NMSA407 Linear Regression

Lecture Notes

Michal Kulich

Last modified on January 2, 2024.

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University

These lecture notes contain the whole contents of the course “NMSA407 Linear Regression”, a compulsory course of the curriculum in the Master’s program “Probability, Mathematical Statistics and Econometrics” at the Faculty of Mathematics and Physics, Charles University.

This document undergoes continuing development. The author will appreciate notifications by the reader of potential typos or misprints.

Michal Kulich
kulich@karlin.mff.cuni.cz

In Karlín on January 2, 2024

Contents

1. Simple Linear Regression: Technical and Historical Review

Consider n measurements of continuous variables (x_i, y_i) for $i = 1, \ldots, n$. Plot them as Carthesian coordinates on a scatterplot (Figure 1.1). The observations seem to be located along a line; there is a perceived linear relationship between the values of x and y, but not an exact one. The goal is to identify a line passing through the observations (see Figure 1.2) so that the line is "optimal" in some way.

Legendre (1805) proposed to find the line by minimizing the sum of squared vertical distances of the observed points from the fitted line (see Figure 1.3). This is called *the least squares method*.[*] It can be also attributed to Gauss, who later claimed (Gauss 1821) that he had been using the method as early as in 1795 but had not published it.

> ***Adrien-Marie Legendre*** *(1752 – 1833) was a French mathematician who made numerous contributions to mathematics. Well-known and important concepts such as the Legendre polynomials and Legendre transformation are named after him.*
> *Source:* `https://en.wikipedia.org/wiki/Adrien-Marie_Legendre`

[*] Česky *Metoda nejmenších čtverců.*

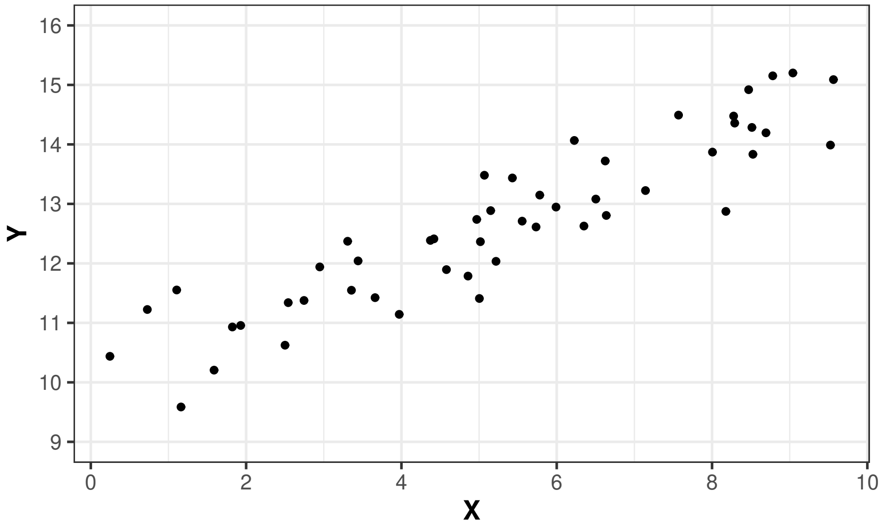


Figure 1.1.: Scatterplot of two continuous variables in $\mathbb{R}^2$.

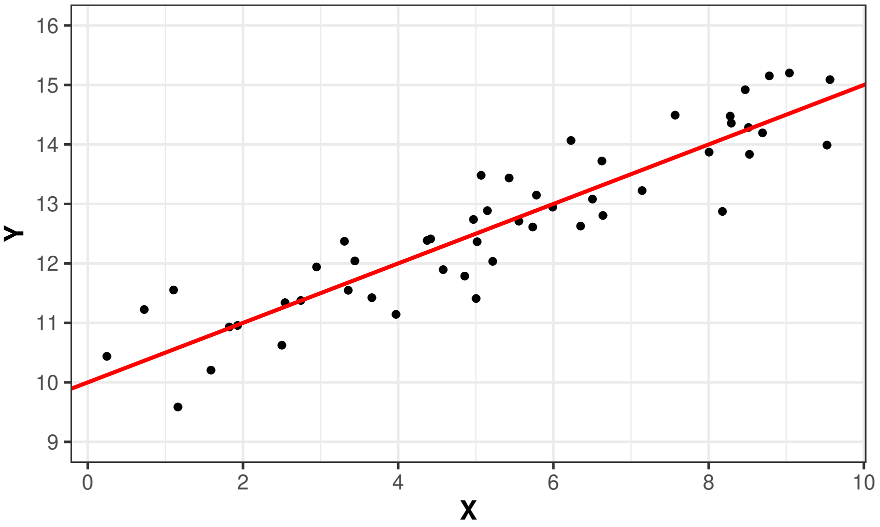


Figure 1.2.: Scatterplot of two continuous variables in $\mathbb{R}^2$ with fitted line.

The least squares method is based on the presumption that the observed values of the variable x_i are measured precisely while y_i are measured with an error that shifts them away from the line that expresses the linear relationship between the two variables. This point of view justifies the minimization of vertical distances instead of e.g. perpendicular distances.

Johann Carl Friedrich Gauss *(1777 – 1855) was a German mathematician, geodesist, and physicist who made significant contributions to many fields in mathematics and science. Gauss published the second and third complete proofs of the fundamental theorem of algebra, made important contributions to number theory and developed the theories of binary and ternary quadratic forms. He is also credited with inventing the fast Fourier transform algorithm and was instrumental in the discovery of the dwarf planet Ceres. His work on the motion of planetoids disturbed by large planets led to the introduction of the Gaussian gravitational constant and the method of least squares, which is still used in all sciences to minimize measurement error.*
Source: https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

Let us show how the idea of Legendre and Gauss works. Consider a line $y = a + bx$ and choose a, b so that

$$SS(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{1.1}$$

is minimized over all $a, b \in \mathbb{R}$. The sum in the expression (1.1) is called *the sum of squares*.[*]

[*] Česky *Součet čtverců*.

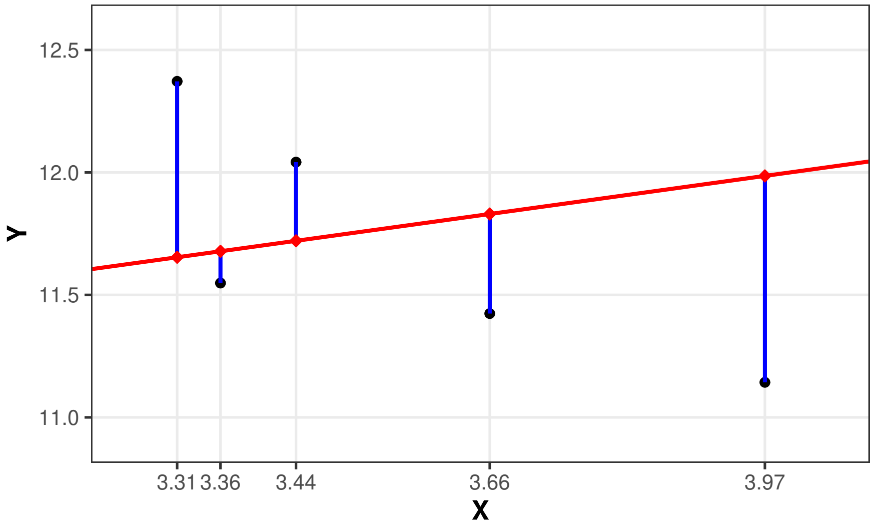


Figure 1.3.: Zoomed subset of data from Figure 1.2 with visualized vertical distances of the points from the line (blue).

The values a, b that minimize the sum of squares are easy to find:

$$\frac{\partial SS(a,b)}{\partial a} = 2\sum_{i=1}^{n}(y_i - a - bx_i)(-1),$$
$$\frac{\partial SS(a,b)}{\partial b} = 2\sum_{i=1}^{n}(y_i - a - bx_i)(-x_i).$$

Thus, a and b are the solutions to the system of two equations

$$\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0,$$
$$\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0.$$

These equations are called the *normal equations*[*].

Introducing the notation $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the normal equations can be solved as follows. From the first equation, we get

$$na = \sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i, \quad \text{hence} \quad a = \overline{y} - b\overline{x}.$$

This shows that the fitted line passes through the point $(\overline{x}, \overline{y})$. Next, substituting in the

[*] Česky *Normální rovnice.*

second equation for the optimal intercept a, we get

$$\begin{aligned} b\frac{1}{n}\sum_{i=1}^{n} x_i^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - a\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} + b\overline{x}^2 \\ b\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right) &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} \\ b\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \end{aligned}$$

Finally,

$$b = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

The former version is more computationally friendly, the latter version provides an insight into the meaning of the slope b. Indeed,

$$b = \frac{\widehat{\mathsf{cov}}(x, y)}{\widehat{\mathsf{var}}(x)} = r_{xy}\sqrt{\frac{\widehat{\mathsf{var}}(y)}{\widehat{\mathsf{var}}(x)}},$$

where $\widehat{\mathsf{cov}}(x, y)$ is the sample covariance of the observations (x_i, y_i), $\widehat{\mathsf{var}}(x)$ is the sample variance of x_i, $\widehat{\mathsf{var}}(y)$ is the sample variance of y_i, and r_{xy} is the sample correlation coefficient of the observations (x_i, y_i).

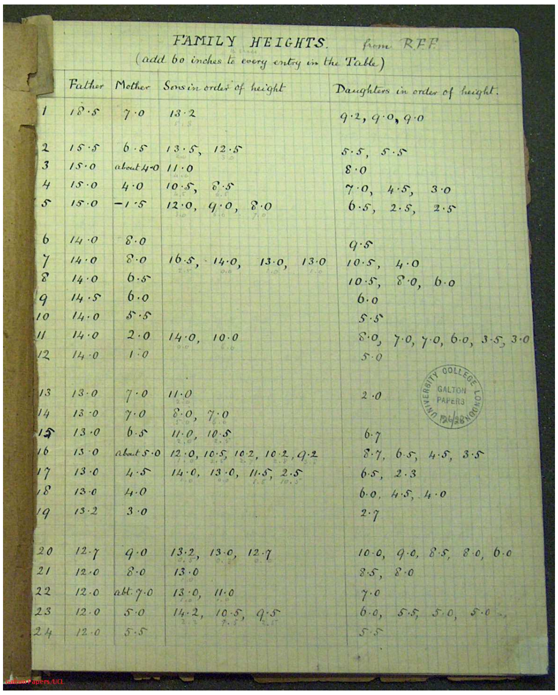
If the observations x_i have the same sample variance as y_i then the slope of the line fitted by least squares is equal to the sample correlation coefficient r_{xy} and therefore lies in the interval $\langle -1, 1\rangle$.

This phenomenon was noticed by sir Francis Galton (Galton 1886). He investigated the relationship of the parents' height with the height of their grown children. The recorded heights (in inches) are shown in Figure 1.4 and Galton's original visualization of the data in Figure 1.5. If we focus on the heights of sons only (to eliminate the fact that daughters are somewhat shorter) and plot them as y_i against the average height of their parents (x_i) we obtain the scatterplot shown in Figure 1.6.

Sir Francis Galton *(1822 – 1911) Darwin's cousin, prodigy child, contributor to the fields of statistics, meteorology, psychology, genetics, co-founder and proponent of eugenics.*
Source: `https://en.wikipedia.org/wiki/Francis_Galton`

The red line in Figure 1.6 was fitted by the method of least squares and its slope is about 0.74.* As explained above, this value corresponds to the sample correlation between the average height of the parents and the height of their son. It means that if the average height of the parents exceeds the population mean by 10 cm the son's height is likely to be above average as well, but only by some 7.4 cm. So, tall parents tend to have tall sons, but

* Galton used a different data set and estimated the slope of the fitted line to be about 0.66.



