NMSA407 Linear Regression

Lecture Notes

Michal Kulich

Last modified on September 9, 2025.

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 September 9, 2025

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.*

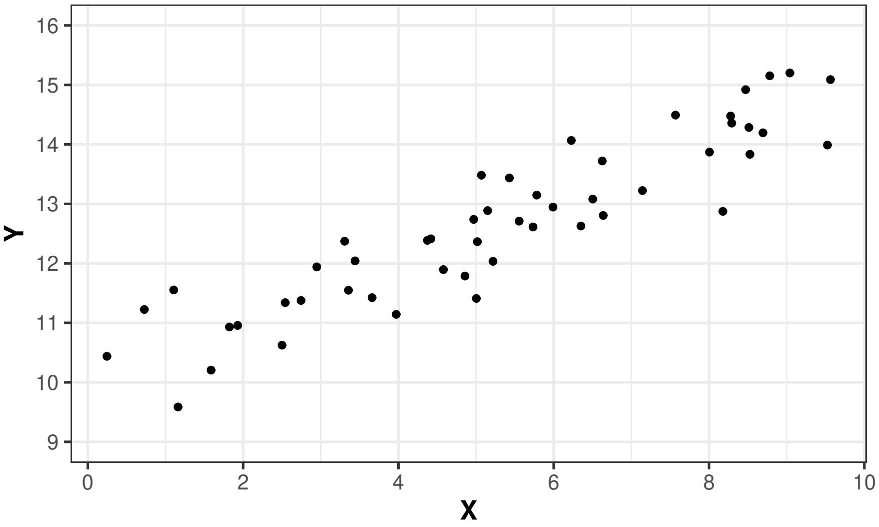


Figure 1.1.: Scatterplot of two continuous variables in $\mathbb{R}^2$.

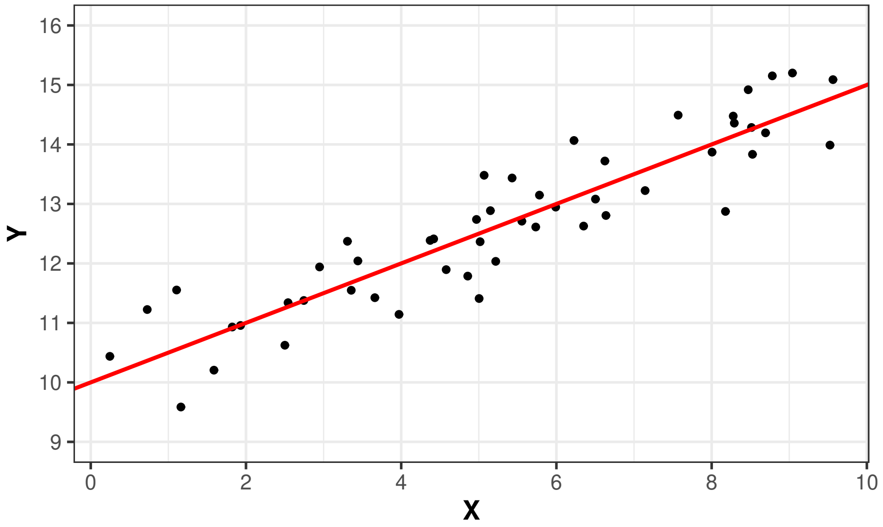


Figure 1.2.: Scatterplot of two continuous variables in $\mathbb{R}^2$ with fitted line.

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

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

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*.

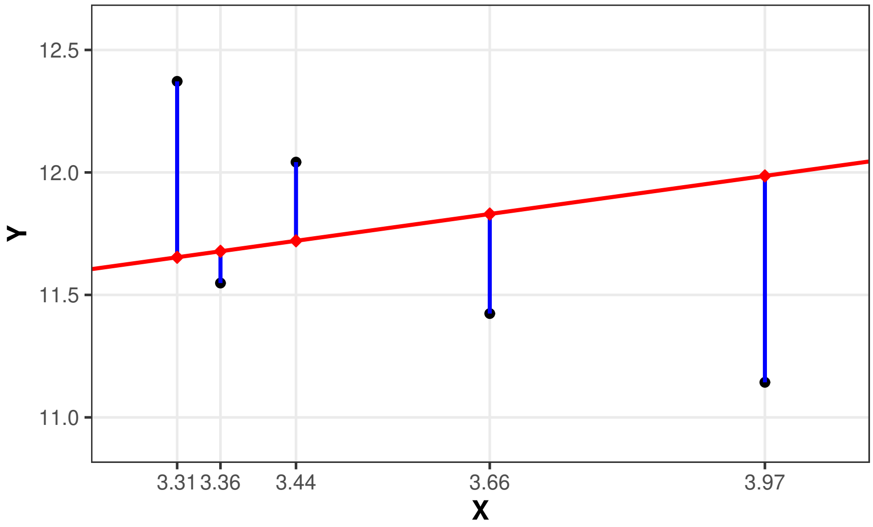


Figure 1.3.: Zoomed subset of data from Figure 1.2 with visualized vertical distances of the points from the line (blue).

