove that whatever parametric model we're dealing with we have that:

- the expected value of the score function is the null vector: $\mathrm{E}\left[U(\boldsymbol{\theta})\right] = \mathbf{0}_k$
- its variance covariance is equal to the expected fisher information matrix: $\mathrm{Var}\left[U(\boldsymbol{\theta})\right] = I(\boldsymbol{\theta})$
- its standardized version converge in probability to a standardized multivariate normal (zero mean and identity variance covariance matrix): $I(\boldsymbol{\theta})^{-\frac{1}{2}} U(\boldsymbol{\theta}) \xrightarrow{d} MVN_k(\mathbf{0}_k, \mathbf{I}_k)$.
  The idea of standardizing the score function is crucial for studying its asymptotic behaviour: from a technical pov it's possible to prove (look intermediate/advanced texts) that standardized version of the score function converge in distribution to multivariate normal.

These results implies that we can always approximate the distribution of the original score funtion using a MVN with zero expected value and expected fisher information matrix as variance covariance

$$U(\boldsymbol{\theta}) \approx MVN_k(\mathbf{0}_k, I(\boldsymbol{\theta}))$$

The quality of the approximation improves as sample size increase.

So we put aside this results for the future: if some general condition are met our score function can be approximated by a MVN

2.3.2 Maximum likelihood estimation

2.3.2.1 Maximum likelihood estimate for β

The vector $\hat{\mathbf{b}}$ is the maximum likelihood (ML) estimate for $\boldsymbol{\beta}$ if and only if

$$l\left(\hat{\mathbf{b}}, \sigma^2\right) = \max_{\mathbf{b} \in \mathbb{R}^{p+1}} l\left(\mathbf{b}, \sigma^2\right)$$

or equivalently

$$\hat{\boldsymbol{b}} = \arg\max_{\mathbf{b}\in\mathbb{R}^{p+1}} l\big(\mathbf{b}, \sigma^2\big)$$

To find it we need to find the value $\hat{\mathbf{b}}$

- at which the score function (first derivative) is equal to 0, or in matrix terms *the log-likelihood gradient with respect to* $\boldsymbol{\beta}$ *evaluated at* $\hat{\mathbf{b}}$ be zero vector

$$U\Big(\hat{\mathbf{b}}\Big) = \frac{\partial}{\partial\boldsymbol{\beta}} l\big(\boldsymbol{\beta}, \sigma^2\big)\Big|_{\boldsymbol{\beta}=\hat{\mathbf{b}}} = \mathbf{0}_{p+1}$$

- among the several point matching the first condition to have a maximum we need htat second partial derivative evaluated at point must be negative, or in matrix terms the Hessian matrix of the log likelihood function (for $\boldsymbol{\beta}$ evaluated at $\hat{\mathbf{b}}$) be negative definite (equivalent for a matrix of a negative scalar)

$$H\Big(\hat{\mathbf{b}}\Big) = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} l\big(\boldsymbol{\beta}, \sigma^2\big)\Big|_{\boldsymbol{\beta}=\hat{\mathbf{b}}}$$

  in this way $\mathbf{z}^\top H\Big(\hat{\mathbf{b}}\Big)\mathbf{z} < 0$, $\forall \mathbf{z} \neq \mathbf{0}_{p+1}$

So to find maximum we have to find b vectors satysifing these condition

- starting from the first one the score vector is null

$$\begin{aligned} U(\mathbf{b}) = \mathbf{0}_{p+1} &\iff \frac{\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{b})}{\sigma^2} = \mathbf{0}_{p+1} \\ &\iff \frac{\mathbf{X}^\top\mathbf{y}}{\sigma^2} = \frac{\mathbf{X}^\top\mathbf{X}\mathbf{b}}{\sigma^2} \end{aligned}$$

  Now if the matrix $\mathbf{X}$ has full column rank (its column are linearly independent) then $\mathbf{X}^\top\mathbf{X}$ is invertible and we can premultiply both terms of the equation for $(X^\top X)^{-1}$ obtaining

$$\mathbf{b} = \big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top\mathbf{y}$$

- for the second partial derivative we have that

$$H(\boldsymbol{\beta}) = -i(\boldsymbol{\beta}) = -\frac{\mathbf{X}^\top\mathbf{X}}{\sigma^2}$$

  σ^2 is unknown but must be positive; if the matrix $\mathbf{X}$ has full column rank any combination of columns will be a vector different from zero vector, so $z^\top X$ will be different from zero, then $z^\top X^\top X z$ will be always strictly positive scalar and with a minus behind the results will be a negative constant. So $\forall \mathbf{b} \in \mathbb{R}^{p+1}$ we have that $H(\boldsymbol{\beta})$ is negative definite.
  So we have the assurance that this is a maximum

Therefore our mle is

$$\hat{\mathbf{b}} = \big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top\mathbf{y}$$

and as far as $\boldsymbol{\beta}$ is concerned, maximum likelihood estimation is equivalent to least square estimation for Gaussian linear models

2.3.2.2 Properties of the ML estimator for β

The MLE estimators $\hat{\mathbf{b}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ depends on:

- $\mathbf{X}$ the regressor values
- $\mathbf{y}$ the observed outcomes (*realisations of the r. v.* $\mathbf{Y}$)

Conditionally on the regressors values $\mathbf{X}$, $\hat{\mathbf{b}}$ is the realisation of a random vector, that can be expressed as a linear transformation of $\mathbf{Y}$

$$\mathbf{A} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top, \mathbf{b} = \mathbf{0}_{p+1} \implies \hat{\mathbf{B}} = \mathbf{AY} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$$

Thus according to the Gaussian linear model assumption, having $\mathbf{Y}|\mathbf{X} \sim MVN_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$ it turns out that the ML estimator is respectively unbiased, having a varcov matrix corresponding to the inverse of the expected information matrix and gaussian:

$$\begin{aligned}
&\mathrm{E}\left[\hat{\mathbf{B}}\middle|\mathbf{X}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\
&\mathrm{Var}\left[\hat{\mathbf{B}}\middle|\mathbf{X}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}_n\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = I(\boldsymbol{\beta})^{-1} \\
&\hat{\mathbf{B}}|\mathbf{X} \sim MVN_{p+1}\left(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}\right)
\end{aligned}$$

Furthermore regarding variability, thanks to Rao-Cramer theorem, having variance covariance matrix coinciding with the inverse of expected information matrix we conclude that the MLE is the efficient estimator for β.
(inverse of cramer-rao lower bound is the minimum variance that we can achieve for an unbiased estimator for a given parameter: if exists an estimator achieving the cramer-rao lower bound, then that estimator is unique, so there are no other estimators with less variability).
So if the model assumption holds the mle estimator for beta are not only unbiased but also efficient.

It is important to check whether the assumptions are adequate for the specific dataset we're dealing with

2.3.2.3 Some general results related to ML method

Some general words on ML method: let

- $\hat{\mathbf{T}}$ be Maximum likelihood estimator for $\boldsymbol{\theta}$ (a random variable on the sample space)
- $\hat{\mathbf{t}} = \arg\max_{\mathbf{t}\in\Theta} l(\mathbf{t})$ be the maximum likelihood estimate for $\boldsymbol{\theta}$ (*sample realisation of* $\hat{\mathbf{T}}$)

Under general regularity conditions (the same for property of the score function) it is possibile to show that:

- the standardized version of the ML estimator has an asymptotic distribution converging to standard MVN $I(\boldsymbol{\theta})^{\frac{1}{2}}\left(\hat{\mathbf{T}} - \boldsymbol{\theta}\right) \xrightarrow{d} MVN_k(\mathbf{0}_k, \mathbf{I}_k)$;
- thus in general $\hat{\mathbf{T}} \approx MVN_k\left(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}\right)$

So no matter what model we are dealing with, if the model satisfies the basic regularity conditions, whatever functional form it takes (even when an explicit analytical form for computing $\hat{\mathbf{t}}$ does not exist)

- the ML estimator for $\boldsymbol{\theta}$ is *asymptotically unbiased* (we have no guarantee that it is unbiased but at least asymptotically when sample size increase it is)
- it is *asymptotically efficient* (having the asymptotic variance covariance matrix coinciding with the inverse of the expected fisher information, the cramer rao lower bound).