FAMILY HEIGHTS. from RFF
(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5
6	14·0	8·0		9·5
7	14·0	8·0	16·5, 14·0, 13·0, 13·0	10·5, 4·0
8	14·0	6·5		10·5, 8·0, 6·0
9	14·5	6·0		6·0
10	14·0	5·5		5·5
11	14·0	2·0	14·0, 10·0	8·0, 7·0, 7·0, 6·0, 3·5, 3·0
12	14·0	1·0		5·0
13	13·0	7·0	11·0	2·0
14	13·0	7·0	8·0, 7·0	
15	13·0	6·5	11·0, 10·5	6·7
16	13·0	about 5·0	12·0, 10·5, 10·2, 10·2, 9·2	8·7, 6·5, 4·5, 3·5
17	13·0	4·5	14·0, 13·0, 11·5, 2·5	6·5, 2·3
18	13·0	4·0		6·0, 4·5, 4·0
19	13·2	3·0		2·7
20	12·7	9·0	13·2, 13·0, 12·7	10·0, 9·0, 8·5, 8·0, 6·0
21	12·0	8·0	13·0	8·5, 8·0
22	12·0	abt. 7·0	13·0, 11·0	7·0
23	12·0	5·0	14·2, 10·5, 9·5	6·0, 5·5, 5·0, 5·0
24	12·0	5·5		5·5

Figure 1.4.: Galton height data: original pen/paper records.
Source: http://www.medicine.mcgill.ca/epidemiology/hanley/galton/

not as tall as the parents were. Galton called this feature *regression towards the mean*. Even though the term *regression*[*] originally referred to this very specific feature that appears only in certain data sets, it began to be used more generally to describe methods and techniques used for fitting lines or curves to observed data.

The least squares method can be easily extended to fit certain non-linear relationships between the two variables. For example, if the relationship is not linear but quadratic we could use the same idea with the function

$$y_i = a + bx_i + cx_i^2.$$

We could find a, b, and c by the method of least squares by minimizing

$$SS(a,b,c) = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2.$$

The estimated parameters a, b, and c are obtained by solving a system of three linear equations.

In this introductory chapter, we approached the problem of fitting a line or a curve through a cloud of bivariate data. We did not introduce any underlying probabilistic model

[*] Česky *Regrese*.

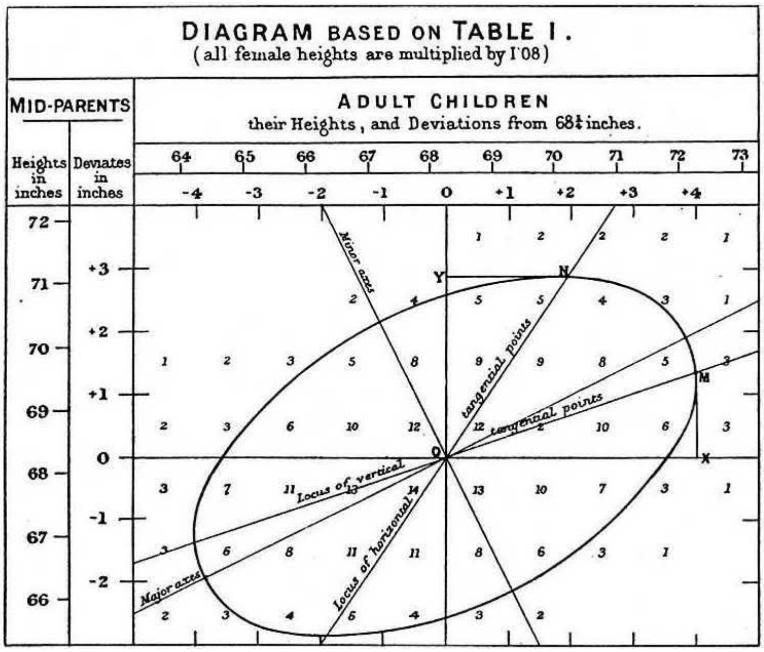


Figure 1.5.: Galton height data: original visualization by the author.
Source: https://en.wikipedia.org

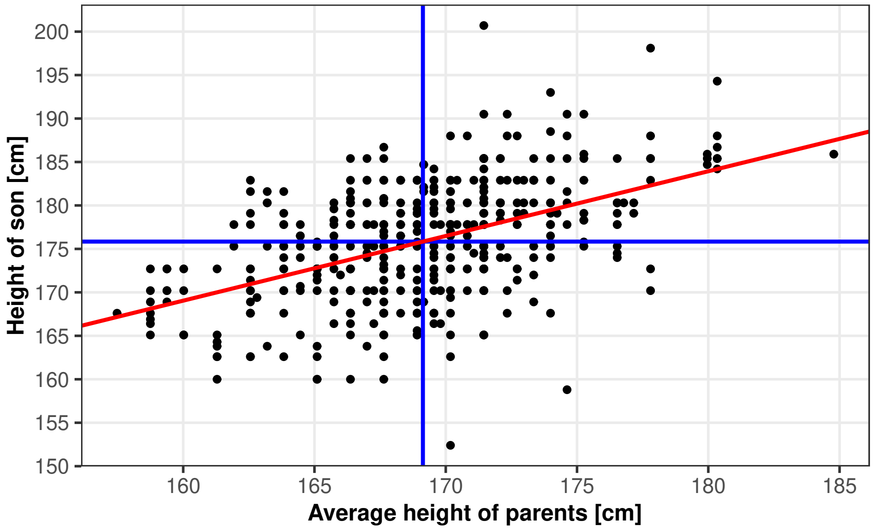


Figure 1.6.: Modified Galton data with fitted least squares line (red). The slope of the line is ≈ 0.74. The means of the two variables are plotted as blue lines.

for the data, did not formulate any assumptions and were not able to find neither an interpretation for the estimates obtained by the least square method nor to investigate their theoretical properties.

2. Linear Regression Model

In this chapter, we formulate a general definition of the linear regression model. We explain the meaning of the regression parameters and derive a general formula for the least squares estimator. We introduce a lot of new technical terms, explain their meaning and investigate some features of linear regression models that will be important for the developments presented in subsequent chapters.

2.1. Definition and Assumptions

Consider a sequence of n independent random vectors $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. The random variable Y_i is called *the response*[*] (also *the dependent variable*[†], *the outcome*). The random vector $\boldsymbol{X}_i$ contains $p < n$ components $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ which are called *the covariates* (also explanatory variables, predictors, regressors)[‡].

Definition 2.1. The independent observations $(Y_i, \boldsymbol{X}_i)$ satisfy *the linear regression model* if the response Y_i can be written as $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$, that is,

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)^\top$ is a vector of unknown *regression parameters (coefficients)*[§] and *the error terms*[¶] $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables such that $\mathsf{E}\left[\varepsilon_i \,\middle|\, \boldsymbol{X}_i\right] = 0$, and $\mathsf{var}\left[\varepsilon_i \,\middle|\, \boldsymbol{X}_i\right] = \sigma_e^2$. ∇

Note. On the covariates:

- The first covariate X_{i1} is usually taken as 1.
- The covariates $\boldsymbol{X}_i$ are often created by a transformation of an originally observed random vector $\boldsymbol{Z}_i$. We suppress this in the notation.
- In certain applications, the covariates are fixed quantities rather than random variables. Because the definition of the linear model only specifies conditional moments given the observed values of the covariates it applies to fixed covariates as well. Most of the developments that follow in this course are not sensitive to differences between fixed and random covariates either. The only occasion when fixed covariates need to be treated differently than random covariates is the investigation of asymptotic properties. This will be discussed in Section **??**.

[*] Česky *odezva* [†] Česky *závislá proměnná* [‡] Česky *regresory, nezávisle proměnné, vysvětlující veličiny, prediktory, kovariáty* [§] Česky *regresní koeficienty* [¶] Česky *chybové členy*

Note. On the error terms:

- The random variables ε_i are required to have zero means and equal variances. It is somewhat misleading to call them *error terms* because they include not only errors in the measurement of the response but also the effects of any factors that influence the mean of the response and are not included in the model. In econometrics, the error terms are often called *disturbances*.
- The variance σ_e^2 of the error terms is called *the residual variance*[*].
- Sometimes, the assumptions on the error terms are strengthened to require that ε_i be independent of $\boldsymbol{X}_i$. Our definition does not require this.

The definition of the linear model can be reformulated in terms of conditional moments of the response as follows:

- $\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$,
- $\mathsf{var}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \sigma_e^2$.

Thus, the model makes assumptions about the first two conditional moments of the response: the conditional mean must be linear in $\boldsymbol{X}_i$ through $\boldsymbol{\beta}$ and the conditional variance must be constant.

The purpose of the linear regression model is not just to fit a line, curve or surface through a cloud of data as it was presented in Chapter 1. Instead, we aim to express how the expected value of the response Y_i changes with different values of $\boldsymbol{X}_i$ and tell what influence the individual covariates have on the expectation.

Notation. Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\mathsf{T} \\ \boldsymbol{X}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{X}_n^\mathsf{T} \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The n by p matrix $\mathbb{X}$ is called *the regression matrix*[†]. It includes the observed covariate vectors in the rows.

Now we can express the model for all the data together

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\mathsf{E}\left[\boldsymbol{\varepsilon} \,\middle|\, \mathbb{X}\right] = \boldsymbol{0}$ and $\mathsf{var}\left[\boldsymbol{\varepsilon} \,\middle|\, \mathbb{X}\right] = \sigma_e^2 \mathbb{I}_n$ or

- $\mathsf{E}\left[\boldsymbol{Y} \,\middle|\, \mathbb{X}\right] = \mathbb{X}\boldsymbol{\beta}$,
- $\mathsf{var}\left[\boldsymbol{Y} \,\middle|\, \mathbb{X}\right] = \sigma_e^2 \mathbb{I}_n$.

Note. From now on, we will often use the notation E, var for the conditional expectation and variance given the covariates. So, we will write $\mathsf{E}\, Y_i$ instead of $\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right]$ and $\mathsf{var}\, Y_i$ instead of $\mathsf{var}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right]$; similarly for $\mathsf{E}\,\varepsilon_i$, $\mathsf{var}\,\varepsilon_i$, $\mathsf{E}\,\boldsymbol{Y}$, $\mathsf{var}\,\boldsymbol{Y}$ etc.

[*] Česky *residuální rozptyl* [†] Česky *regresní matice*

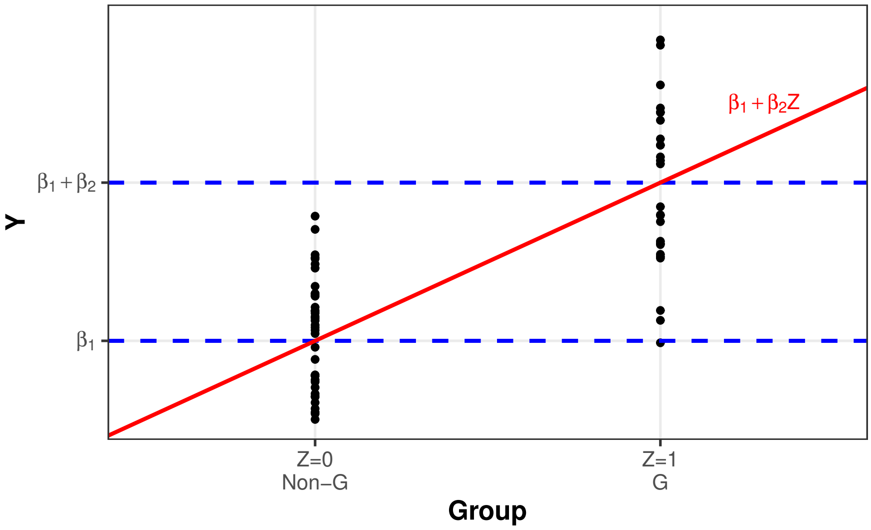


Figure 2.1.: Two sample problem expressed as a linear regression model $\mathsf{E}\,Y = \beta_1 + \beta_2 Z$, where $Z = \mathbb{1}(G)$. The regression line has no interpretation except at $Z = 0$ or $Z = 1$.

Example 2.1 (Linear model for iid data). Suppose the responses $Y_1, \ldots, Y_n$ represent a random sample of independent identically distributed random variables with $\mathsf{E}\,Y_i = \mu$ and $\operatorname{var} Y_i = \sigma^2$. Then

$$Y_i = \mu + \varepsilon_i,$$

where ε_i, $i = 1, \ldots, n$ are iid with zero mean and variance σ^2. Thus, Y_i satisfies a linear regression model with $\boldsymbol{X}_i = 1$, $\boldsymbol{\beta} = \mu$ and $\sigma_e^2 = \sigma^2$. △

Example 2.2 (Simple linear regression). Suppose we observe a random sample of (Y_i, Z_i), where Z_i is univariate. Define the covariate vector as $\boldsymbol{X}_i = (1, Z_i)^\mathsf{T}$. This leads to the regression matrix

$$\mathbb{X} = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ \vdots & \vdots \\ 1 & Z_n \end{pmatrix},$$

and the simple linear regression model (recall Chapter 1)

$$Y_i = \beta_1 + \beta_2 Z_i + \varepsilon_i,$$

with $\mathsf{E}\,Y_i = \beta_1 + \beta_2 Z_i$ and $\operatorname{var} Y_i = \sigma_e^2$. △

Example 2.3 (Two sample problem). In the previous example, take a special case with a binary covariate Z_i, which attains only values 0 or 1. Suppose that Z_i indicates a membership of the observation in some subgroup G, that is $Z_i = \mathbb{1}(i \in G)$.