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

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

Finally,

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

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).

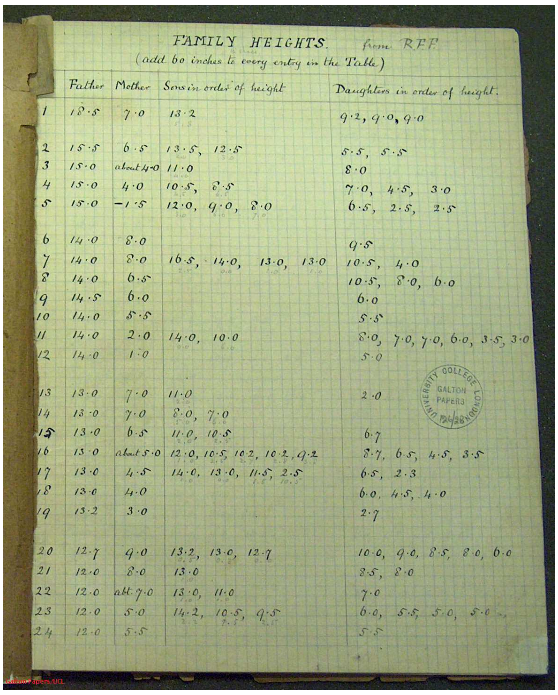
If the observations x_i have the same sample variance as y_i then the slope of the line fitted by least squares is equal to the sample correlation coefficient r_{xy} and therefore lies in the interval $\langle -1, 1\rangle$.

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 R.F.F.
(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

Galton Papers/UCL

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*.

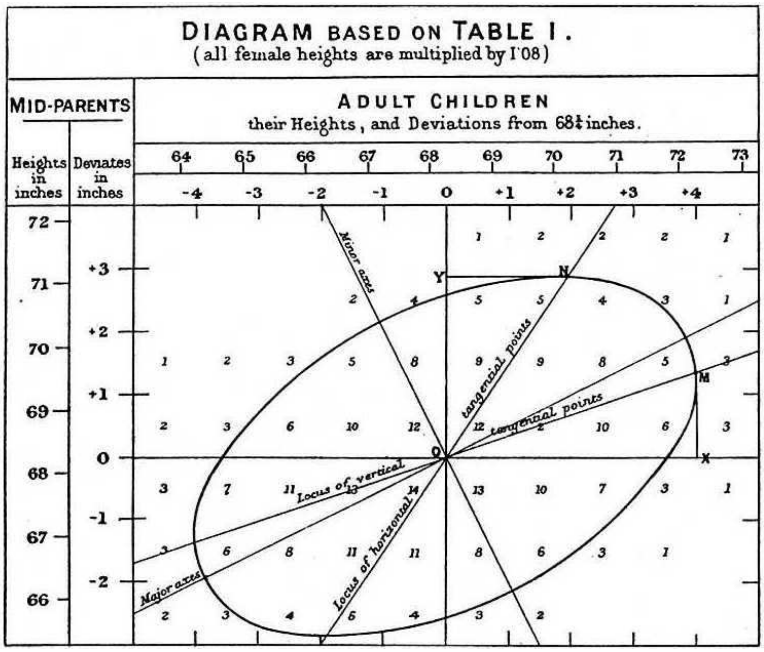


Figure 1.5.: Galton height data: original visualization by the author.
Source: https://en.wikipedia.org