For gaussian linear model they are unbiased and efficient as well (for any sample size).

2.3.2.4 Maximum likelihood estimate for σ^2

The other parameter of the gaussian model is σ^2; when performing regression analysis main focus is the betas, but we still have a variance so we need to compute the estimate.
Once we've found the ML estimates for the regression coefficients we can look for a ML estimate for the σ^2. It is possible to prove (compute the first and second partial derivative of loglikelihood with respect to σ^2, set the first equal to zero and select among them where second partial derivative is negative: here is a scalar so it's a simple real valued function) that the estimator of σ^2, s^2 is basically the variance of sample raw residuals

$$\hat{s}^2 = \underset{s^2 \in \mathbb{R}^+}{\arg\max}\, l\left(\hat{\mathbf{b}}, s^2\right) = \frac{\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}\right)^2}{n} = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^\top \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)}{n} = \frac{\mathbf{e}^\top \mathbf{e}}{n}$$

There will be a score function, an observed and expected information matrix as well but we don't focus on it being less interesting for our purpose.
A single raw residual is something like

$$e_i = y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}, \quad i = 1, \dots, n$$

while exploiting matrix algebra we see the vector of raw residuals $\mathbf{e}$ can be expressed as

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\underbrace{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}}_{\hat{\mathbf{b}}} = \left[\mathbf{I}_n - \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top}_{\mathbf{H}}\right]\mathbf{y} = \underbrace{\left[\mathbf{I}_n - \mathbf{H}\right]}_{\mathbf{M}}\mathbf{y}$$

It ends with $\mathbf{e}$ being expressed as linear transformation of the vector $\mathbf{y}$

- $\mathbf{H}$ the so called *hat matrix* (which transforms the observed $\mathbf{y}$ in the fitted $\hat{\mathbf{y}} = X\hat{\mathbf{b}}$
- $\mathbf{M} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$, sometimes referred as *residual maker matrix* (transforming $\mathbf{y}$ into $\mathbf{e}$), is
  - symmetric
  - idempotent ($\mathbf{MM} = \mathbf{M}$)
  - usually not diagonal
  - not invertible

2.3.2.5 Properties of raw residuals

Given the Gaussian linear model assumptions, it is possible to prove that:

- being a linear trasformation of vector $\mathbf{y}$ will have a multivariate gaussian distribution (if the gaussian assumption are ok its mean is 0 while the varcov is not diagonal):

$$\mathbf{e}|\mathbf{X} \sim MVN_n\left(\mathbf{0}_n, \sigma^2\mathbf{M}\right)$$

- it can be proved that if model assumptions holds, the sum of the squares of residuals divided by σ^2 (conditional on the value of the regressors), is distributed as a Chi square with $n-(p+1)$ (where p is the number of regressors) degrees of freedom:

$$\left.\frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2}\right|\mathbf{X} \sim \chi^2_{n-p-1}$$

- therefore the expected value of the sum of the squares of residuals conditional to the regressor is (taking

$$\mathrm{E}\left[\mathbf{e}^\top\mathbf{e}|\mathbf{X}\right] = \sigma^2(n-p-1)$$

- $\dfrac{\mathbf{e}^\top\mathbf{e}}{\sigma^2}$ is independent of $\hat{\mathbf{B}}$ (ML estimates of $\boldsymbol{\beta}$)

These properties are crucial for establishing the properties of ML estimator for σ^2

2.3.2.6 Properties of the maximum likelihood estimator for σ^2

The estimator for σ^2 is

$$\hat{S}^2 = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}\right)^\top\left(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}\right)}{n}$$

Given the Gaussian linear model assumptions, and exploiting the properties of the raw residuals, it is possible to prove that:

- the maximum likelihood estimator for σ^2 is biased since

$$\mathrm{E}\left[\left.\hat{S}^2\right|\mathbf{X}\right] = \sigma^2\frac{n-p-1}{n} \neq \sigma^2$$

  The (negative) bias we have (the ML estimate tend to underestimate the true σ^2) get cancelled out as n increases (so ML estimator is asymptotically unbiased as seen before), that is

$$\mathrm{E}\left[\left.\hat{S}^2\right|\mathbf{X}\right] \underset{n\to\infty}{\longrightarrow} \sigma^2$$

- if we are interested in unbiased estimator for σ^2 a corrected expression is obtaining by dividing by $n-p-1$ (degrees of freedom) instead of n

$$S^2 = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}\right)^\top\left(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}\right)}{n-p-1}$$

(this is nothing but a generalization of what occurs with the sample variance where using the mean make the degrees of freedom to decrease of 1) The unbiased version (not the ML one) is typically what is returned from software to estimate σ^2

- both ML and unbiased estimator $\hat{S}^2$, S^2 are independent of $\hat{\mathbf{B}}$.
  This is important property when we want to compute test statistics (all the most relevant test statistics can be expressed as ratio between numerator depending on ML estimates of beta and the denominator depending on an estimate of σ^2: knowing that num and denom are independent makes the derivation of distribution of test statistics under null hypothesis much easier.

2.3.2.7 Standardised residuals

We have seen that raw residual have a more or less known distribution (not considering σ^2)

$$\mathbf{e}|\mathbf{X} \sim MVN_n\big(\mathbf{0}_n, \sigma^2\mathbf{M}\big)$$

In general $\mathbf{M}$:

- is not diagonal, so resids are not independent
- is usually not homoschedastic: its diagonal elements of $\mathbf{M}$ differ from one another. They differ because the diagonal elements of M depend of the value of the regressors associated with each unit in the sample, eg for the i-th unit it will be

  $$\mathbf{M}_{ii} = 1 - \mathbf{x}_i^\top \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}_i = 1 - \mathbf{H}_{ii}$$

  units with different covariate pattern will have residuals whose variance is different

The variance of each residual depends on σ^2 (for which we have an unbiased estimator) and the $\mathbf{M}$ diagonal entry; these latter are a function of the regressors, so condition on the regressor we can compute the exact value.
Starting from the raw residual we can come up with two refined version:

- **Pearson residuals**: $e_i^P = \dfrac{e_i}{\sqrt{s^2}} \quad i = 1, \ldots, n$ obtained dividing each raw residual for the square root of unbiased estimate of σ^2.
  In this way however we don't obtain yet a residual which is homoschedasti but it will show up as well in the GLM
- **Standardised residuals**: $r_i = \dfrac{e_i}{\sqrt{s^2(1-\mathbf{H}_{ii})}} \quad i = 1, \ldots, n$ which divided the residual for a measure that take into account the i-th diagonal element on the residual maker matrix).

If all the assumptions of gaussian model are met it is possible to prove that we have that the asymptotic is the following

$$\mathbf{r} = (r_1, r_2, \ldots, r_n)^\top \Big| \mathbf{X} \xrightarrow{d} MVN_n(\mathbf{0}_n, \mathbf{I}_n)$$

Approximately (by n fixed), standardised residuals from a Gaussian linear models are equivalent to an observed sample drawn from an n-dimensional standardised Gaussian random vector.

Important remark 4. Idea: once we fitted the model we can inspect the standardized residuals (eg by plots) and if we find some deviation of behaviour from standard MVN (IID random vector of standard gaussians), then we can conclude that the model assumptions are not adequate.
When assumption holds this would not happen.
This is one of the most crucial step to do

Chapter 3

Linear hypotheses

Important remark 5. Typically we are interested in test hypotheses on parameters: we will focus on linear hypotheses, so called because they can be expressed as a linear system (involving the regression coefficient betas).