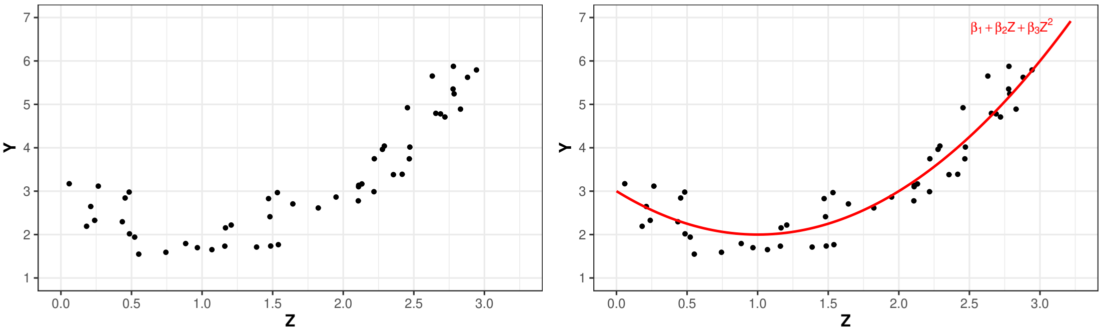


Figure 2.2.: Data following a quadratic association with a fitted quadratic curve.

The simple linear regression model has the form

$$Y_i = \beta_1 + \beta_2 \mathbb{1}(i \in G) + \varepsilon_i,$$

that is,

$$\mathsf{E}\, Y_i = \begin{cases} \beta_1 & \text{when } i \notin G, \\ \beta_1 + \beta_2 & \text{when } i \in G, \end{cases} \qquad \mathsf{var}\, Y_i = \sigma_e^2.$$

This model specifies a two-sample location problem with equal variances in both groups and possibly different expectations. The regression parameter β_2 expresses the difference in expectations between the groups.

An illustration of the two-sample location problem is provided by Figure 2.1. The regression line is shown in red color but realize that it can only be interpreted at points that actually appear in the data, that is $Z = 1$ (group G) or $Z = 0$ (group $\neg G$). △

The end of lecture 1 (Oct 3, 2023)

Example 2.4 (Quadratic regression). Suppose we observe a random sample of (Y_i, Z_i), where Z_i is univariate. Define the covariate vector as $\boldsymbol{X}_i = (1, Z_i, Z_i^2)^\mathsf{T}$. This leads to the regression matrix

$$\mathbb{X} = \begin{pmatrix} 1 & Z_1 & Z_1^2 \\ 1 & Z_2 & Z_2^2 \\ \vdots & \vdots & \vdots \\ 1 & Z_n & Z_n^2 \end{pmatrix},$$

and the quadratic regression model (recall Chapter 1)

$$Y_i = \beta_1 + \beta_2 Z_i + \beta_3 Z_i^2 + \varepsilon_i,$$

with $\mathsf{E}\, Y_i = \beta_1 + \beta_2 Z_i + \beta_3 Z_i^2$ (a quadratic function of Z_i) and $\mathsf{var}\, Y_i = \sigma_e^2$.

An illustration of the quadratic regression model is provided by Figure 2.2. △

2.2. Interpretation of Regression Coefficients

Recall how the regression coefficients are related to the expectation of the response:

$$\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = (x_{i1},\dots,x_{ip})\right] = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

Thus, the regression coefficients capture and express the influence of $\boldsymbol{X}_i$ on $\mathsf{E}\,Y_i$.

Suppose that $X_{i1} = 1\ \forall i \in \{1,\dots,n\}$. Then the coefficient pertaining to this covariate is called *the intercept* (or *the absolute term*[*]). Obviously,

$$\beta_1 = \mathsf{E}\left[Y_i \,\middle|\, X_{i2} = 0, X_{i3} = 0, \dots, X_{ip} = 0\right].$$

The intercept provides the expectation of the response for an observation with zero values of all covariates (except the first).

Next, take an observation with any covariate vector $\boldsymbol{x} = (1, x_2, \dots, x_p)$ and denote the j-th unit vector of dimension p by $\boldsymbol{e}_j = (0,\dots,0,1,0,\dots,0)^\mathsf{T}$ with 1 at the j-th position ($j = 2,\dots,p$). We have

$$\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = \boldsymbol{x}\right] = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

and

$$\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = \boldsymbol{x} + \boldsymbol{e}_j\right] = \beta_1 + \beta_2 x_2 + \cdots + \beta_j(x_j + 1) + \ldots + \beta_p x_p.$$

After subtracting the top equation from the bottom one, we get

$$\beta_j = \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = \boldsymbol{x} + \boldsymbol{e}_j\right] - \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = \boldsymbol{x}\right], \quad j = 2,\dots,p.$$

So, β_j **expresses the increase in** EY_i **after the** j**-th covariate is increased by one unit and all other covariates stay the same.**[†]

It is important to realize that these interpretations do not always make sense.
Obviously, the intercept cannot be interpreted if an observation with all covariates equal to zero does not exist.
In quadratic regression $\mathsf{E}\left[Y_i \,\middle|\, Z_i\right] = \beta_1 + \beta_2 Z_i + \beta_3 Z_i^2$, with $X_{i2} = Z_i$ and $X_{i3} = Z_i^2$, one cannot increase X_{i2} by a single unit while keeping X_{i3} the same and vice versa. So, β_2 and β_3 cannot be interpreted either. This is because in this model a single variable Z_i affects the values of several covariates simultaneously.
Another cautionary note applies to interpretation of the absolute value of β_j. It is not true that a covariate with a very large value of β_j affects the response more strongly than a covariate with a parameter close to zero. The strength of the influence of the covariate also depends on the units of measurement. By rescaling all values of X_{ij} to mX_{ij}, the coefficient β_j is made m-times smaller because $\beta_j X_{ij} = \frac{\beta_j}{m} \cdot mX_{ij}$. Thus, rescaling a measurement made in kilometers into meters makes the regression coefficient 1000 times smaller without changing anything about the strength of the influence of that covariate on the response.

[*] Česky *absolutní člen*

[†] Of course, if $\beta_j < 0$, it expresses a decrease in the expectation.

2.3. Least Squares Estimation

Definition of the least squares estimator

Consider the model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ and $\mathsf{var}\,\boldsymbol{\varepsilon} = \sigma_e^2\mathbb{I}_n$. The regression matrix $\mathbb{X}$ has n rows and p columns, with $p < n$, and the dimension of $\boldsymbol{\beta}$ is p.

Definition 2.2 (Least Squares Estimator). The *the least squares estimator* (LSE) $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the point in $\mathbb{R}^p$ that minimizes the sum of squares

$$SS_e(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2.$$

∇

In order to make the LSE unique, we will make the following assumption.

Assumption. Let the regression matrix $\mathbb{X}_{n\times p}$ be of full rank, that is, $r(\mathbb{X}) = p$.

If the regression matrix did not have full rank there would exist at least one covariate (a column of $\mathbb{X}$) that can be expressed as a linear combination of other covariates. Under such circumstances the regression coefficients are not identifiable and the LSE $\widehat{\boldsymbol{\beta}}$ does not have a unique value.

Example 2.5. Consider the model $\mathsf{E}\,Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$ and suppose that $X_4 = X_2 + X_3$. Then there are infinitely many values of $\boldsymbol{\beta}$ that always generate the same expectation for the response:

$$\begin{aligned}
\mathsf{E}\,Y &= \beta_1 + \beta_2X_2 + \beta_3X_3 + \beta_4(X_2 + X_3) = & \boldsymbol{\beta} &= (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathsf{T}}\\
&= \beta_1 + (\beta_2 + \beta_4)X_2 + (\beta_3 + \beta_4)X_3 = & \boldsymbol{\beta} &= (\beta_1, \beta_2 + \beta_4, \beta_3 + \beta_4, 0)^{\mathsf{T}}\\
&= \beta_1 + \big(\beta_2 + \frac{\beta_4}{2}\big)X_2 + \big(\beta_3 + \frac{\beta_4}{2}\big)X_3 + \frac{\beta_4}{2}(X_2 + X_3) & \boldsymbol{\beta} &= \big(\beta_1, \beta_2 + \frac{\beta_4}{2}, \beta_3 + \frac{\beta_4}{2}, \frac{\beta_4}{2}\big)^{\mathsf{T}}
\end{aligned}$$

et cetera. When the regression coefficients $\boldsymbol{\beta}$ do not have a unique value the model is called *unidentifiable*[*]. $\triangle$

Through the entire course, **we will avoid regression matrices that are not of full rank**. It makes little sense to deal with them because such models cannot be used in practice. We can always satisfy our assumption by dropping the columns that can be expressed as linear combination of other columns and so reducing the dimension of the model and the number of parameters p until the regression matrix has a full rank.

[*] Česky *neidentifikovatelný*

Note. One could raise an objection that we consider $\mathbb{X}$ random and hence its rank is also a random variable. The following simple example shows that it is possible to end up with a singular regression matrix by mere bad luck.

Suppose $\mathsf{E}\, Y_i = \beta_1 + \beta_2 X_i$ where $X_i \in \{0, 1\}$ is an indicator of membership of the individual in some subgroup $\mathscr{G}$. The rank of the regression matrix should be equal to $p = 2$. Let $\mathsf{P}\left[X_i = 1\right] \equiv \pi \in (0, 1)$. If $\pi = 0$ or $\pi = 1$, the covariate generates the same value for all observations and the regression matrix is of rank 1. But even if we exclude these cases by requiring $\pi \in (0, 1)$, we still get

$$\begin{aligned}\mathsf{P}\left[X_i = 1\ \forall i \in \{1, \dots, n\}\right] &= \pi^n > 0\\ \mathsf{P}\left[X_i = 0\ \forall i \in \{1, \dots, n\}\right] &= (1-\pi)^n > 0,\end{aligned}$$

so for any finite sample size n there is a positive probability of $r(\mathbb{X}) = 1$. The probability, however, converges to zero fairly quickly as n increases.

If it happens in practice, it means that either the group $\mathscr{G}$ or the complement $\mathscr{G}^{\mathscr{C}}$ are not represented in the data at all and we cannot estimate the effect of the group on the expectation of the response. We have no choice but to drop the indicator of the group from the model and reduce the number of columns of the regression matrix.

Note. In the general case, express $\boldsymbol{X}_i = (1, \boldsymbol{X}_i^M)$ (separate the intercept from the rest of the covariates). Then it holds: If $\mathsf{var}\, \boldsymbol{X}_i^M > 0$ then $\mathsf{P}\left[r(\mathbb{X}) = p\right] \to 1$ as $n \to \infty$.

Derivation of the explicit form of the LSE

Let us derive the explicit form of the least squares estimator. Decompose $SS_e(\boldsymbol{\beta})$ into several parts.

$$SS_e(\boldsymbol{\beta}) = (\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta}) = \boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y}-\boldsymbol{\beta}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\boldsymbol{Y}-\boldsymbol{Y}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\beta}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y}-2\boldsymbol{\beta}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\boldsymbol{Y}+\boldsymbol{\beta}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta}.$$

We will use rules for matrix differentiation. In particular, for any vector $\boldsymbol{c}$ and any symmetric matrix $\mathbb{A}$

$$\frac{\partial \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{c}}{\partial \boldsymbol{\beta}} = \boldsymbol{c} \quad \text{and} \quad \frac{\partial \boldsymbol{\beta}^{\mathsf{T}}\mathbb{A}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\mathbb{A}\boldsymbol{\beta}.$$

We have,

$$\frac{\partial \boldsymbol{\beta}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\boldsymbol{Y}}{\partial \boldsymbol{\beta}} = \mathbb{X}^{\mathsf{T}}\boldsymbol{Y} \quad \text{and} \quad \frac{\partial \boldsymbol{\beta}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\mathbb{X}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta},$$

and hence

$$\frac{\partial SS_e(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbb{X}^{\mathsf{T}}\boldsymbol{Y} + 2\mathbb{X}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta}.$$

The LSE $\widehat{\boldsymbol{\beta}}$ solves the system of p linear equations

$$\mathbb{X}^{\mathsf{T}}\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^{\mathsf{T}}\boldsymbol{Y}, \tag{2.1}$$

which is called *the normal equations*[*] in this context.