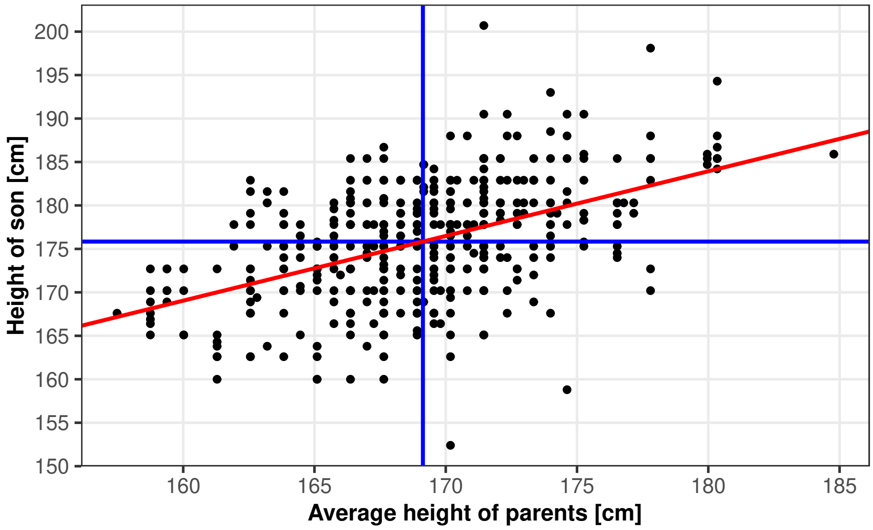


Figure 1.6.: Modified Galton data with fitted least squares line (red). The slope of the line is ≈ 0.74. The means of the two variables are plotted as blue lines.

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 $(Y_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$. 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 $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\top$ 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 $\boldsymbol{\beta} = (\beta_1, \dots \beta_p)^\top$ is a vector of unknown *regression parameters (coefficients)*[§] and *the error terms*[¶] $\varepsilon_1, \dots, \varepsilon_n$ 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*

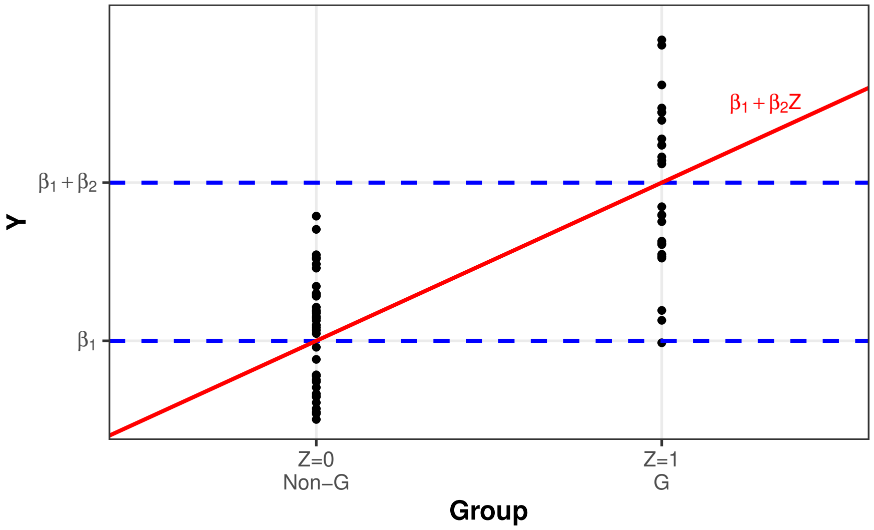


Figure 2.1.: Two sample problem expressed as a linear regression model $\operatorname{E} Y = \beta_1 + \beta_2 Z$, where $Z = \mathbb{1}(G)$. The regression line has no interpretation except at $Z = 0$ or $Z = 1$.

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

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

where ε_i, $i = 1, \dots, n$ 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 $\operatorname{E} Y_i = \beta_1 + \beta_2 Z_i$ 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 $Z_i = \mathbb{1}(i \in G)$.

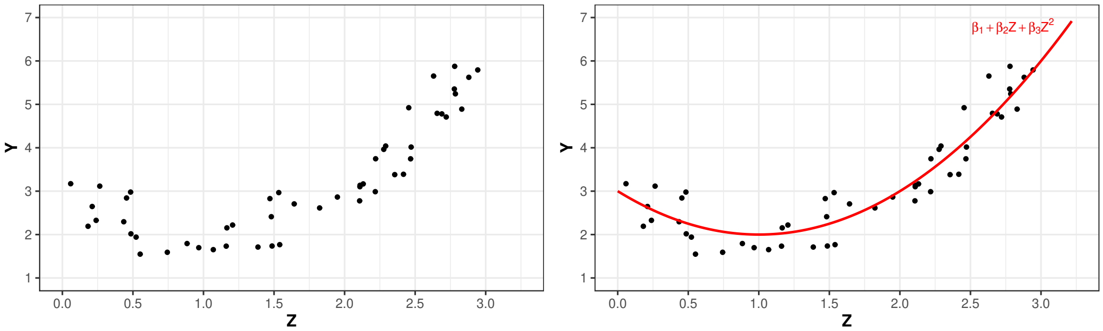


Figure 2.2.: Data following a quadratic association with a fitted quadratic curve.