3.1 Linear hypotheses

3.1.1 Linear hypotheses on β

Our setup

- Gaussian linear model: $\boldsymbol{Y}|\mathbf{X} \sim MVN_n\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\big)$
- the $(p+1) \times 1$ parameter vector

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

- suppose we have
  - $\mathbf{K}$, $q \times (p+1)$ matrix, composed of known constants with full row rank q (rows are linearly independent: q must be smaller or equal to $p+1$)
  - $\mathbf{t}$, $q \times 1$ vector composed of known constants

  Any linear hypotheses on $\boldsymbol{\beta}$ can be expressed as a system of linear equations:

$$H_0 : \mathbf{K}\boldsymbol{\beta} = \mathbf{t}$$

  In the latter we're specifying that linear combinations of regressors are equal to given constants

Example 3.1.1 (Linear hypotheses on β: some examples)**.** Supposing $p = 3$, the following are different systems of hypotheses to be tested

(A) if $\mathbf{K} = [0\ 1\ 0\ 0]$, and $\mathbf{t} = 0$ then we obtain a simple test on a single coefficient

$$H_0 : \beta_1 = 0$$

(B) if $\mathbf{K} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\mathbf{t} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ then we pick two coefficients and put them equal to 0 we have

$$H_0 : \begin{cases} \beta_1 = 0 \\ \beta_3 = 0 \end{cases}$$

or in a more common/compact way

$$\beta_1 = \beta_3 = 0$$

(C) if $\mathbf{K} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\mathbf{t} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ then

$$H_0 : \begin{cases} \beta_1 = 0 \\ \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$$

or

$$\beta_1 = \beta_2 = \beta_3 = 0$$

which is called *linear independence* test (if null is true there's independence between the dependent variable and the regressors: the latters have no effect on the dependent variable)

(D) if $\mathbf{K} = [0\ 1\ 0\ -1]$, $\mathbf{t} = 0$ then

$$H_0 : \beta_1 = \beta_3$$

In this case by chosing a different $\mathbf{K}$ we relate coefficient between them, not only to constants.
Note that $H_0 : \beta_1 = \beta_3$ is much more general than $H_0 : \beta_1 = \beta_3 = 0$ seen in (B); here they can be equal no matter what value they take. In some situations is useful to test hypothesis on equivalence of regression coefficients without specifying the given value

(E) we can test that a coefficient to be a constant, not necessarily 0: eg if $\mathbf{K} = [0\ 1\ 0\ 0]$ and $\mathbf{t} = 3$

$$H_0 : \beta_1 = 3$$

3.1.2 Nested linear models

Linear hypotheses (A), (B) and (C) in the previous example lead to Gaussian linear models that can be obtained by removing some regressors from the starting model: eg if the starting model is

$$\mathrm{E}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

then:

(A) $\Rightarrow \mathrm{E}_{H_0}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} = \mathrm{E}\left[Y_i|x_{2i}, x_{3i}\right]$

(B) $\Rightarrow \mathrm{E}_{H_0}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 + \beta_2 x_{2i} = \mathrm{E}\left[Y_i|x_{2i}\right]$

(C) $\Rightarrow \mathrm{E}_{H_0}\left[Y_i|x_{1i}, x_{2i}, x_{3i}\right] = \beta_0 = \mathrm{E}\left[Y_i\right]$

These are nested models, that is models that are obtained after removing one or more regressors from a starting one.

3.1.3 Likelihood ratio test (LRT) statistics - 1

To compare models we use the LRT statistics, which is a general test which we can use to test any kind of hypothesis on the parameter of a parametric statistical model.
It's defined as

$$LRT = \frac{L\left(\hat{\mathbf{b}}, \sigma^2\right)}{L\left(\hat{\mathbf{b}}_{H_0}, \sigma^2\right)}$$

so as the ratio between

- $\hat{\mathbf{b}} = \arg\max_{\mathbb{R}^{(p+1)}} l\left(\mathbf{b}, \sigma^2\right)$, that is the maximized likelihood (value of likelihood function at the ML estimates)
- $\hat{\mathbf{b}}_{H_0} = \arg\max_{\{\mathbf{b}:\mathbf{Kb}=\mathbf{t}\}\subset\mathbb{R}^{(p+1)}} l\left(\mathbf{b}, \sigma^2\right)$ the maximized likelihood under the restriction imposed by the system of linear hypothesis (that is considering in the parameter space only those elements introduced by the linear restriction, we have a constrainted maximization)

Equivalently, using the loglikelihood we have the differences, that is

$$2\left[l\left(\hat{\mathbf{b}}, \sigma^2\right) - l\left(\hat{\mathbf{b}}_{H_0}, \sigma^2\right)\right]$$

The point now we focus on is how to find the denominator of the LRT

3.2 Constrained maximum likelihood estimation

3.2.1 The Method of Lagrange multipliers

This is a generic method for constrainted optimization: the idea is to work on a slightly different version of the function to be optimized.
$\hat{\mathbf{b}}_{H_0}$ maximises $l\left(\boldsymbol{\beta}, \sigma^2\right)$ in the parameter subspace $\{\mathbf{b} : \mathbf{Kb} = \mathbf{t}\} \subset \mathbb{R}^{(p+1)}$. Rather than maximizing l

- we maximize a modified version l^*

$$l^*(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) = l(\boldsymbol{\beta}, \sigma^2) - \boldsymbol{\alpha}^\top (\mathbf{K}\boldsymbol{\beta} - \mathbf{t})$$

  where $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{bmatrix}$ is a $q \times 1$ vector containing unknown *Lagrange multipliers*. So the original likelihood is modified using $\boldsymbol{\alpha}$ and something regarding the linear restriction we're interested in

- the maximization is done with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$: what we found of out with respect of $\boldsymbol{\beta}$ is a set of value satisfying the conditions and maximizing the likelihood under them

Important remark 6. Some technical passages follows: take the main message above and don't worry

To do the maximization, the following system equations must be solved:

$$\begin{cases} U(\mathbf{b}) = \dfrac{\partial}{\partial \boldsymbol{\beta}} l^*(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) \Big|_{\boldsymbol{\beta}=\mathbf{b}} = \mathbf{0}_{p+1} \\ \\ U(\mathbf{a}) = \dfrac{\partial}{\partial \boldsymbol{\alpha}} l^*(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) \Big|_{\boldsymbol{\alpha}=\mathbf{a}} = \mathbf{0}_q \end{cases}$$

It's a $p + 1 + q$ equation sistem where the first $p + 1$ are related to betas and last q to alphas).
We have to compute the first partial derivatives with respect both to beta and alphas:

$$\frac{\partial}{\partial \boldsymbol{\beta}} l^*(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) = U(\boldsymbol{\beta}) - \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\alpha}^\top (\mathbf{K}\boldsymbol{\beta} - \mathbf{t}) = \frac{\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} - \mathbf{K}^\top \boldsymbol{\alpha}$$
$$\frac{\partial}{\partial \boldsymbol{\alpha}} l^*(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) = -\frac{\partial}{\partial \boldsymbol{\alpha}} \boldsymbol{\alpha}^\top (\mathbf{K}\boldsymbol{\beta} - \mathbf{t}) = -(\mathbf{K}\boldsymbol{\beta} - \mathbf{t})$$

Remembering general rules:

$$\frac{\partial}{\partial \boldsymbol{\delta}} \mathbf{A}\boldsymbol{\delta} = \mathbf{A}^\top$$
$$\frac{\partial}{\partial \boldsymbol{\delta}} \boldsymbol{\delta}^\top \mathbf{A} = \mathbf{A}$$

The idea is to solve first the first $p+1$ equations with respect to the betas, then we plug the solutions in the remaining q equations.
Consider the first $p + 1$ equations:

$$\frac{\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{b})}{\sigma^2} - \mathbf{K}^\top \mathbf{a} = \mathbf{0}_{p+1}$$
$$\frac{\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{b})}{\sigma^2} = \mathbf{K}^\top \mathbf{a}$$
$$\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\mathbf{b} = \sigma^2 \mathbf{K}^\top \mathbf{a}$$
$$\mathbf{X}^\top \mathbf{X}\mathbf{b} = \mathbf{X}^\top \mathbf{y} - \sigma^2 \mathbf{K}^\top \mathbf{a}$$