When $\mathbb{X}$ is of rank p, as we assume, $\mathbb{X}^\top\mathbb{X}$ is a $p \times p$ matrix of rank p and therefore its inverse exists and is unique. It follows that the normal equations have a single solution, which is

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{Y}. \tag{2.2}$$

This is the explicit form of the least squares estimator in linear regression.

To show that this estimator really minimizes the least squares criterion, we calculate the Hessian matrix:

$$\frac{\partial}{\partial\boldsymbol{\beta}^\top}\frac{\partial SS_e(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{\partial}{\partial\boldsymbol{\beta}^\top}\big(-2\mathbb{X}^\top\boldsymbol{Y} + 2\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta}\big) = 2\mathbb{X}^\top\mathbb{X},$$

which is a positive definite matrix at any argument $\boldsymbol{\beta} \in \mathbb{R}^p$. Thus, the function $SS_e(\boldsymbol{\beta})$ is strictly convex and we have found its global minimum.

Alternative verification that $\widehat{\boldsymbol{\beta}}$ is the LSE

There is another way how to verify that the solution $\widehat{\boldsymbol{\beta}}$ to the system of normal equations (2.1) is the LSE. Take any $\boldsymbol{\beta} \in \mathbb{R}^p$ and write

$$\begin{aligned} SS_e(\boldsymbol{\beta}) &= \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}} + \mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\boldsymbol{\beta}\|^2 \\ &= \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 + 2(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\top\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \end{aligned}$$

where the last term is zero because

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbb{X}^\top(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\mathbb{X}^\top\boldsymbol{Y} - \mathbb{X}^\top\mathbb{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}$$

using the fact that $\widehat{\boldsymbol{\beta}}$ solves the normal equations $\mathbb{X}^\top\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\top\boldsymbol{Y}$.

Hence, at any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$SS_e(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \geq \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|^2 = SS_e(\widehat{\boldsymbol{\beta}})$$

and equality is attained if and only if

$$\|\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbb{X}^\top\mathbb{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{0}.$$

When $\mathbb{X}^\top\mathbb{X}$ is of full rank, this is equivalent to $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Thus, $\widehat{\boldsymbol{\beta}}$ is the unique minimizer of $SS_e(\boldsymbol{\beta})$.

Fitted values and residuals

Definition 2.3 (Fitted values, residuals).

[*] Česky *normální rovnice*

(a) $\widehat{\boldsymbol{Y}} \equiv \mathbb{X}\widehat{\boldsymbol{\beta}}$ are called *the fitted values*[*].

(b) $\boldsymbol{u} \equiv \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}$ are called *the residuals*[†].

Recall the definition of the linear regression model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbb{X}\boldsymbol{\beta}$ is the conditional mean of $\boldsymbol{Y}$ given the covariates and $\boldsymbol{\varepsilon}$ is random noise, and compare it with the decomposition

$$\boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{u},$$

where the fitted values $\mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{Y}}$ represent the estimated mean of $\boldsymbol{Y}$ and the residuals $\boldsymbol{u}$ represent the estimated noise. The fitted values are the “best” approximations (or predictions) of the responses that can be calculated from the covariates alone.

We can write $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = \mathbb{H}\boldsymbol{Y}$, where $\mathbb{H} \equiv \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ is a square $n \times n$ matrix. The matrix $\mathbb{H}$ is called *the hat matrix*[‡] . It is symmetric, $r(\mathbb{H}) = p$ because $r(\mathbb{X}) = p$, and it is idempotent:

$$\mathbb{H}\mathbb{H} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\,\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T} = \mathbb{H}.$$

Recall that any idempotent matrix satisfies $r(\mathbb{H}) = \operatorname{tr}(\mathbb{H})$.

Throughout the whole course, we will frequently use the following trivial identities:

$$\mathbb{H}\mathbb{X} = \mathbb{X}, \quad (\mathbb{I} - \mathbb{H})\mathbb{X} = \mathbf{0}.$$

The end of lecture 2 (Oct. 6, 2023)

The main linear properties of fitted values and residuals are summarized in the following note.

Note.

(a) $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$ where $\mathbb{H} \equiv \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ is a symmetric, idempotent $n \times n$ matrix of rank p.

(b) $\boldsymbol{u} = (\mathbb{I} - \mathbb{H})\boldsymbol{Y}$ where $\mathbb{I} - \mathbb{H}$ is a symmetric, idempotent $n \times n$ matrix of rank $n - p$. Also, $\boldsymbol{u} = (\mathbb{I} - \mathbb{H})\boldsymbol{\varepsilon}$.

(c) $\widehat{\boldsymbol{Y}}$, $\boldsymbol{u}$, and $\widehat{\boldsymbol{\beta}}$ are all linear transformations of bY.

(d) $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{u}$ are always orthogonal.

Parts (a) and (c) of the note are trivial or have been proven above. As for part (b), $(\mathbb{I} - \mathbb{H})(\mathbb{I} - \mathbb{H}) = \mathbb{I} - 2\mathbb{H} + \mathbb{H}\mathbb{H} = \mathbb{I} - \mathbb{H}$, so $(\mathbb{I} - \mathbb{H})$ is indeed idempotent. Its rank can be calculated using $r(\mathbb{A}) = \operatorname{tr}(\mathbb{A})$ for any idempotent $\mathbb{A}$:

$$r(\mathbb{I} - \mathbb{H}) = \operatorname{tr}(\mathbb{I} - \mathbb{H}) = \operatorname{tr}(\mathbb{I}) - \operatorname{tr}(\mathbb{H}) = n - r(\mathbb{H}) = n - p. \tag{2.3}$$

[*] Česky *vyrovnané hodnoty* [†] Česky *residua (sing. residuum)* [‡] Česky *nemá český ekvivalent*

Finally, using the definition of the linear model and $(\mathbb{I}-\mathbb{H})\mathbb{X} = \mathbf{0}$,

$$\boldsymbol{u} = (\mathbb{I}-\mathbb{H})\boldsymbol{Y} = (\mathbb{I}-\mathbb{H})(\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = (\mathbb{I}-\mathbb{H})\mathbb{X}\boldsymbol{\beta} + (\mathbb{I}-\mathbb{H})\boldsymbol{\varepsilon} = (\mathbb{I}-\mathbb{H})\boldsymbol{\varepsilon}.$$

As for (d), it is easy to verify that

$$\widehat{\boldsymbol{Y}}^\top \boldsymbol{u} = \boldsymbol{Y}^\top \mathbb{H}(\mathbb{I}-\mathbb{H})\boldsymbol{Y} = \boldsymbol{Y}^\top(\mathbb{H}-\mathbb{H}\mathbb{H})\boldsymbol{Y} = 0.$$

Geometric interpretation of the LSE

From Linear Algebra:

Consider a vector space V and two subspaces U and W such that $V = U \cap W$. U and W are orthogonal iff $\boldsymbol{u}^\top \boldsymbol{w} = 0$ for any $\boldsymbol{u} \in U$, $\boldsymbol{w} \in W$. Then we denote $W = U^\perp$.

Any vector $\boldsymbol{v} \in V$ can be uniquely decomposed as $\boldsymbol{u}_v + \boldsymbol{w}_v$, where $\boldsymbol{u}_v \in U$ and $\boldsymbol{w}_v \in U^\perp$. This is called *orthogonal projection*. Projection is a linear transformation of the vector through a projection matrix $\mathbb{P}$. The columns of $\mathbb{P}$ are the projections of basis vectors of V, and U is the image of $\mathbb{P}$.

A square matrix $\mathbb{P}$ is a projection matrix if and only if it is idempotent.

Let $\mathbb{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)$ be any basis of a subspace U of V. Then $\mathbb{A}(\mathbb{A}^\top\mathbb{A})^{-1}\mathbb{A}^\top$ is a projection matrix of V onto U.

Let $\mathscr{M}(\mathbb{X})$ be the linear subspace of $\mathbb{R}^n$ generated by the columns of the regression matrix $\mathbb{X}$ (denote them by $\boldsymbol{x}_j$, $j = 1, \ldots, p$):

$$\mathscr{M}(\mathbb{X}) = \left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \sum_{j=1}^{p} q_j \boldsymbol{x}_j, q_j \in \mathbb{R}\right\}.$$

Let $\mathscr{M}(\mathbb{X})^\perp$ be the subspace orthogonal to $\mathscr{M}(\mathbb{X})$:

$$\mathscr{M}(\mathbb{X})^\perp = \left\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^\top \boldsymbol{x} = 0 \ \forall \boldsymbol{x} \in \mathscr{M}(\mathbb{X})\right\}.$$

Then

- $\widehat{\boldsymbol{Y}}$ is the orthogonal projection of $\boldsymbol{Y} \in \mathbb{R}^n$ to the p-dimensional subspace $\mathscr{M}(\mathbb{X})$, with the projection matrix $\mathbb{H}$;
- $\boldsymbol{u}$ is the orthogonal projection of $\boldsymbol{Y} \in \mathbb{R}^n$ to the $n-p$-dimensional subspace $\mathscr{M}(\mathbb{X})^\perp$, with the projection matrix $\mathbb{I}-\mathbb{H}$.

So, $\mathbb{H}$ and $\mathbb{I}-\mathbb{H}$ are projection matrices to the two orthogonal subspaces, $\mathscr{M}(\mathbb{X})$ and $\mathscr{M}(\mathbb{X})^\perp$, respectively.

2.4. Residual Sum of Squares

The residual sum of squares, denoted by SS_e, is the sum of squared residuals and at the same time the minimized value of the least squares criterion $SS_e(\boldsymbol{\beta})$. There are several alternative ways how to express it.

$$SS_e \equiv SS_e(\widehat{\boldsymbol{\beta}}) = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2 = \|\boldsymbol{u}\|^2 = \sum_{i=1}^{n} u_i^2.$$

According to the note on p. 19, part (b), $\boldsymbol{u} = (\mathbb{I}-\mathbb{H})\boldsymbol{Y} = (\mathbb{I}-\mathbb{H})\boldsymbol{\varepsilon}$. Because $\mathbb{I}-\mathbb{H}$ is idempotent, SS_e can be expressed as a quadratic form in two alternative ways:

$$SS_e = \boldsymbol{Y}^\mathsf{T}(\mathbb{I}-\mathbb{H})\boldsymbol{Y} = \boldsymbol{\varepsilon}^\mathsf{T}(\mathbb{I}-\mathbb{H})\boldsymbol{\varepsilon}.$$

Another way to express residual sum of squares is this:

$$\begin{aligned} SS_e = (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}^\mathsf{T}\boldsymbol{Y} - \boldsymbol{Y}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\mathsf{T}\mathbb{X}^\mathsf{T}\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})\widehat{\boldsymbol{\beta}} = \\ = \boldsymbol{Y}^\mathsf{T}\boldsymbol{Y} - \boldsymbol{Y}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y}^\mathsf{T}\boldsymbol{Y} - \boldsymbol{Y}^\mathsf{T}\widehat{\boldsymbol{Y}}. \end{aligned} \tag{2.4}$$

2.5. Equivalence of Regression Models

Consider two different regression models for the same response $\boldsymbol{Y}$:

$$\begin{aligned} \boldsymbol{Y} &= \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, && \text{where } \mathbb{X}_{n\times p} \text{ and } \boldsymbol{\beta}_{p\times 1}, \\ \text{and}\quad \boldsymbol{Y} &= \mathbb{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*, && \text{where } \mathbb{X}^*_{n\times q} \text{ and } \boldsymbol{\beta}^*_{q\times 1}. \end{aligned}$$

The two models are called *equivalent* if and only if $\mathscr{M}(\mathbb{X}) = \mathscr{M}(\mathbb{X}^*)$, that is, the linear subspaces generated by the columns of $\mathbb{X}$ and $\mathbb{X}^*$, respectively, are the same. This is true if and only if there exists a $q \times p$ matrix $\mathbb{C}$ such that $\mathbb{X} = \mathbb{X}^*\mathbb{C}$. For this particular $\mathbb{C}$, it follows that $\mathbb{X}\boldsymbol{\beta} = \mathbb{X}^*\mathbb{C}\boldsymbol{\beta}$ and hence $\boldsymbol{\beta}^* = \mathbb{C}\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}^* = \boldsymbol{\varepsilon}$.