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 $\neg G$). △

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 $\boldsymbol{e}_j = (0,\dots,0,1,0,\dots,0)^\top$ 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 $\beta_j X_{ij} = \frac{\beta_j}{m} \cdot m X_{ij}$. 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 $\mathsf{P}[X_i = 1] \equiv \pi \in (0,1)$. 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

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

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 $\mathsf{P}[r(\mathbb{X}) = p] \to 1$ 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 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 $\boldsymbol{Y}$.

(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

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

Note. Any symmetric idempotent matrix is positive semi-definite. *Prove this yourself.*

Geometric interpretation of the LSE

From Linear Algebra:

Consider a vector space V and two subspaces U and W such that $V=U\oplus W$. 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 $\mathbb{A}=(\boldsymbol{a}_1,\dots,\boldsymbol{a}_p)$ 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 $\mathcal{M}(\mathbb{X})$ be the linear subspace of $\mathbb{R}^n$ generated by the columns of the regression matrix $\mathbb{X}$ (denote them by $\boldsymbol{x}_j$, $j=1,\dots,p$):

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

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

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

Then

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

So, $\mathbb{H}$ and $\mathbb{I}-\mathbb{H}$ are projection matrices to the two orthogonal subspaces, $\mathcal{M}(\mathbb{X})$ and $\mathcal{M}(\mathbb{X})^{\perp}$, 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. 20, 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 $\boldsymbol{Y}$:

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

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 $Y_1, \dots, Y_n$ 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 $\operatorname{var} \varepsilon_i = \sigma_Y^2$. 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 $\mathbb{X} = (1, \ldots, 1)^\mathsf{T} \equiv \mathbf{1}_n$, $\boldsymbol{\beta} = \mu$.

Notation.

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

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

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

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 $SS_e = \boldsymbol{u}^\mathsf{T}\boldsymbol{u} = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$.

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 $j = 2, \ldots, p$. 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)

Take $\mathbb{J}_n = \mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, 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, $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}$. Then for all $i=1,\dots,n$,

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_1 + \widehat{\beta}_2X_{i2} + \dots + \widehat{\beta}_pX_{ip}\\ &= \widehat{\alpha} + \widehat{\beta}_2(X_{i2}-\overline{X}_2) + \dots + \widehat{\beta}_p(X_{ip}-\overline{X}_p),\end{aligned}$$

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) there is no change in either the regression parameter for that covariate or in its LSE.

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

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

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, $\boldsymbol{1}_n \in \mathscr{M}(\mathbb{X})$. 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

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

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, $\mathbb{H}\boldsymbol{1}_n = \boldsymbol{1}_n$, hence $\mathbb{H}\mathbb{J}_n = \mathbb{J}_n$, hence $(\mathbb{I}-\mathbb{H})\mathbb{J}_n = \boldsymbol{0}$. 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

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

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, $0 \leq R^2 \leq 1$.
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:

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

[*] 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 $\mathsf{var}\,Y_i$*.

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

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

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

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

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

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

This result will play an important role in Chapter 4.

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

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

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

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 $(\mathbb{I}-\mathbb{H})\mathbb{X} = \boldsymbol{0}$. Hence

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

Next,

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

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

$$\begin{aligned} \boldsymbol{Y} &= \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ \boldsymbol{Y} &= \mathbb{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{u} \end{aligned}$$

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.

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

Lemma 3.3. $\mathsf{E}\,SS_e = (n-p)\sigma_e^2$. ♢

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

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

Definition 3.1.

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

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 3.2. $\widehat{\boldsymbol{\theta}}$ 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 $\mathsf{E}\,\mathbb{B}\boldsymbol{Y} = \boldsymbol{\theta}$

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

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

Theorem 3.4 (Gauss-Markov, version I). *Let the linear regression model specified in Section 3.1 on page 31 be satisfied, let $\widehat{\boldsymbol{\beta}}$ be the LSE. Then $\boldsymbol{c}^\top\widehat{\boldsymbol{\beta}}$ is the unique best linear unbiased estimator of $\boldsymbol{c}^\top\boldsymbol{\beta}$ for any $\boldsymbol{0} \neq \boldsymbol{c} \in \mathbb{R}^p$.* ♢

Proof.