If $\mathbf{X}$ has full column rank $(p+1)$ then

$$\begin{aligned}\hat{\mathbf{b}}_{H_0} &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} - \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\mathbf{a} \\ &= \hat{\mathbf{b}} - \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\mathbf{a}\end{aligned}$$

So this last is what we'll use in the remaining equations; note that σ^2 and $\mathbf{a}$ are unknown.
Now exploiting the formula for $\hat{\mathbf{b}}_{H_0}$ in the last q equations:

$$\begin{aligned}\mathbf{K}\left[\hat{\mathbf{b}} - \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\mathbf{a}\right] &= \mathbf{t} \\ \mathbf{K}\hat{\mathbf{b}} - \sigma^2\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\mathbf{a} &= \mathbf{t} \\ \sigma^2\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\mathbf{a} &= \mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\end{aligned}$$

If $\mathbf{K}$ has full row rank (q)

$$\hat{\mathbf{a}} = \frac{1}{\sigma^2}\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)$$

So finally, to obtain the constrained maximum likelihood estimate, by substituting $\hat{\mathbf{a}}$ for $\boldsymbol{\alpha}$ in the formula for $\hat{\mathbf{b}}_{H_0}$:

$$\begin{aligned}\hat{\mathbf{b}}_{H_0} &= \hat{\mathbf{b}} - \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\frac{1}{\sigma^2}\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) \\ &= \hat{\mathbf{b}} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) \qquad (3.1)\end{aligned}$$

So we can get a general analytical expression (complicated ok but don't worry) to compute the constrainted betas for any possible system of hypotheses.
Note that, even the constrained maximum likelihood estimate $\hat{\mathbf{b}}_{H_0}$ can be computed without knowing the true value of σ^2. As expected the returned betas satisfy the systems of constraints/linear hypotheses since:

$$\mathbf{K}\hat{\mathbf{b}}_{H_0} = \mathbf{K}\hat{\mathbf{b}} - \underbrace{\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}}_{\mathbf{I}_q}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) = \mathbf{K}\hat{\mathbf{b}} - \mathbf{K}\hat{\mathbf{b}} + \mathbf{t} = \mathbf{t}$$

3.2.2 Residuals of the constrained model

In order to have the expression to compute the LRT for gaussian models it is worth look at the residuals associated with the constrainted model. Basic matrix algebra show that:

$$\begin{aligned}\mathbf{e}_{H_0} &= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{H_0} \\ &= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{H_0} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}\hat{\mathbf{b}} \\ &= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) = \mathbf{e} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)\end{aligned}$$

Looking at the last we conclude that the residual of the constrainted model are equal to the residual of the unconstrainted plus something else (depending on

data and difference between constraint and unconstrainted estimates).
What if we compute the sum of the squared constrainted residuals? we have:

$$\begin{aligned}\mathbf{e}_{H_0}^\top \mathbf{e}_{H_0} &= \left[\mathbf{e} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)\right]^\top \left[\mathbf{e} + \mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)\right] \\ &= \mathbf{e}^\top\mathbf{e} + \mathbf{e}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) + \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top\mathbf{X}^\top\mathbf{e} + \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)\end{aligned}$$

so we end with four terms. Now with basic algebra we can show that in general:

$$\begin{aligned}\mathbf{X}^\top\mathbf{e} &= \mathbf{X}^\top\left[\mathbf{y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\right] = \mathbf{X}^\top\mathbf{y} - \underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}}_{\mathbf{I}_{p+1}}\mathbf{X}^\top\mathbf{y} \\ &= \mathbf{0}_{p+1}\end{aligned}$$

So coming back to the sum of square of constrainted residuals it simplifies to the sum of squares of unconstrainted residuals plus a quadratic function

$$\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} = \mathbf{e}^\top\mathbf{e} + \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)$$

Now if $\mathbf{X}$ has full column rank, then $\mathbf{X}^\top\mathbf{X}$ is positive definite and so if $\hat{\mathbf{b}} \neq \hat{\mathbf{b}}_{H_0}$:

$$\begin{aligned}&\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) > 0 \\ &\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} > \mathbf{e}^\top\mathbf{e}\end{aligned}$$

So when we introduce linear restriction we end up with a constrainted models where squared sum of residuals is always larger than unrestricted one (by introducing restriction we deteriorate the model). The amount of difference depends on the difference between constraited and undconsrtaited estimates and the matrix $\mathbf{X}$.
We can see what happens if we replace $\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}$ with what found before (in 3.1) that is:

$$\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{K}^\top\left[\mathbf{K}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)$$

Therefore the difference between the sum of squared residuals is:

$$\begin{aligned}&\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} - \mathbf{e}^\top\mathbf{e} \\ &= \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{H_0}\right) \\ &= \left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)^\top\left[\mathbf{K}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{K}^\top\right]^{-1}\mathbf{K}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{K}^\top\left[\mathbf{K}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right) \\ &= \left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)^\top\left[\mathbf{K}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}} - \mathbf{t}\right)\end{aligned}$$

In the end, to know how much the errors increase we do not actually need to fit the model under the restriction because $\hat{\mathbf{b}}_{H_0}$ is not in the final formula.
Thanks to this results we're able to compute the value of the lrt simply by starting from the unconstrainted ML estimate

3.3 Likelihood ratio properties

3.3.1 LRT statistics - 2

Developing a bit the loglikelihood version we have

$$\Delta l = 2\ln\left[\frac{L\left(\hat{\mathbf{b}},\sigma^2\right)}{L\left(\hat{\mathbf{b}}_{H_0},\sigma^2\right)}\right] = -n\ln 2\pi\sigma^2 - \frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^\top\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)}{\sigma^2} + n\ln 2\pi\sigma^2 + \frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}_{H_0}\right)^\top\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}_{H_0}\right)}{\sigma^2}$$

$$= \frac{\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} - \mathbf{e}^\top\mathbf{e}}{\sigma^2} = \frac{\left(\mathbf{K}\hat{\mathbf{b}}-\mathbf{t}\right)^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{b}}-\mathbf{t}\right)}{\sigma^2}$$

So it turns out it depends on the difference of sum of squared residuals and, as we've anticipated, it is not necessary to know $\hat{\mathbf{b}}_{H_0}$ in order to compute the LR test statistic and to derive its distribution.
Once computed $\hat{\mathbf{b}}$, once chosen $\mathbf{K}$ and $\mathbf{t}$, in order to compute LRT statistics we don't need the model under restriction (we can forget about lagrange multiplier)
In the quadratic form at the numerator the closer $K\hat{\mathbf{b}}-\mathbf{t}$ is to $\mathbf{0}$ (so the closer is $\hat{\mathbf{b}}$ to $\hat{\mathbf{b}}_{H_0}$) the smaller will be the value of the test statistics; on the contrary the larger the difference between $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}_{H_0}$ the larger will be the value of the test stastics.
So the closer the unconstrainted model is to the constrainted one the smaller will be the value of the test statistics

Important remark 7. per il prof importante la formula finale dell'LRT

Important remark 8. To do proper test, we need to know the distribution of the LRT.