Because the fitted values $\widehat{\boldsymbol{Y}}$ in the two models are projections of the same vector $\boldsymbol{Y}$ into the same linear subspace, they must be the same in both models. The same is true for the residuals $\boldsymbol{u}$ and the residual sum of squares SS_e.

When $\mathbb{X}^*_{n\times q}$ is a matrix of rank $p < q$ then there exists a full rank matrix $\mathbb{X}_{n\times p}$ that generates an equivalent model. This is the mechanism how to avoid ever considering non-full rank regression matrices. If a regression matrix is not of full rank we work instead with an equivalent model, which is of full rank.

2.6. Model for iid Response

The simplest special case of a regression model describes independent and identically distributed responses. Let $Y_1, \ldots, Y_n$ be iid random variables with $\mathsf{E}\,Y_i = \mu$ and $\mathsf{var}\,Y_i = \sigma_Y^2$.

Write

$$Y_i = \mu + (Y_i - \mu) \equiv X_i \beta + \varepsilon_i,$$

where $X_i = 1$ for all i, $\beta = \mu$, $\mathsf{E}\,\varepsilon_i = 0$, and $\mathsf{var}\,\varepsilon_i = \sigma_Y^2$. This is a linear model. We can write the vector containing all the responses in the form

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbb{X} = (1, \dots, 1)^\top \equiv \mathbf{1}_n$, $\boldsymbol{\beta} = \mu$.

Notation.

- Let $\mathbf{1}_n$ be a column n-vector of ones; $\mathbf{1}_n = (1, \dots, 1)^\top$.
- Let $\mathbb{J}_n = \mathbf{1}_n \mathbf{1}_n^\top$ be an $n \times n$ matrix of ones.

Let us now calculate the least squares estimator and residual sum of squares. We have

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{Y} = (\mathbf{1}_n^\top \mathbf{1}_n)^{-1} (\mathbf{1}_n^\top \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i \equiv \overline{Y}_n.$$

So, the least squares estimate of the common expectation is the arithmetic average. Next, the fitted values are $\widehat{\boldsymbol{Y}} = \overline{Y}_n \mathbf{1}_n$ and the residuals are $\boldsymbol{u} = \boldsymbol{Y} - \overline{Y}_n \mathbf{1}_n$. The residual sum of squares is $SS_e = \boldsymbol{u}^\top \boldsymbol{u} = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$.

2.7. Model With Centered Covariates

In order to gain further insights into the meaning of the LSE procedure, we need to center the covariates. Consider the model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the first column of $\mathbb{X}$ is $\mathbf{1}_n$ (the intercept column). Denote the rest of the regression matrix as $\mathbb{X}_R$, that is, $\mathbb{X} = (\mathbf{1}_n | \mathbb{X}_R)$. The vector $\boldsymbol{\beta}$ is divided similarly into $\boldsymbol{\beta} = \binom{\beta_1}{\boldsymbol{\beta}_R}$.

Each observation can be expressed as

$$Y_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i.$$

Let $\overline{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$ for $j = 2, \dots, n$. Now, subtract from the value of each covariate the respective mean (except for the intercept). We get

$$Y_i = \alpha + \beta_2 (X_{i2} - \overline{X}_2) + \cdots + \beta_p (X_{ip} - \overline{X}_p) + \varepsilon_i,$$

where $\alpha = \beta_1 + \beta_2 \overline{X}_2 + \cdots + \beta_p \overline{X}_p$ to maintain the equality. This is the model with *centered covariates* (shortly, the "centered model"). It is an equivalent model (the subspaces generated by the columns of the regression matrix have not changed) and the parameters $\beta_2, \dots, \beta_p$ are the same. Only the intercept parameter is different. The new intercept has the interpretation $\mathsf{E}\big[Y_i \,\big|X_{i2} = \overline{X}_2, \dots, X_{ip} = \overline{X}_p\big]$, the expected response for an individual with average value in all covariates.

Message: If any covariate is shifted by a constant (the same number is added to/subtracted from all values of the covariate) the regression parameter for that covariate is not changed.

Take $\mathbb{J}_n = \mathbf{1}_n \mathbf{1}_n^\top$, an $n \times n$ matrix with 1 at all positions. The centered covariates can be created by multiplication by the column centering matrix: $\mathbb{X}_C = (\mathbb{I}_n - n^{-1}\mathbb{J}_n)\mathbb{X}_R$. The centered model can be written as

$$\boldsymbol{Y} = (\mathbf{1}_n | \mathbb{X}_C)\begin{pmatrix}\alpha\\ \boldsymbol{\beta}_R\end{pmatrix} + \boldsymbol{\varepsilon}.$$

Let us find the least squares estimate of $(\alpha, \boldsymbol{\beta}_R)$. The original model and the centered model are equivalent, they have the same fitted values $\widehat{Y}_i$. Let $\widehat{\boldsymbol{\beta}} = \binom{\widehat{\beta}_1}{\widehat{\boldsymbol{\beta}}_R}$ be the LSE in the original model, $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top \boldsymbol{Y}$. Then for all $i = 1, \dots, n$,

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_1 + \widehat{\beta}_2 X_{i2} + \dots + \widehat{\beta}_p X_{ip}\\ &= \widehat{\alpha} + \widehat{\beta}_2(X_{i2} - \overline{X}_2) + \dots + \widehat{\beta}_p(X_{ip} - \overline{X}_p),\end{aligned}$$

where $\widehat{\alpha} = \widehat{\beta}_1 + \widehat{\beta}_2\overline{X}_2 + \dots + \widehat{\beta}_p\overline{X}_p$. Because $\binom{\widehat{\alpha}}{\widehat{\boldsymbol{\beta}}_R}$ generates the same fitted values, residuals and SS_e as the LSE of the original model, it must be the unique LSE in the centered model.

Message: If any covariate is shifted by a constant (the same number is added to/subtracted from all values of the covariate) the LSE of the regression parameter for that covariate is not changed.

The end of lecture 3 (Oct. 10, 2023)

Now, apply the LSE formula to the centered model. We have

$$\begin{aligned}\begin{pmatrix}\widehat{\alpha}\\ \widehat{\boldsymbol{\beta}}_R\end{pmatrix} &= \big[(\mathbf{1}_n|\mathbb{X}_C)^\top(\mathbf{1}_n|\mathbb{X}_C)\big]^{-1}(\mathbf{1}_n|\mathbb{X}_C)^\top \boldsymbol{Y}\\ &= \begin{pmatrix} n & \mathbf{1}_n^\top\mathbb{X}_C\\ \mathbb{X}_C^\top\mathbf{1}_n & \mathbb{X}_C^\top\mathbb{X}_C\end{pmatrix}^{-1}\begin{pmatrix}\sum_{i=1}^n Y_i\\ \mathbb{X}_C^\top\boldsymbol{Y}\end{pmatrix} = \begin{pmatrix}\frac{1}{n} & \mathbf{0}\\ \mathbf{0} & (\mathbb{X}_C^\top\mathbb{X}_C)^{-1}\end{pmatrix}\begin{pmatrix}\sum_{i=1}^n Y_i\\ \mathbb{X}_C^\top\boldsymbol{Y}\end{pmatrix} = \begin{pmatrix}\overline{Y}\\ (\mathbb{X}_C^\top\mathbb{X}_C)^{-1}(\mathbb{X}_C^\top\boldsymbol{Y})\end{pmatrix}.\end{aligned}$$

We have verified that $\widehat{\alpha} = \overline{Y}$. The fitted values in the centered model are

$$\widehat{Y}_i = \overline{Y} + \widehat{\beta}_2(X_{i2} - \overline{X}_2) + \dots + \widehat{\beta}_p(X_{ip} - \overline{X}_p).$$

Because the original model has the same fitted values, we have the following conclusion.

Conclusion: If the model includes the intercept column, *the fitted value* evaluated at the average value of each of the remaining covariates is equal to *the average of the responses*.

We can also construct an additional way to express the residual sum of squares in a model with intercept. In the original model, we have $SS_e = \boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{Y}^\top\mathbb{X}\widehat{\boldsymbol{\beta}}$, see (2.4). When we apply this to the centered model, we get

$$SS_e = \boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{Y}^\top(\mathbf{1}_n|\mathbb{X}_C)\begin{pmatrix}\overline{Y}\\ \widehat{\boldsymbol{\beta}}_R\end{pmatrix} = \boldsymbol{Y}^\top\boldsymbol{Y} - n\overline{Y}^2 - \boldsymbol{Y}^\top\mathbb{X}_C\widehat{\boldsymbol{\beta}}_R = \sum_{i=1}^{n}(Y_i-\overline{Y})^2 - \boldsymbol{Y}^\top\mathbb{X}_C\widehat{\boldsymbol{\beta}}_R.$$

2.8. Relationship to Sample Covariance Matrices

In this section, we still work under the assumption that $\mathbf{1}_n \in \mathscr{M}(\mathbb{X})$ (the intercept is included in the model). Denote by $\mathbb{S}_{XX}$ the sample covariance matrix of the columns of $\mathbb{X}_R$ (the remaining columns of the regression matrix after excluding the intercept). It is a $(p-1)\times(p-1)$ matrix with diagonal elements $\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)^2$ and off-diagonal elements $\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$. Obviously, $\mathbb{S}_{XX} = \frac{1}{n-1}\mathbb{X}_C^\top\mathbb{X}_C$.

Now consider the sample covariance matrix[*] $\mathbb{S}_{XY}$ of the columns of $\mathbb{X}_R$ with the response vector $\boldsymbol{Y}$, a $(p-1)\times 1$ matrix with elements $\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(Y_i-\overline{Y})$. Because $\sum_{i=1}^{n}(X_{ij}-\overline{X})\overline{Y} = 0$, we have $\mathbb{S}_{XY} = \frac{1}{n-1}\mathbb{X}_C^\top\boldsymbol{Y}$.

Conclusion: If the model includes the intercept column, the LSE of the non-intercept parameters can be expressed in terms of sample covariance matrices as follows: $\widehat{\boldsymbol{\beta}}_R = \mathbb{S}_{XX}^{-1}\mathbb{S}_{XY}$.

We can also express the LSE of the intercept parameter using the results of the previous section.

$$\widehat{\beta}_1 = \widehat{\alpha} - \frac{1}{n}\mathbf{1}_n^\top\mathbb{X}_R\widehat{\boldsymbol{\beta}}_R = \overline{Y} - \frac{1}{n}\mathbf{1}_n^\top\mathbb{X}_R\mathbb{S}_{XX}^{-1}\mathbb{S}_{XY}.$$

2.9. Decomposition of Sums of Squares

This can be done in two ways – for non-centered or centered response. The first decomposition is universally valid but less useful. The second is more useful but holds only if the intercept is included in the model.

Decomposition of sums of squares with non-centered response

Start with the sum of squared responses

$$\|\boldsymbol{Y}\|^2 = \boldsymbol{Y}^\top\boldsymbol{Y} = \boldsymbol{Y}^\top\mathbb{H}\boldsymbol{Y} + \boldsymbol{Y}^\top(\mathbb{I}-\mathbb{H})\boldsymbol{Y}.$$

The last term on the right-hand side can be recognized as the *residual sum of squares* SS_e. The left-hand side is called *the non-centered total sum of squares*, denoted by SS_T^*. The remaining

[*] actually, it is a vector

term, $\boldsymbol{Y}^{\mathsf{T}}\mathbb{H}\boldsymbol{Y}$, is called *the non-centered regression sum of squares*, denoted by SS_R^*. We have

$$SS_R^* = \boldsymbol{Y}^{\mathsf{T}}\mathbb{H}\boldsymbol{Y} = \boldsymbol{Y}^{\mathsf{T}}\mathbb{H}\mathbb{H}\boldsymbol{Y} = \|\mathbb{H}\boldsymbol{Y}\|^2 = \|\widehat{\boldsymbol{Y}}\|^2 = \|\mathbb{X}\widehat{\boldsymbol{\beta}}\|^2 = \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\mathbb{X}\widehat{\boldsymbol{\beta}}$$

The non-centered decomposition is

$$\underbrace{\sum_{i=1}^{n} Y_i^2}_{SS_T^*} = \underbrace{\sum_{i=1}^{n} \widehat{Y}_i^2}_{SS_R^*} + \underbrace{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}_{SS_e}.$$

Decomposition of sums of squares with centered response

Assume that the model contains the intercept, $\mathbf{1}_n \in \mathscr{M}(\mathbb{X})$. Calculate the mean response $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and subtract the mean from all responses, that is, take

$$\boldsymbol{Y} - \mathbf{1}_n\overline{Y} = \boldsymbol{Y} - \mathbf{1}_n\tfrac{1}{n}\mathbf{1}_n^{\mathsf{T}}\boldsymbol{Y} = \boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y}.$$

Now apply the decomposition of sums of squares to these centered responses.