3.3.2 LRT statistic distribution - σ^2 known

There are different way to come up with the distribution. One is the following: start by hypothesizing σ^2 is known.
By recalling properties of the maximum likelihood estimator for $\boldsymbol{\beta}$ we have that

$$\hat{\mathbf{B}} \sim MVN_{p+1}\left(\boldsymbol{\beta},\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

the variance is the inverse of the expected fisher information matrix. If we apply a linear transformation then

$$\mathbf{K}\hat{\mathbf{B}}-\mathbf{t} \sim MVN_q\left(\mathbf{K}\boldsymbol{\beta}-\mathbf{t},\sigma^2\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right)$$

If by hypothesis H_0 is true $\mathbf{K}\boldsymbol{\beta}=\mathbf{t}$ so $\mathbf{K}\boldsymbol{\beta}-\mathbf{t}=\mathbf{0}_q$ therefore

$$\mathbf{K}\hat{\mathbf{B}}-\mathbf{t}|H_0 \sim MVN_q\left(\mathbf{0}_q,\sigma^2\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right)$$

Applying the standardization for a MVN, by dividing for the square root of variance/covariance matrix, we end up with a standardized MVN

$$\boldsymbol{Z} = \frac{1}{\sigma}\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-\frac{1}{2}}\left(\mathbf{K}\hat{\mathbf{B}}-\mathbf{t}\right)|H_0 \sim MVN_q(\mathbf{0}_q,\mathbf{I}_q)$$

It turns out that the sum of squares of the vector $\mathbf{Z}$, that is $\boldsymbol{Z}^\top \boldsymbol{Z}$, is exactly the expression for the LRT statstics we found before, that is we end up with the previous quadratic form (sum of squares). But since we're taking the sum of q standardized independent gaussian rv, we obtain that are distributed as a χ^2 with q decgrees of freedom (element in $\mathbf{Z}$)

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \frac{\left(\mathbf{K}\hat{\mathbf{B}} - \mathbf{t}\right)^\top \left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1} \left(\mathbf{K}\hat{\mathbf{B}} - \mathbf{t}\right)}{\sigma^2} = \Delta l | H_0 \sim \chi^2_q$$

If σ^2 were known than LRT would be $\sim \chi^2_q$ under null hypothesis. In the more realistic situation where σ^2 is unknown we replace it with an estimate. We now see what is the impact of this replacing in the distribution.

3.3.3 LRT statistic distribution - σ^2 unknown

We know that (*Proprieties of raw residuals*) the sum of raw residual squared over σ^2, $\frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2}$:

- is chi squared distributed $\frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2} \sim \chi^2_{n-p-1}$
- is independent between $\hat{\mathbf{B}}$ and S^2

we can in some sense replace the unknown σ^2 with an estimate based on the sum of the squares of residuals, and in doing so we end with a new test statistic representable as a ratio of independent chi-squares.
If H_0 is true

$$\begin{aligned}\frac{\Delta l}{\frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2}}\frac{n-p-1}{q} &= \frac{\frac{\mathbf{e}_{H_0}^\top\mathbf{e}_{H_0} - \mathbf{e}^\top\mathbf{e}}{\sigma^2}}{\frac{\mathbf{e}^\top\mathbf{e}}{\sigma^2}}\frac{n-p-1}{q} \\ &= \frac{\frac{\left(\mathbf{K}\hat{\mathbf{B}}-\mathbf{t}\right)^\top\left[\mathbf{K}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{K}^\top\right]^{-1}\left(\mathbf{K}\hat{\mathbf{B}}-\mathbf{t}\right)}{q}}{\frac{\mathbf{e}^\top\mathbf{e}}{n-p-1}} | H_0 \sim \frac{\frac{\chi^2_q}{q}}{\frac{\chi^2_{n-p-1}}{n-p-1}} = F_{(q,n-p-1)}\end{aligned}$$

From a technical pov what we can do is divide the LRT and the sum of squared residuals by their degrees of freedom (lrt has q degrees of freedom, sum of squares of residuals over $sigma^2$ has $n-p-1$) and then divide the obtained quantity: by doing this we cancel out σ^2 and so we are left with the ratio of two independent chi square distributed statistics (at denominator we have the unbiased estimator of σ^2) divided by they degrees of freedom.
So here we have a test statistics which involves only known quantities (once we have our sample) and is distributed as an F with q and $n-p-1$ degrees of freedom.
This test statistics formally speaking is not the LRT (which is at the numerator) but basically we can compute it and we now its distribution under null so we can use it for inference

3.3.4 Applications

Comparison between complete and reduced models When linear hypotheses lead to the removal of q regressors, raw residuals $\mathbf{e}_{H_0}$ correspond to the residuals of a reduced model (nested in the complete model) and the previous general test becomes writable as

$$\frac{\Delta l}{\frac{\mathbf{e}^\top \mathbf{e}}{\sigma^2}} \frac{n-p-1}{q} = \frac{\frac{SSE_{M_{H_0}} - SSE_{M_C}}{q}}{\frac{SSE_{M_C}}{n-p-1}}$$

where

- SSE_{M_C} is the residual sum of squares for the complete model (with all regressors)
- $SSE_{M_{H_0}}$ is the residual sum of squares for the reduced model (after excluding q regressors)

Wald test statistics There is an interesting property of the LRT for hypotheses such $H_0 : \beta_j = 0$. Here we have that LRT takes a simplified expression, as ratio of the estimate of interest and the square root of its variance

$$\Delta l = \frac{\hat{B}_j^2}{\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}} = \left[\frac{\hat{B}_j}{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} \right]^2$$

where $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}_{jj}$ is the j-th element on the main diagonal of $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$. In case:

- σ^2 *known* then $\frac{\hat{B}_j}{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} | H_0 \sim N(0,1)$
- σ^2 *unknown* we replace it with unbiased estimator we end up with the test statistics $\frac{\hat{B}_j}{S \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} | H_0 \sim t_{n-p-1}$

3.4 Confidence intervals

Starting from properies of ML estimators we can come up with confidence intervals for a parameter

Considering the pivotal quantity for β_j: we know that ML estimator has a gaussian distribution we can standardize it by subtracting its expected value (being unbiased its the real beta) and divide by the square root of its variance will have a standard gaussian distribution

$$\frac{\hat{B}_j - \beta_j}{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} \sim N(0,1)$$

then we can end up with

- a gaussian intervals (if σ^2 known) at a $1-\alpha$ confidence level:

$$\left[\hat{b}_j - z_{\frac{\alpha}{2}}\sqrt{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}, \hat{b}_j + z_{\frac{\alpha}{2}}\sqrt{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}\right]$$

- a student-t intervals (if σ^2 unknown) at a $1-\alpha$ confidence level:

$$\left[\hat{b}_j - t_{\frac{\alpha}{2},n-p-1}\sqrt{s^2(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}, \hat{b}_j + t_{\frac{\alpha}{2},n-p-1}\sqrt{s^2(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}\right]$$

Important remark 9. We seen main tools for linear hypothesis testing (for betas, we could test variance as well but less interesting). Now we see some example of use of this tools, as well as the use of categorical regressors.

Chapter 4

Use of categorical regressors

4.1 Unordered categories

4.1.1 Motivating example

Example 4.1.1 (Glucose level in blood and ethnic origin)**.** A glucose level in blood between 100 and 125 mg/dL represents a risk factor for developing diabetes.
Aim: Are there systematic differences in the glucose level among people with different *ethnic origins*?
Available information: Glucose level and ethnic origin (White/African American/other) on a sample of 2020 women not affected by diabetes after menopause. Some basic info in graphs 4.1: in the sample most of the women are Caucasian. In terms of conditional distribution more or less the boxplox show similar variability with strong overlap, with some differences in location. The zoom on conditional means + CI highlight the differences in average glocose level (there's differences in the width due to differences in group sample sizes).
Hypothesis of interest: Absence of significant differences in the average glucose level among different ethnic groups. We want to check that the conditional expected value is the same for different groups:

$$H_0 : \mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{White}\right] = \mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{Other}\right] = \mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{African American}\right]$$
$$i = 1, \ldots, 2020$$

Remark 6. Inferential tools we can adopt for the hypothesis of interest are:

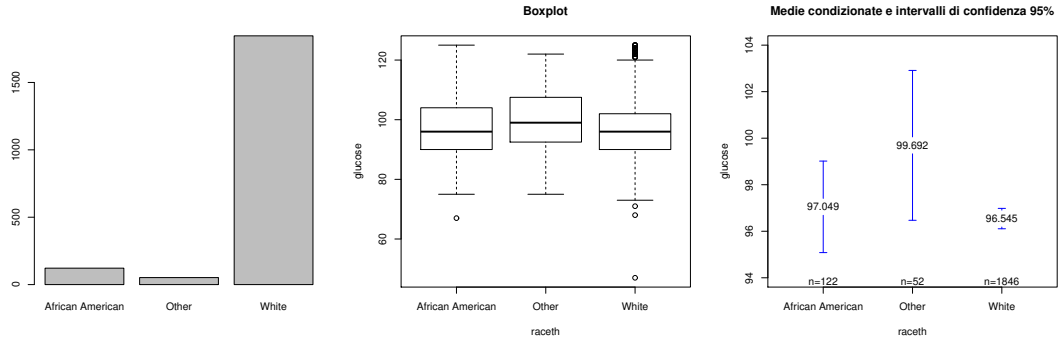
- One-way ANOVA
- Gaussian linear models with indicator/dummy variables

4.1.2 One-way ANOVA

```
> summary(aov(glucose ~ raceth,data=hers.nod))
              Df   Sum Sq   Mean Sq   F value   Pr(>F)
 raceth        2      521    260.51     2.747   0.0643
 Residuals  2017   191259     94.82
```

4.1.3 Linear regression with a qualitative regressor

First we have to do numeric coding of categorical regressor :



(a) (b) (c)

Figure 4.1: Glucose level and ethnic origins

- we can introduce 3 dummy variables, one for each category as follows (1 for the category considered, 0 for the others):

	x_{Ai}	x_{Oi}	x_{Wi}
	$\texttt{African American}_i$	$\texttt{Other}_i$	$\texttt{White}_i$
$\texttt{raceth}_i = \texttt{African American}$	1	0	0
$\texttt{raceth}_i = \texttt{Other}$	0	1	0
$\texttt{raceth}_i = \texttt{White}$	0	0	1

  So 3 indicator variables allow to code a qualitative regressor with 3 categories

- however it is necessary to consider the context of a multiple linear regression model. These 3 indicator variables sums up to 1, for any sample unit:

$$x_{Ai} + x_{Oi} + x_{Wi} = 1$$

  If they are included in a linear model along with an intercept term, the corresponding regressor matrix $\mathbf{X}$ will not have full column rank (being the intercept regressor a linear combination, simple sum of, the dummies introduced)

- what we can do is basically two thing:

  1. exclude one of the indicator variables: the corresponding category is termed *baseline/reference category.*
  2. exclude the intercept from the estimation and leave all the three dummies

4.1.3.1 Using a baseline category

For the first strategy suppose we exclude the african american dummy from the estimation, obtaining the model

$$\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i\right] = \beta_0 + \beta_1 \texttt{Other}_i + \beta_2 \texttt{White}_i$$

then:

$$\begin{aligned}
\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{African American}\right] &= \beta_0 \\
\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{Other}\right] &= \beta_0 + \beta_1 \\
\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{White}\right] &= \beta_0 + \beta_2
\end{aligned}$$

African american becomes the reference category since each regression coefficient (β_1, β_2) represents the difference between the conditional expected value given the

corresponding category and the conditional expected value given the *reference category* (which is β_0).

In this context our hypothesis the absence of significant differences in the average glucose level among different ethnic groups

$$H_0 : \mathrm{E}[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{White}] = \mathrm{E}[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{Other}] = \mathrm{E}[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{African American}]$$

can be translated in a system of linear hypotheses on parameters of the gaussian model like

$$H_0 : \begin{cases} \beta_0 = \beta_0 + \beta_1 \\ \beta_0 = \beta_0 + \beta_2 \\ (\beta_0 + \beta_1 = \beta_0 + \beta_2) \end{cases}$$

basic algebra leads to the equivalent $H_0 : \beta_1 = \beta_2 = 0$.
The results of the gaussian linear regression model are presented below

```
> summary(modello1)
Call:
lm(formula = glucose ~ raceth, data = hers.nod)
...
 Coefficients:
               Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)    97.0492        0.8816   110.081     <2e-16
 racethOther     2.6431        1.6127     1.639      0.101
 racethWhite    -0.5042        0.9103    -0.554      0.580

Residual standard error:  9.738 on 2017 degrees of freedom


Multiple R-squared:  0.002717, Adjusted R-squared:  0.001728
F-statistic:  2.747 on 2 and 2017 DF, p-value:  0.06434
```

The relevant test statistic:

- t test statistics allow to evaluate differences between each category and the reference category: the regression coefficients for the two indicator variables $\texttt{Other}_i$ and $\texttt{White}_i$ are not significantly different from 0;
- last row reports the F test statistic of our interest in this case (the linear independence hypothesis) which test that all the betas are 0. In this case we cannot refuse the null hypothesis so there's no evidence on effect of race on the dependent variable (despite being near the 0.05 treshold).
  The test can be reproduced using

```
> K1
     1   2   3
 1   0   1   0
 2   0   0   1

> t1

[1] 0 0

> linearHypothesis(modello1, K1, t1, test="F")
Linear hypothesis test

Hypothesis:

racethOther = 0
racethWhite = 0
Model 1:  restricted model

Model 2:  glucose ~ raceth
     Res.Df      RSS   Df   Sum of Sq        F   Pr(>F)
 1     2019   191780
 2     2017   191259    2      521.02   2.7473   0.06434
```

In this example:

- $\mathbf{e}_{H_0}^\top \mathbf{e}_{H_0} = 191780$ (sum of squared residuals of the restricted model)
- $\mathbf{e}^\top \mathbf{e} = 191259$ (sum of squared residuals of the unrestricted model),
- $q = 2$,
- $\mathbf{e}_{H_0}^\top \mathbf{e}_{H_0} - \mathbf{e}^\top \mathbf{e} = 521.02$
- the statistic is
$$\frac{\mathbf{e}_{H_0}^\top \mathbf{e}_{H_0} - \mathbf{e}^\top \mathbf{e}}{\mathbf{e}^\top \mathbf{e}} \frac{n-p-1}{q} = 2.743$$
- finally we see that the p-value coincides with the p-value reported in the model above

Choice of the reference category Regarding:

- the choice of the reference category is arbitrary
- the estimates for the regression coefficients will change, but the global measures remains the same

The default choice in `R` is the first category, in alphabetical order:

```
                  Other   White
African American      0       0
           Other      1       0
           White      0       1
```

Instead if we use caucasian women as reference category:

```
                       1   2
    African American   1   0
               Other   0   1
               White   0   0
```

The meaning of the regression coefficients changes accordingly (we use different symbols δ to denote it):

$$\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i\right] = \delta_0 + \delta_1 \texttt{raceth1}_i + \delta_2 \texttt{raceth2}_i + \varepsilon_i, \quad i = 1, \ldots, 2020$$

with

$$\begin{aligned}
\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{African American}\right] &= \delta_0 + \delta_1 \\
\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{Other}\right] &= \delta_0 + \delta_2 \\
\mathrm{E}\left[\texttt{glucose}_i | \texttt{raceth}_i = \texttt{White}\right] &= \delta_0
\end{aligned}$$

The results of changing the categories are the following

```
> summary(modello2)
Call:
lm(formula = glucose ~ raceth, data = hers.nod)
...
 Coefficients:
                Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     96.5450        0.2266   425.979     <2e-16
 raceth1          0.5042        0.9103     0.554     0.5797
 raceth2          3.1473        1.3693     2.299     0.0216

Residual standard error:  9.738 on 2017 degrees of freedom

Multiple R-squared:  0.002717, Adjusted R-squared:  0.001728
F-statistic:  2.747 on 2 and 2017 DF, p-value:  0.06434
```

So we've seen:

- the estimates for the regression coefficients (table above) has changed (intercept is the estimate of the expected value for the reference group which is changed, same for the others): we note the difference of raceth2 which was highlighted by changing the reference category.
  This can happen in real life: if we have a significant F and nonsignificant t maybe switching the reference category helps finding the difference
- the global measures (last paragraph below) remains the same regardless the reference category chosen