The total (centered) sum of squares is

$$SS_T \equiv \|\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y}\|^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

This can be decomposed as

$$SS_T = (\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y})^{\mathsf{T}}\mathbb{H}(\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y}) + (\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y})^{\mathsf{T}}(\mathbb{I} - \mathbb{H})(\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y}).$$

Because the model contains the intercept, $\mathbb{H}\mathbf{1}_n = \mathbf{1}_n$, hence $\mathbb{H}\mathbb{J}_n = \mathbb{J}_n$, hence $(\mathbb{I} - \mathbb{H})\mathbb{J}_n = \mathbf{0}$. Thus, the last term on the right-hand side is still the *residual sum of squares SS_e*.

The remaining term, $(\boldsymbol{Y} - \frac{1}{n}\mathbb{J}_n)^{\mathsf{T}}\mathbb{H}(\boldsymbol{Y} - \frac{1}{n}\mathbb{J}_n)$, is *the (centered) regression sum of squares*, denoted by SS_R. We have

$$\begin{aligned} SS_R &= (\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y})^{\mathsf{T}}\mathbb{H}(\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y}) = \|\mathbb{H}(\boldsymbol{Y} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y})\|^2 = \|\mathbb{H}\boldsymbol{Y} - \tfrac{1}{n}\underbrace{\mathbb{H}\mathbb{J}_n}_{\mathbb{J}_n}\boldsymbol{Y}\|^2 \\ &= \|\widehat{\boldsymbol{Y}} - \tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y}\|^2 = \|\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}\|^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2. \end{aligned}$$

The centered decomposition of sums of squares is

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}_{SS_R} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{SS_e}.$$

This can be interpreted as follows. The total sum of squares SS_T captures the total variability in the response. This is decomposed into SS_R, the variability that is explained by the regression model (using the covariates), and into SS_e, which is the part of variability that could not be explained.

Notice that we have the mean of all responses in the expression for SS_R instead of the mean of the fitted values.

2.10. Coefficient of Determination

We continue to assume that the model contains the intercept, $\mathbf{1}_n \in \mathscr{M}(\mathbb{X})$, and recall the centered decomposition of sums of squares $SS_T = SS_R + SS_e$ derived in the previous section.

Definition 2.4 (Coefficient of determination). The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_e}{SS_T} \qquad \nabla$$

is called *the coefficient of determination**.

If we interpret SS_T as the total variability of the response and SS_R as the variability explained by the covariates included in the model, we can view R^2 as the fraction of the variability of the response that was explained by the regression model.

Notes on coefficient of determination

1. Obviously, $0 \le R^2 \le 1$.
2. $\sqrt{R^2}$ is sometimes called *multiple correlation coefficient*† between the random variable Y and random vector $\boldsymbol{X}$.
3. R^2 is equal to the square of the estimated correlation coefficient between $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$.

   Proof.

   $$R^2 = \frac{||\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}||^2}{||\boldsymbol{Y} - \mathbf{1}_n\overline{Y}||^2} = \frac{||\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}||^4}{||\boldsymbol{Y} - \mathbf{1}_n\overline{Y}||^2||\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}||^2}$$

   Now express the norm in the numerator differently:

   $$\begin{aligned} ||\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}||^2 &= (\widehat{\boldsymbol{Y}} - \boldsymbol{Y} + \boldsymbol{Y} - \mathbf{1}_n\overline{Y})^\top(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}) \\ &= \underbrace{(\widehat{\boldsymbol{Y}} - \boldsymbol{Y})^\top(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y})}_{=0} + (\boldsymbol{Y} - \mathbf{1}_n\overline{Y})^\top(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}) \qquad (2.5) \\ &= (\boldsymbol{Y} - \mathbf{1}_n\overline{Y})^\top(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}) \end{aligned}$$

* Česky *koeficient determinace* † Česky *koeficient mnohonásobné korelace*

The first term on the second line is zero because

$$(\widehat{\boldsymbol{Y}}-\boldsymbol{Y})^{\mathsf{T}}(\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y})=(\mathbb{H}\boldsymbol{Y}-\boldsymbol{Y})^{\mathsf{T}}(\mathbb{H}\boldsymbol{Y}-\tfrac{1}{n}\mathbb{J}_n\boldsymbol{Y})=-\boldsymbol{Y}^{\mathsf{T}}(\mathbb{I}-\mathbb{H})(\mathbb{H}-\tfrac{1}{n}\mathbb{J}_n)\boldsymbol{Y}$$

and

$$(\mathbb{I}-\mathbb{H})(\mathbb{H}-\tfrac{1}{n}\mathbb{J}_n)=(\mathbb{I}-\mathbb{H})\mathbb{H}-\tfrac{1}{n}(\mathbb{I}-\mathbb{H})\mathbb{J}_n=0$$

because $\mathbf{1}_n\in\mathscr{M}(\mathbb{X})$. So,

$$R^2=\left[\frac{(\boldsymbol{Y}-\mathbf{1}_n\overline{Y})^{\mathsf{T}}(\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y})}{\sqrt{\|\boldsymbol{Y}-\mathbf{1}_n\overline{Y}\|^2\|\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y}\|^2}}\right]^2=\widehat{\mathsf{cor}}^2(\boldsymbol{Y},\widehat{\boldsymbol{Y}}).$$

□

4. Another variant of the coefficient of determination is so called *adjusted* R^2 defined as

$$R_a^2=1-\frac{n-1}{n-p}\frac{SS_e}{SS_T}.$$

The motivation for this is to subtract the ratio of two unbiased estimators of $\mathsf{var}\,\varepsilon_i$ and $\mathsf{var}\,Y_i$[*].

The end of lecture 4 (Oct. 13, 2023)

2.11. LSE Under Linear Restrictions

Consider the linear model $\boldsymbol{Y}=\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ with $\mathbb{X}$ of full rank. The least squares estimator $\widehat{\boldsymbol{\beta}}=(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\boldsymbol{Y}$ minimizes the residual sum of squares $SS_e(\boldsymbol{\beta})=\|\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta}\|^2$ over all $\boldsymbol{\beta}\in\mathbb{R}^p$.

Now we impose an additional set of linear restrictions on the parameters: let $\mathbb{C}\boldsymbol{\beta}=\boldsymbol{c}$, where $\mathbb{C}$ is a $q\times p$ matrix with rank $r(\mathbb{C})=q<p$ and $\boldsymbol{c}\in\mathbb{R}^q$. We will minimize $SS_e(\boldsymbol{\beta})$ over the set $\mathscr{B}=\{\boldsymbol{\beta}\in\mathbb{R}^p:\mathbb{C}\boldsymbol{\beta}=\boldsymbol{c}\}$. Denote $\widehat{\boldsymbol{\beta}}_C=\arg\min_{\boldsymbol{\beta}\in\mathscr{B}}\|\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta}\|^2$.

We can use the method of Lagrange multipliers to calculate $\widehat{\boldsymbol{\beta}}_C$. Introduce the objective function

$$S(\boldsymbol{\beta},\boldsymbol{\lambda})=SS_e(\boldsymbol{\beta})+\boldsymbol{\lambda}^{\mathsf{T}}(\mathbb{C}\boldsymbol{\beta}-\boldsymbol{c}),$$

where $\boldsymbol{\lambda}\in\mathbb{R}^q$. Calculate

$$\frac{\partial S(\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}}=-2\mathbb{X}^{\mathsf{T}}\boldsymbol{Y}+2\mathbb{X}^{\mathsf{T}}\mathbb{X}\boldsymbol{\beta}+\mathbb{C}^{\mathsf{T}}\boldsymbol{\lambda}$$

and set it equal to zero to find $\widehat{\boldsymbol{\beta}}_C$. We get

$$\mathbb{X}^{\mathsf{T}}\mathbb{X}\widehat{\boldsymbol{\beta}}_C=\mathbb{X}^{\mathsf{T}}\boldsymbol{Y}-\frac{1}{2}\mathbb{C}^{\mathsf{T}}\boldsymbol{\lambda}$$

[*] The fact that $SS_e/(n-p)$ is an unbiased estimator of σ_e^2 will be established in Section **??**.

and hence

$$\widehat{\beta}_C = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}[\mathbb{X}^\mathsf{T}\boldsymbol{Y} - \frac{1}{2}\mathbb{C}^\mathsf{T}\boldsymbol{\lambda}] = \widehat{\boldsymbol{\beta}} - \frac{1}{2}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}\boldsymbol{\lambda}. \tag{2.6}$$

The solution must satisfy the constraint $\mathbb{C}\widehat{\boldsymbol{\beta}}_C = \boldsymbol{c}$, i.e.,

$$\mathbb{C}\widehat{\boldsymbol{\beta}} - \frac{1}{2}\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}\boldsymbol{\lambda} = \boldsymbol{c}.$$

Use this to identify $\boldsymbol{\lambda}$: it is a solution to the system of linear equations

$$\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c} = \frac{1}{2}\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}\boldsymbol{\lambda}.$$

Since $r(\mathbb{X}) = p$ and $r(\mathbb{C}) = q < p$, the $q \times q$ matrix $\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}$ is of rank q, therefore regular and invertible. Thus,

$$\boldsymbol{\lambda} = 2[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}).$$

Plug this into (2.6) to obtain the result

$$\widehat{\boldsymbol{\beta}}_C = \widehat{\boldsymbol{\beta}} - (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}). \tag{2.7}$$

However, this is only a suspicious point. We still need to show that it really minimizes $SS_e(\boldsymbol{\beta})$ over $\boldsymbol{\beta} \in \mathscr{B}$. So, take any $\boldsymbol{\beta} \in \mathscr{B}$ and write

$$\begin{aligned} SS_e(\boldsymbol{\beta}) &= \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C + \mathbb{X}\widehat{\boldsymbol{\beta}}_C - \mathbb{X}\boldsymbol{\beta}\|^2 \\ &= \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C\|^2 + \|\mathbb{X}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})\|^2 + 2(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C)^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}) \end{aligned}$$

Look at the last term. From (2.7), we have

$$\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}} + \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})$$

and

$$\begin{aligned} (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C)^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}) &= \underbrace{(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}\mathbb{X}}_{=\boldsymbol{u}^\mathsf{T}\mathbb{X}=\boldsymbol{0}}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}) \\ &\quad + (\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}\underbrace{\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})}_{=\mathbb{C}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}) = \boldsymbol{c} - \boldsymbol{c} = \boldsymbol{0}} \\ &= \boldsymbol{0}. \end{aligned}$$

Thus, for any $\boldsymbol{\beta} \in \mathscr{B}$,

$$SS_e(\boldsymbol{\beta}) = SS_e(\widehat{\boldsymbol{\beta}}_C) + (\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})^\mathsf{T}\mathbb{X}^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}) \geq SS_e(\widehat{\boldsymbol{\beta}}_C)$$

and equality is attained if and only if $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_C$. Thus, $\widehat{\boldsymbol{\beta}}_C$ is the unique minimizer of $SS_e(\boldsymbol{\beta})$ over $\boldsymbol{\beta} \in \mathscr{B}$ and therefore it is the restricted LSE.

Now evaluate the difference between $SS_e = SS_e(\widehat{\boldsymbol{\beta}})$ and $SS_e(\widehat{\boldsymbol{\beta}}_C)$. Since $\widehat{\boldsymbol{\beta}}_C$ minimizes SS_e over a subspace of $\mathbb{R}^p$, $SS_e \leq SS_e(\widehat{\boldsymbol{\beta}}_C)$. Write

$$\begin{aligned} SS_e(\widehat{\boldsymbol{\beta}}_C) &= \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C\|^2 = \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}} + \mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widehat{\boldsymbol{\beta}}_C\|^2 \\ &= \|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\mathbb{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_C)\|^2 + 2(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_C)^\mathsf{T} \underbrace{\mathbb{X}^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})}_{=\mathbb{X}^\mathsf{T}\boldsymbol{u}=\boldsymbol{0}}. \end{aligned}$$

Hence

$$SS_e(\widehat{\boldsymbol{\beta}}_C) = SS_e + (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_C)^\mathsf{T}\mathbb{X}^\mathsf{T}\mathbb{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_C).$$

From (2.7) we know that

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_C = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}).$$

Plug it into the previous expression and after canceling unnecessary terms we get

$$SS_e(\widehat{\boldsymbol{\beta}}_C) = SS_e + (\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}). \tag{2.8}$$

Summary: In this section, we have derived two important results for the least squares estimator $\widehat{\boldsymbol{\beta}}_C$ calculated under linear restrictions $\mathbb{C}\boldsymbol{\beta} = \boldsymbol{c}$:

$$\begin{gathered} \widehat{\boldsymbol{\beta}}_C = \widehat{\boldsymbol{\beta}} - (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}), \\ SS_e(\widehat{\boldsymbol{\beta}}_C) = SS_e + (\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}). \end{gathered}$$

3. Properties of the Least Squares Estimator

In this chapter, we start investigating probabilistic and statistical properties of the quantities that were introduced in the previous chapter. The first two sections apply to the general linear regression model, the third section requires the additional condition of normality of the responses (or of the error terms).