4.1.3.2 Exclusion of the intercept

If one consider a regression model without intercept, it is possible to include all the 3 indicator variables (without choosing a reference category). The model fitted will be (again different symbols)

$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i\right] = \mu_1\texttt{African American}_i + \mu_2\texttt{Other}_i + \mu_3\texttt{White}_i \quad i = 1, \ldots, 2020$$

and regarding the interpretation of coefficients

$$\begin{aligned}
&\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{African American}\right] = \mu_1\\
&\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{Other}\right] = \mu_2\\
&\mathrm{E}\left[\texttt{glucose}_i|\texttt{raceth}_i = \texttt{White}\right] = \mu_3
\end{aligned}$$

In this setup the hypothesis we're interested is is

$$H_0 : \begin{cases}\mu_1 = \mu_2\\ \mu_1 = \mu_3\end{cases}$$

or simply $H_0 : \mu_1 = \mu_2 = \mu_3$, so here we don't need to set anything equal to 0.
The estimation without intercept is done by putting `-1` in the formula:

```
> summary(modello3)
Call:
lm(formula = glucose ~ raceth - 1, data = hers.nod)
...
 Coefficients:
                         Estimate   Std.  Error   t value   Pr(>|t|)
 racethAfrican American    97.0492        0.8816    110.08     <2e-16
 racethOther               99.6923        1.3504     73.83     <2e-16
 racethWhite               96.5450        0.2266    425.98     <2e-16

Residual standard error:  9.738 on 2017 degrees of freedom


Multiple R-squared:  0.99, Adjusted R-squared:  0.99
F-statistic:  6.634e+04 on 3 and 2017 DF, p-value:  < 2.2e-16
```

There are some **WARNING** in removing intercept in R: the t test is against a null of beta to be 0 which in this case (and often) is non interesting.
Removing intercept messes up things especially in the summary part

- In this setting the function `lm` computes R^2 using $\sum_{i=1}^n y_i^2$ as denominator instead of total variability $\sum_{i=1}^n (y_i - \mu_y)^2$
- the F test statistic is referred to the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$ where the last $= 0$ is not of interest/meaningless in our case and is easily rejected
  To compute the proper test we can rely on the general approach as follows

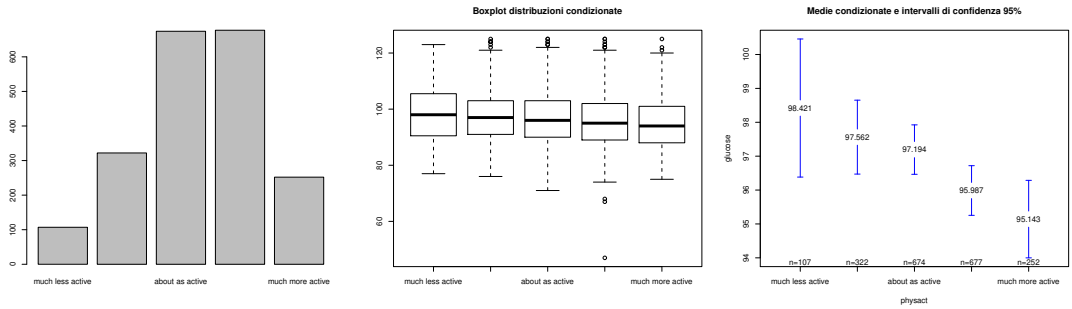
```
> K3
     1    2    3
 1   1   -1    0
 2   1    0   -1

> t3

[1] 0 0
```

With the previous K3 matrix we picked

$$\begin{bmatrix}1 & -1 & 0\\ 1 & 0 & -1\end{bmatrix}$$

in order to obtain

$$\begin{cases}\mu_1 = \mu_2\\ \mu_1 = \mu_2\end{cases}$$



(a) (b) (c)

Figure 4.2: Glucose level and physical activity

By the way we would get exactly the same by setting either one of the following matrices

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & -1 \\ -1 & 1 & 0 \end{bmatrix}$$

only changing the way the test is constructed not the results. eg the left matrix would be

$$\begin{cases} \mu_1 = \mu_2 \\ \mu_2 = \mu_3 \end{cases}$$

Going with the estimates we end up with the same results as the first model

```
> linearHypothesis(modello3,K3,t3,test="F")
Linear hypothesis test

Hypothesis:

racethAfrican American - racethOther = 0
racethAfrican American - racethWhite = 0
Model 1:  restricted model

Model 2:  glucose ~ raceth
     Res.Df      RSS   Df   Sum of Sq        F   Pr(>F)
 1     2019   191780
 2     2017   191259    2      521.02   2.7473   0.06434
```

4.2 Ordered Categories

4.2.1 Motivating example

Example 4.2.1 (Glucose level in blood and physical activity)**.** A glucose level in blood between 100 and 125 mg/dL represents a risk factor for developing diabetes.
Aim: Does physical activity (a modifiable factor related to life style) contribute to the reduction of the glucose level, thus preventing a severe disease?
Available information: Glucose level and physical activity level (`much less active`, `somewhat less active`, `about as active`, `somewhat more active`, `much more active`) on a sample of 2032 women not affected by diabetes after menopause.
The boxplots (fig 4.2) has more or less the same variability (strong overlap of distribution) and there's a decreasing trend of glucose mean level of as physical activity increases (maybe it's not that clinically relevant btw).

Remark 7. We can deal with this kind of data

- employing the same strategy of reference category seen for unordered data
- choosing a coding which acknowledge the ordering

4.2.2 Model with reference category

The estimated model with `much less active` as reference category

```
> physact1<-lm(glucose~physact,data=hers.nod)
> summary(physact1)
...
                              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)                    98.421         0.939   104.784      0.000
 physactsomewhat less active    -0.858         1.084    -0.792      0.429
 physactabout as active         -1.226         1.011    -1.213      0.225
 physactsomewhat more active    -2.434         1.011    -2.408      0.016
 physactmuch more active        -3.278         1.121    -2.924      0.003
...

Residual standard error:  9.716 on 2027 degrees of freedom

Multiple R-squared:  0.008668, Adjusted R-squared:  0.006712
F-statistic:  4.431 on 4 and 2027 DF, p-value:  0.001441
```

There seems to be a significant impact of the physical activity on glucose level. Looking both at the coefficient and their P-values it seems to be a scaletta.
To reproduce the F test

```
> K1
     1   2   3   4   5
 1   0   1   0   0   0
 2   0   0   1   0   0
 3   0   0   0   1   0
 4   0   0   0   0   1

> t1

[1] 0 0 0 0


   > linearHypothesis(physact1,K1,t1,test="F")
Linear hypothesis test

Hypothesis:

physactsomewhat less active = 0
physactabout as active = 0
physactsomewhat more active = 0
physactmuch more active = 0
Model 1:  restricted model

Model 2:  glucose ~ physact
     Res.Df         RSS   Df   Sum of Sq      F   Pr(>F)
 1     2031   193017.70
 2     2027   191344.61    4     1673.09   4.43   0.0014
```

Important remark 10. linear models in statistics, rencher, consigliato per le proprieta varie dei modelli

4.2.3 Model with incremental/split coding

Before we used the same coding as used in the unordered categorical groups; an alternative coding scheme can be used if there is a "natural" order among the categories such as in this case

	x_{Bi}	x_{Ci}	x_{Di}	x_{Ei}
much less active	0	0	0	0
somewhat less active	1	0	0	0
about as active	1	1	0	0
somewhat more active	1	1	1	0
much more active	1	1	1	1

The number of dummy is still 4 to represent 5 categories, but they're defined differently (the first dummy take value 0 for the first category and 1 for the other and so on).

It is possible to show that these alternative indicator variables can be obtained by linear combination (summing subsets) of the indicator variables introduced above.
It's called split coding because implicitly we split categories into two subsets.
What happens with such coding? the model has still 5 parameters ...

$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i\right] = \beta_0 + \beta_B x_{Bi} + \beta_C x_{Ci} + \beta_D x_{Di} + \beta_E x_{Ei} i = 1, \ldots, 2032$$

but their interpretation changes in the sense that each regression coefficient represents the difference between the conditional expected values associated with two consecutive categories

$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{much less active}\right] = \beta_0$$
$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{somewhat less active}\right] = \beta_0 + \beta_B$$
$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{about as active}\right] = \beta_0 + \beta_B + \beta_C$$
$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{somewhat more active}\right] = \beta_0 + \beta_B + \beta_C + \beta_D$$
$$\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{much more active}\right] = \beta_0 + \beta_B + \beta_C + \beta_D + \beta_E$$

When it comes to estimates ...

```
> physact2 <-lm(glucose~physact,data=hers.nod)
> summary(physact2)
...
              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     98.421         0.939   104.784      0.000
 physactd1       -0.858         1.084    -0.792      0.429
 physactd2       -0.368         0.658    -0.559      0.576
 physactd3       -1.208         0.529    -2.284      0.022
 physactd4       -0.844         0.717    -1.177      0.239
...

Residual standard error:  9.716 on 2027 degrees of freedom

Multiple R-squared:  0.008668, Adjusted R-squared:  0.006712
F-statistic:  4.431 on 4 and 2027 DF, p-value:  0.001441
```

In the interprestation:

- we have the same intercept as before (the expected value of the first category)
- the same happen for `physactd1` which is comparing the second group to the first one
- the remaining betas have different estimates because comparing to considered category to the previous one (instead of the first one).
- looking at the F test (all dummy variable = 0) we have the exact *same results* as the other coding scheme

In order to reproduce the F in the general context the matrix $\mathbf{K}$ is equal. The results will be the same as previously seen

```
> K2
     1   2   3   4   5
 1   0   1   0   0   0
 2   0   0   1   0   0
 3   0   0   0   1   0
 4   0   0   0   0   1

> t2

[1] 0 0 0 0

> linearHypothesis(physact2,K2,t2,test="F")
Linear hypothesis test

Hypothesis:

physactd1 = 0
physactd2 = 0
physactd3 = 0
physactd4 = 0
Model 1:  restricted model

Model 2:  glucose ~ physact
     Res.Df         RSS   Df   Sum of Sq      F   Pr(>F)
 1     2031   193017.70
 2     2027   191344.61    4     1673.09   4.43   0.0014
```

4.2.4 Linear trend hypothesis

What's the advantage of using the incremental coding? It matters if we're interested in some hypothesis in which the natural ordering of the categs is involved.
One of the typical hypothesis we could be interested is the so called linear trend hypothesis: it assumes that the change in conditional expected values given is constant, as we move from category to the next, no matter which consecutive couple of category we compare.
This is done by introduction of suitable linear constraints in the regression coefficients associated with the incremental coding scheme as

$$H_0 : \beta_B = \beta_C = \beta_D = \beta_E = \beta(\neq 0)$$

if $\beta > 0$ we'll have a constant increase in the conditional expected value or contrary for $\beta < 0$. By testing the hypothesis we check if we can replace the categories dummies with a numerical regressor taking the values 0 to 4 and by using a single coefficient β

$$\begin{aligned}
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{much less active}\right] | H_0 &= \beta_0 + 0 \cdot \beta = \beta_0 \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{somewhat less active}\right] | H_0 &= \beta_0 + 1 \cdot \beta \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{about as active}\right] | H_0 &= \beta_0 + 2 \cdot \beta \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{somewhat as active}\right] | H_0 &= \beta_0 + 3 \cdot \beta \\
\mathrm{E}\left[\texttt{glucose}_i|\texttt{physact}_i = \texttt{much more active}\right] | H_0 &= \beta_0 + 4 \cdot \beta
\end{aligned}$$

To implement this, after fitting the model, the linear constraints on coefficients in the general framework are (for four coefficient we need three equalities)

$$H_0 : \beta_B = \beta_C = \beta_D = \beta_E = \beta(\neq 0) \Longrightarrow H_0 : \begin{cases} \beta_B = \beta_C \\ \beta_C = \beta_D \\ \beta_D = \beta_E \end{cases}$$

which can be implemented as

```
> K.lin
     1   2   3   4   5
 1   0   1  -1   0   0
 2   0   0   1  -1   0
 3   0   0   0   1  -1

> t.lin

[1] 0 0 0
```

In the previous scheme we could have implemented three other constraints as well (obtainin same results) eg

$$\begin{cases} \beta_B = \beta_C \\ \beta_B = \beta_D \\ \beta_B = \beta_E \end{cases}$$

However In our case the results are as follows

```
> linearHypothesis(physact2,K.lin,t.lin,test="F")
Linear hypothesis test

Hypothesis:

physactd1 - physactd2 = 0
physactd2 - physactd3 = 0
physactd3 - physactd4 = 0
Model 1:  restricted model

Model 2:  glucose ~ physact
     Res.Df         RSS   Df   Sum of Sq      F   Pr(>F)
 1     2030   191419.47
 2     2027   191344.61    3       74.86   0.26    0.8511
```

So the linear trend hypothesis is not rejected (the fourth parameters -0.85, -0.36, -1.2, -0.84 are not significantly different from each other) and we can simplify the data by substituting numerical coding. There seems to be a constant difference in the expected value when we

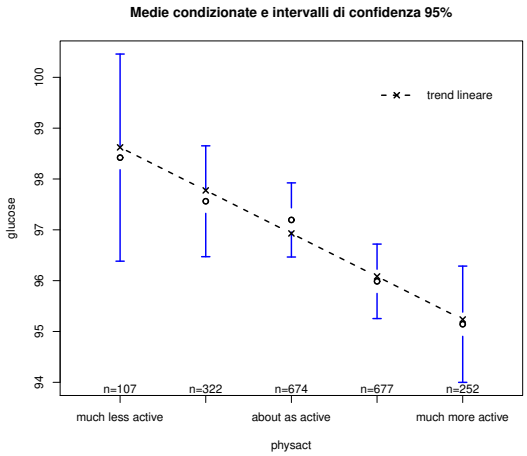


Figure 4.3: Estimated glucose conditional mean by activity level

compare consecutive pairs of groups.
The parameters of the constrained model can be estimated by coding the ordered categorical regressor using integer scores from 0 to 4 and by fitting a new Gaussian linear model

```
> hers.nod$physact.num<-as.numeric(hers.nod$physact)-1
> physact3<-lm(glucose physact.num,data=hers.nod)
> summary(physact3)
...
              Estimate   Std.  Error   t value   Pr(>|t|)
 (Intercept)     98.622         0.523   188.592      0.000
 physact.num     -0.847         0.206    -4.117      0.000
...

Residual standard error:  9.711 on 2030 degrees of freedom

Multiple R-squared:  0.00828, Adjusted R-squared:  0.007792
F-statistic:  16.95 on 1 and 2030 DF, p-value:  3.993e-05
```

As we can see the estimate -0.84 is basically a mean of the estimated coefficient of the model with categorical coding.
So the estimated conditional expected value is plotted in figure 4.3 where it overlap well with descriptive data.
We could get linear hypothesis test models using `anova` as well and comparing the two estimates (restricted and unrestricted models) - btw `> anova(physact3,physact2)`

```
Analysis of Variance Table

Model 1:  glucose ~ physact.num

Model 2:  glucose ~ physact
     Res.Df         RSS   Df   Sum of Sq      F   Pr(>F)
 1     2030   191419.47
 2     2027   191344.61    3       74.86   0.26    0.8511
```

So in general to obtain the test we need either:

- the starting model and the set of restrictions with `linearHypothesis`;
- the starting model and reduced model with `anova`.

If fitting the constrainted model is trivial (eg remove some regressors) we can go with `anova`, otoh more complex hypotheses/constrainted models can be tackled with `linearHypothesis`