3.1. Moment Properties of the Least Squares Estimator

Consider the regression model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\mathsf{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$ and $\mathsf{var}\,\boldsymbol{\varepsilon} = \sigma_e^2 \mathbb{I}_n$ or, equivalently, $\mathsf{E}\,\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}$ and $\mathsf{var}\,\boldsymbol{Y} = \sigma_e^2 \mathbb{I}_n$. Let the regression matrix $\mathbb{X}_{n\times p}$ have a full rank $p < n$. The least squares estimator $\widehat{\boldsymbol{\beta}}$ can be expressed as

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}.$$

The first lemma specifies the first and the second moment of $\widehat{\boldsymbol{\beta}}$ (conditionally on the covariates).

Lemma 3.1.

(i) $\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, i.e., $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

(ii) $\mathsf{var}\,\widehat{\boldsymbol{\beta}} = \sigma_e^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$. ♢

Proof. Treating $\mathbb{X}$ as a matrix of constants and $\boldsymbol{Y}$ as a random vector, we get:

$$\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \mathsf{E}\,(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathsf{E}\,\boldsymbol{Y} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

and

$$\begin{aligned}\mathsf{var}\,\widehat{\boldsymbol{\beta}} &= \mathsf{var}\,(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathsf{var}\,\boldsymbol{Y}\,\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1} \\ &= \sigma_e^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}(\mathbb{X}^\mathsf{T}\mathbb{X})\,(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1} = \sigma_e^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}.\end{aligned}$$ □

The second lemma specifies the first and the second moments of the fitted values and residuals. Its proof is also straightforward.

Lemma 3.2.

(i) $E\widehat{Y} = EY = \mathbb{X}\boldsymbol{\beta}$,
(ii) $E\boldsymbol{u} = \mathbf{0}$,
(iii) $var\,\widehat{Y} = \sigma_e^2\mathbb{H}$,
(iv) $var\,\boldsymbol{u} = \sigma_e^2(\mathbb{I}-\mathbb{H})$. ◇

Proof. We have $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$ and $\boldsymbol{u} = (\mathbb{I}-\mathbb{H})\boldsymbol{Y}$, where $\mathbb{H} = \mathbb{X}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}$ is the projection matrix to the subspace $\mathscr{M}(\mathbb{X})$. $\mathbb{H}$ is symmetric, idempotent, and satisfies $\mathbb{H}\mathbb{X} = \mathbb{X}$ and $(\mathbb{I}-\mathbb{H})\mathbb{X} = \mathbf{0}$. Hence

$$\mathsf{E}\,\widehat{\boldsymbol{Y}} = \mathsf{E}\,\mathbb{H}\boldsymbol{Y} = \mathbb{H}\mathsf{E}\,\boldsymbol{Y} = \mathbb{H}\mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta},$$
$$\mathsf{var}\,\widehat{\boldsymbol{Y}} = \mathsf{var}\,\mathbb{H}\boldsymbol{Y} = \mathbb{H}\mathsf{var}\,\boldsymbol{Y}\mathbb{H} = \sigma_e^2\mathbb{H}\mathbb{H} = \sigma_e^2\mathbb{H}.$$

Next,

$$\mathsf{E}\,\boldsymbol{u} = \mathsf{E}\,(\mathbb{I}-\mathbb{H})\boldsymbol{Y} = (\mathbb{I}-\mathbb{H})\mathsf{E}\,\boldsymbol{Y} = (\mathbb{I}-\mathbb{H})\mathbb{X}\boldsymbol{\beta} = \mathbf{0},$$
$$\mathsf{var}\,\boldsymbol{u} = \mathsf{var}\,(\mathbb{I}-\mathbb{H})\boldsymbol{Y} = (\mathbb{I}-\mathbb{H})\mathsf{var}\,\boldsymbol{Y}(\mathbb{I}-\mathbb{H}) = \sigma_e^2(\mathbb{I}-\mathbb{H})(\mathbb{I}-\mathbb{H}) = \sigma_e^2(\mathbb{I}-\mathbb{H}).$$ □

It is important to realize one substantial difference. We can write the responses in two different ways:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$
$$\boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{u}$$

In the first case, the error terms $\boldsymbol{\varepsilon}$ are independent and have equal variances. However, in the second case, the residuals $\boldsymbol{u}$ do not share these properties: they are not independent (because the matrix $\mathbb{I}-\mathbb{H}$ is not diagonal) and they do not have equal variances.

One can fix the unequal variances of the raw residuals by introducing so called *standardized residuals*. Standardized residuals have equal variances but are not independent.

Definition 3.1.

$$u_i^* = \frac{u_i}{\sqrt{1-h_{ii}}},$$

where h_{ii} is the i-th diagonal element of the matrix $\mathbb{H}$, are called *standardized residuals*. ∇

Finally, we calculate the expectation of the residual sum of squares and derive an unbiased estimator for the residual variance.

Lemma 3.3. $E\,SS_e = (n-p)\sigma_e^2$. ◇

Proof. Remembering the results from Section 2.4, we can write $SS_e = \boldsymbol{u}^\mathsf{T}\boldsymbol{u} = \boldsymbol{\varepsilon}^\mathsf{T}(\mathbb{I}-\mathbb{H})\varepsilon$. By Lemma A.1 in the Appendix and using the fact that $\mathbb{I}-\mathbb{H}$ is idempotent of rank $n-p$ — see equation (2.3) — we get

$$\begin{aligned} \mathsf{E}\, SS_e &= \mathsf{E}\, \boldsymbol{\varepsilon}^\mathsf{T}(\mathbb{I}-\mathbb{H})\varepsilon = 0 + \operatorname{tr}\left[(\mathbb{I}-\mathbb{H})\operatorname{\mathsf{var}} \boldsymbol{\varepsilon}\right] \\ &= \sigma_e^2 \operatorname{tr}(\mathbb{I}-\mathbb{H}) = \sigma_e^2 r(\mathbb{I}-\mathbb{H}) = \sigma_e^2(n-p). \end{aligned}$$ □

Definition 3.2.

$$\widehat{\sigma}_e^2 = \frac{SS_e}{n-p} \equiv MS_e$$

is called *the estimated residual variance*. The symbol MS_e is just an alternative notation for the expression $SS_e/(n-p)$. ∇

By Lemma 3.3, $\widehat{\sigma}_e^2$ is an unbiased estimator of the residual variance.

3.2. Gauss-Markov Theorem

The Gauss-Markov theorem shows that the least squares estimator is in a certain sense optimal. It was originally formulated by Carl Friedrich Gauss in 1821 (Gauss 1821) under the assumption of normality. It was extended to the general case by Andrey Andreyevich Markov in 1912 (Markov 1912). Further extension to correlated errors of unequal variance was provided by Aitken (1936).*

> ***Andrey Andreyevich Markov*** *(1856 – 1922) was a Russian mathematician, who became particularly famous for his pioneering work on stochastic processes (Markov property, Markov chains, etc.).*
> *Source:* `https://en.wikipedia.org/wiki/Andrey_Markov`

Here we state the Gauss-Markov theorem in three different ways, after we introduce and explain the optimality criterion needed for all three versions.

Definition 3.3. $\widehat{\boldsymbol{\theta}}$ is *best linear unbiased estimator* (BLUE) of $\boldsymbol{\theta}$ based on the data vector $\boldsymbol{Y}$ if and only if the following three conditions hold:

(i) $\widehat{\boldsymbol{\theta}}$ is linear, i.e., $\widehat{\boldsymbol{\theta}} = \mathbb{A}\boldsymbol{Y}$.

(ii) $\widehat{\boldsymbol{\theta}}$ is unbiased, i.e., $\mathsf{E}\,\widehat{\boldsymbol{\theta}} = \mathsf{E}\,\mathbb{A}\boldsymbol{Y} = \boldsymbol{\theta}$.

(iii) For any matrix $\mathbb{B}$ (of the same dimension as $\mathbb{A}$) that satisfies $\mathsf{E}\,\mathbb{B}\boldsymbol{Y} = \boldsymbol{\theta}$

$$\operatorname{\mathsf{var}} \mathbb{B}\boldsymbol{Y} - \operatorname{\mathsf{var}} \widehat{\boldsymbol{\theta}} \geq 0,$$

that is, the matrix on the left-hand side is positive semi-definite. ∇

[*] we do not talk about that extension in this course

Theorem 3.4 (Gauss-Markov, version I). *Let the linear regression model specified in Section 3.1 on page 30 be satisfied, let $\widehat{\boldsymbol{\beta}}$ be the LSE. Then $\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ is the unique best linear unbiased estimator of $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$ for any $\mathbf{0} \neq \boldsymbol{c} \in \mathbb{R}^p$.* ◇

Proof.

- $\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ is linear: $\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} = \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}$.
- $\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ is unbiased: $\mathsf{E}\,\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} = \boldsymbol{c}^\mathsf{T}\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$.
- $\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators:

  Take another linear unbiased estimator $\boldsymbol{a}^\mathsf{T}\boldsymbol{Y}$ of $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$, where $\boldsymbol{a} \in \mathbb{R}^n$. We have $\mathsf{E}\,\boldsymbol{a}^\mathsf{T}\boldsymbol{Y} = \boldsymbol{a}^\mathsf{T}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$. Hence, $\boldsymbol{a}^\mathsf{T}\mathbb{X} = \boldsymbol{c}^\mathsf{T}$. Now,

  $$\begin{aligned} \mathsf{var}\,\boldsymbol{a}^\mathsf{T}\boldsymbol{Y} &= \sigma_e^2\boldsymbol{a}^\mathsf{T}\boldsymbol{a}, \\ \mathsf{var}\,\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} &= \sigma_e^2\boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\boldsymbol{c} = \sigma_e^2\boldsymbol{a}^\mathsf{T}\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{a} = \sigma_e^2\boldsymbol{a}^\mathsf{T}\mathbb{H}\boldsymbol{a}. \end{aligned}$$

  Finally,

  $$\mathsf{var}\,\boldsymbol{a}^\mathsf{T}\boldsymbol{Y} - \mathsf{var}\,\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} = \sigma_e^2\boldsymbol{a}^\mathsf{T}(\mathbb{I}-\mathbb{H})\boldsymbol{a} \geq 0$$

  because $\mathbb{I}-\mathbb{H}$ is positive semi-definite. The variances of both estimators are equal if and only if $(\mathbb{I}-\mathbb{H})\boldsymbol{a} = \mathbf{0}$, which is equivalent to $\boldsymbol{a} = \mathbb{H}\boldsymbol{a}$ or $\boldsymbol{a}^\mathsf{T} = \boldsymbol{a}^\mathsf{T}\mathbb{H}$. It follows that the estimator $\boldsymbol{a}^\mathsf{T}\boldsymbol{Y}$ can be rewritten as

  $$\boldsymbol{a}^\mathsf{T}\boldsymbol{Y} = \boldsymbol{a}^\mathsf{T}\mathbb{H}\boldsymbol{Y} = \boldsymbol{a}^\mathsf{T}\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = \boldsymbol{a}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}.$$ □

Theorem 3.5 (Gauss-Markov, version II). *Let the linear regression model specified in Section 3.1 on page 30 be satisfied, let $\widehat{\boldsymbol{\beta}}$ be the LSE and $\mathbb{C}$ any $q \times p$ matrix. Then $\mathbb{C}\widehat{\boldsymbol{\beta}}$ is a best linear unbiased estimator of $\mathbb{C}\boldsymbol{\beta}$.* ◇

Proof. This is an easy corollary of the preceding theorem. $\mathbb{C}\widehat{\boldsymbol{\beta}}$ is obviously a linear and unbiased estimator of $\mathbb{C}\boldsymbol{\beta}$. Consider another linear unbiased estimator $\mathbb{A}\boldsymbol{Y}$ with $\mathbb{A}_{q\times n}$. To be unbiased, it must satisfy $\mathsf{E}\,\mathbb{A}\boldsymbol{Y} = \mathbb{A}\mathbb{X}\boldsymbol{\beta} = \mathbb{C}\boldsymbol{\beta}$ and hence $\mathbb{A}\mathbb{X} = \mathbb{C}$ and $r(\mathbb{A}) = r(\mathbb{C})$.

Denote $\mathbb{D} = \mathsf{var}\,\mathbb{A}\boldsymbol{Y} - \mathsf{var}\,\mathbb{C}\widehat{\boldsymbol{\beta}}$ and prove that $\mathbb{D} \geq 0$ by taking any non-zero vector $\boldsymbol{d} \in \mathbb{R}^q$ and showing that $\boldsymbol{d}^\mathsf{T}\mathbb{D}\boldsymbol{d} \geq 0$. We have $\boldsymbol{d}^\mathsf{T}\mathbb{D}\boldsymbol{d} = \mathsf{var}\,\boldsymbol{d}^\mathsf{T}\mathbb{A}\boldsymbol{Y} - \mathsf{var}\,\boldsymbol{d}^\mathsf{T}\mathbb{C}\widehat{\boldsymbol{\beta}}$. Also, $\mathsf{E}\,\boldsymbol{d}^\mathsf{T}\mathbb{A}\boldsymbol{Y} = \boldsymbol{d}^\mathsf{T}\mathbb{A}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{d}^\mathsf{T}\mathbb{C}\boldsymbol{\beta}$.

Hence, $\boldsymbol{d}^\mathsf{T}\mathbb{A}\boldsymbol{Y}$ is a linear unbiased estimator of $\boldsymbol{d}^\mathsf{T}\mathbb{C}\boldsymbol{\beta}$ and it follows from Theorem 3.4 that $\boldsymbol{d}^\mathsf{T}\mathbb{D}\boldsymbol{d} \geq 0$. □

Note. It follows from Theorem 3.5 that $\widehat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$. Just take a special case with $\mathbb{C} = \mathbb{I}$.

Another special case of Theorem 3.5, with $\mathbb{C} = \mathbb{X}$, shows that $\widehat{\boldsymbol{Y}}$ is the BLUE of $\mathsf{E}\,\boldsymbol{Y}$. This produces the third version of the Gauss-Markov theorem.

Theorem 3.6 (Gauss-Markov, version III). *Let the linear regression model specified in Section 3.1 on page 30 be satisfied, let $\widehat{\boldsymbol{\beta}}$ be the LSE. Then $\widehat{Y}$ is the best linear unbiased estimator of* $\mathsf{E}\, Y$. ◇

3.3. Properties of the Least Squares Estimator Under Normality

In this section, we consider the linear regression model with the normality assumption. In particular,

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with

$$\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma_e^2 \mathbb{I}_n) \quad \text{or, equivalently,} \quad \boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2 \mathbb{I}_n).$$

The regression matrix $\mathbb{X}_{n\times p}$ still has a full rank $p < n$. All the results of the previous two sections are still valid. Under normality, we can derive additional results about distributions of various quantities, which are summarized in the following lemma.

Lemma 3.7. *Under the assumptions of the current section,*

(i) $\widehat{\boldsymbol{\beta}} \sim \mathsf{N}_p(\boldsymbol{\beta}, \sigma_e^2(\mathbb{X}^T\mathbb{X})^{-1})$*;*

(ii) $\widehat{\boldsymbol{Y}} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2 \mathbb{H})$*;*

(iii) $\boldsymbol{u} \sim \mathsf{N}_n(\mathbf{0}, \sigma_e^2(\mathbb{I} - \mathbb{H}))$*;*

(iv) $\dfrac{SS_e}{\sigma_e^2} \sim \chi^2_{n-p}$*;*

(v) $\widehat{\boldsymbol{\beta}}$ *and* SS_e *are independent.* ◇

Proof.

(i)–(iii) This is obvious: $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Y}}$, and $\boldsymbol{u}$ are just linear transformations of $\boldsymbol{Y}$. The first and second moments have been provided by Lemmas 3.1 and 3.2.

(iv) As shown in Section 2.4, $SS_e = \boldsymbol{\varepsilon}^\mathsf{T}(\mathbb{I} - \mathbb{H})\varepsilon$, where $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma_e^2 \mathbb{I}_n)$. Because $\mathbb{I} - \mathbb{H}$ is idempotent of rank $n - p$ it follows from Lemma A.2 in the Appendix that $SS_e/\sigma_e^2 \sim \chi^2_{n-p}$.

(v) We have $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} \equiv \mathbb{B}\boldsymbol{Y}$ and $\sigma_e^2 = \boldsymbol{Y}^\mathsf{T}(\mathbb{I} - \mathbb{H})\boldsymbol{Y} \equiv \boldsymbol{Y}\mathbb{A}\boldsymbol{Y}$. By Lemma A.3 in the Appendix it suffices to show that $\mathbb{B}\mathbb{A} = 0$. But

$$\mathbb{B}\mathbb{A} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}(\mathbb{I} - \mathbb{H}) = \mathbf{0}$$

because $(\mathbb{I} - \mathbb{H})\mathbb{X} = \mathbf{0}$. □

The linear regression model with normally distributed responses is a parametric model. Let us derive the maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ and σ_e^2.

We have $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2 \mathbb{I}_n)$ with unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \sigma_e^2)^\mathsf{T}$. The likelihood is

$$L(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \frac{1}{(2\pi)^{n/2}(\sigma_e^2)^{n/2}} \mathrm{e}^{-\frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})}$$

and the log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma_e^2 \mid \boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_e^2 - \frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta}).$$

Regardless of σ_e^2, to maximize this over $\boldsymbol{\beta}$ it is enough to minimize $\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|^2 = SS_e(\boldsymbol{\beta})$. So, the least squares estimator $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ in the normal linear regression model. Plug this into the log-likelihood to find the MLE of σ_e^2:

$$\ell(\widehat{\boldsymbol{\beta}}, \sigma_e^2 \mid \boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_e^2 - \frac{1}{2\sigma_e^2}SS_e.$$

Now,

$$\frac{\partial \ell(\widehat{\boldsymbol{\beta}}, \sigma_e^2 \mid \boldsymbol{Y})}{\partial \sigma_e^2} = -\frac{n}{2}\cdot\frac{1}{\sigma_e^2} + \frac{1}{2}\frac{SS_e}{(\sigma_e^2)^2}$$

The MLE solves the equation

$$\frac{n}{\sigma_e^2} = \frac{SS_e}{(\sigma_e^2)^2}$$

and the solution is SS_e/n. We have proven the following lemma.

Lemma 3.8. *In the normal linear regression model, the maximum likelihood estimator of $\boldsymbol{\beta}$ is the LSE $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\boldsymbol{Y})$ and the maximum likelihood estimator of σ_e^2 is SS_e/n.* ♢

Note. The MLE of σ_e^2 differs from the unbiased estimator $\widehat{\sigma}_e^2$ of Definition 3.2 by dividing SS_e with n instead of $n-1$. This difference becomes negligible as n increases.

Notation

Here we list symbols that are consistently used in the same meaning throughout the whole text (perhaps with a few exceptions). Symbols that are introduced and used locally (e.g., in one section) are usually not listed here.

$\boldsymbol{\varepsilon}$	column vector of error terms
$\mathbb{H}$	hat matrix
h_{ii}	the i-th diagonal element of the hat matrix $\mathbb{H}$
$\mathbf{1}_n$	column vector of ones of length n
$\mathbb{J}_n$	$\mathbf{1}_n\mathbf{1}_n^\top$, $n \times n$ matrix of ones
$\mathscr{M}(\mathbb{X})$	subspace generated by the columns of $\mathbb{X}$
$\mathscr{M}(\mathbb{X})^\perp$	subspace orthogonal to the columns of $\mathbb{X}$
R^2	coefficient of determination
$SS_e(\boldsymbol{\beta})$	sum of squares taken as a function of $\boldsymbol{\beta}$
SS_e	residual sum of squares (minimized over $\boldsymbol{\beta}$)
SS_R	regression sum of squares (centered)
SS_T	total sum of squares (centered)
$\boldsymbol{u}$	column vector of residuals
u_i^*	the i-th standardized residual
$\mathbb{X}$	regression matrix containing covariate vectors in rows
$\boldsymbol{Y}$	column vector of responses
$\widehat{\boldsymbol{Y}}$	column vector of fitted values

List of Figures

Bibliography

Aitken, A. C. (1936). IV.—on least squares and linear combination of observations, *Proceedings of the Royal Society of Edinburgh* **55**: 42–48.

Galton, F. (1886). Regression towards mediocrity in hereditary stature, *The Journal of the Anthropological Institute of Great Britain and Ireland* **15**: 246–263.

Gauss, C. F. (1821). *Theoria combinationis observationum erroribus minimis obnoxiae*, Heinrich Dieterich, Göttingen.

Legendre, A.-M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, F. Didot, Paris.

Markov, A. A. (1912). *Wahrscheinlichkeitsrechnung*, 2nd edn, Leipzig and Berlin.

Index

A. Appendix

The Appendix presents some useful results that are used in this course.

Lemma A.1. *Let $\boldsymbol{X}$ be any random vector of dimension n with mean $\boldsymbol{\mu}$ and finite variance matrix Σ. Let $\mathbb{A}$ be any $n \times n$ matrix. Then*

$$E\,\boldsymbol{X}^T \mathbb{A} \boldsymbol{X} = \boldsymbol{\mu}^T \mathbb{A} \boldsymbol{\mu} + \operatorname{tr}(\mathbb{A}\Sigma). \qquad \diamondsuit$$

Proof.

$$\begin{aligned}
\mathsf{E}\,\boldsymbol{X}^\mathsf{T}\mathbb{A}\boldsymbol{X} &= \mathsf{E}\,(\boldsymbol{X}-\boldsymbol{\mu}+\boldsymbol{\mu})^\mathsf{T}\mathbb{A}(\boldsymbol{X}-\boldsymbol{\mu}+\boldsymbol{\mu}) \\
&= \mathsf{E}\operatorname{tr}\left[(\boldsymbol{X}-\boldsymbol{\mu})^\mathsf{T}\mathbb{A}(\boldsymbol{X}-\boldsymbol{\mu})\right] + \mathsf{E}\,(\boldsymbol{X}-\boldsymbol{\mu})^\mathsf{T}\mathbb{A}\boldsymbol{\mu} + \mathsf{E}\,\boldsymbol{\mu}^\mathsf{T}\mathbb{A}(\boldsymbol{X}-\boldsymbol{\mu}) + \mathsf{E}\,\boldsymbol{\mu}^\mathsf{T}\mathbb{A}\boldsymbol{\mu} \\
&= \operatorname{tr}\left[\mathsf{E}\,(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\mathsf{T}\mathbb{A}\right] + 0 + 0 + \boldsymbol{\mu}^\mathsf{T}\mathbb{A}\boldsymbol{\mu} \\
&= \operatorname{tr}\left[(\mathsf{var}\,\boldsymbol{X})\mathbb{A}\right] + \boldsymbol{\mu}^\mathsf{T}\mathbb{A}\boldsymbol{\mu} = \boldsymbol{\mu}^\mathsf{T}\mathbb{A}\boldsymbol{\mu} + \operatorname{tr}(\mathbb{A}\Sigma). \qquad \square
\end{aligned}$$

Lemma A.2. *Let $\boldsymbol{X} \sim \mathsf{N}_n(\mathbf{0}, \Sigma)$. Let $\mathbb{A}$ be an $n \times n$ matrix such that $\mathbb{A}\Sigma$ is idempotent. Then*

$$\boldsymbol{X}^T \mathbb{A} \boldsymbol{X} = \chi^2_{\operatorname{tr}(\mathbb{A}\Sigma)}. \qquad \diamondsuit$$

Lemma A.3. *Let $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \Sigma)$. Then $\boldsymbol{X}^T \mathbb{A} \boldsymbol{X}$ and $\mathbb{B}\boldsymbol{X}$ are independent if and only if*

$$\mathbb{B}\Sigma\mathbb{A} = \mathbf{0}. \qquad \diamondsuit$$