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

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

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

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

Finally,

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

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 $\boldsymbol{a}^\top = \boldsymbol{a}^\top \mathbb{H}$. It follows that the estimator $\boldsymbol{a}^\top \boldsymbol{Y}$ can be rewritten as

$$\boldsymbol{a}^\top \boldsymbol{Y} = \boldsymbol{a}^\top \mathbb{H} \boldsymbol{Y} = \boldsymbol{a}^\top \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{Y} = \boldsymbol{a}^\top \mathbb{X} \widehat{\boldsymbol{\beta}} = \boldsymbol{c}^\top \widehat{\boldsymbol{\beta}}. \qquad \square$$

Theorem 3.5 (Gauss-Markov, version II). *Let the linear regression model specified in Section 3.1 on page 31 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 $\mathbb{D} = \operatorname{var} \mathbb{A}\boldsymbol{Y} - \operatorname{var} \mathbb{C}\widehat{\boldsymbol{\beta}}$ and prove that $\mathbb{D} \geq 0$ by taking any non-zero vector $\boldsymbol{d} \in \mathbb{R}^q$ and showing that $\boldsymbol{d}^\top \mathbb{D} \boldsymbol{d} \geq 0$. We have $\boldsymbol{d}^\top \mathbb{D} \boldsymbol{d} = \operatorname{var} \boldsymbol{d}^\top \mathbb{A}\boldsymbol{Y} - \operatorname{var} \boldsymbol{d}^\top \mathbb{C}\widehat{\boldsymbol{\beta}}$. Also, $\mathsf{E}\,\boldsymbol{d}^\top \mathbb{A}\boldsymbol{Y} = \boldsymbol{d}^\top \mathbb{A}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{d}^\top \mathbb{C}\boldsymbol{\beta}$.

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

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 31 be satisfied, let $\widehat{\boldsymbol{\beta}}$ be the LSE. Then $\widehat{\boldsymbol{Y}}$ is the best linear unbiased estimator of $\mathsf{E}\,\boldsymbol{Y}$.* ♢

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) $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma_e^2(\mathbb{X}^T\mathbb{X})^{-1})$;

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

(iii) $\boldsymbol{u} \sim N_n(\boldsymbol{0}, \sigma_e^2(\mathbb{I}-\mathbb{H}))$;

(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 $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma_e^2\mathbb{I}_n)$. 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

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

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

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 $\boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2\mathbb{I}_n)$ 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:

$$\ell(\widehat{\boldsymbol{\beta}}, \sigma_e^2 \mid \boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_e^2 - \frac{1}{2\sigma_e^2}SS_e. \tag{3.1}$$

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 3.1 by dividing SS_e with n instead of $n-1$. This difference becomes negligible as n increases.

4. Statistical Inference in the Linear Regression Model

4.1. Exact Inference Under Normality

In this section, we work under the assumption of normality, when the regression model can be formulated as

$$\boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2 \mathbb{I}_n)$$

and we use the results of Section 3.3, in particular, Lemma 3.7. First, we formulate the exact distribution of the normalized linear combination of estimated regression coefficients.

Lemma 4.1. *Under the assumptions of the current section, for any* $\mathbf{c} \neq \mathbf{0}$*,*

$$\frac{\mathbf{c}^\top\widehat{\boldsymbol{\beta}} - \mathbf{c}^\top\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2 \mathbf{c}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbf{c}}} \sim t_{n-p}.$$

t_{n-p} *is the Student's t-distribution with* $n-p$ *degrees of freedom.* ♢

Proof. By Lemma 3.7, part (i),

$$\mathbf{c}^\top\widehat{\boldsymbol{\beta}} \sim \mathsf{N}(\mathbf{c}^\top\boldsymbol{\beta}, \sigma_e^2 \mathbf{c}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbf{c})$$

and hence

$$\frac{\mathbf{c}^\top\widehat{\boldsymbol{\beta}} - \mathbf{c}^\top\boldsymbol{\beta}}{\sigma_e\sqrt{\mathbf{c}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbf{c}}} \sim \mathsf{N}(0,1).$$

By Lemma 3.7, parts (iv) and (v), $\dfrac{SS_e}{\sigma_e^2} \sim \chi^2_{n-p}$ and $\widehat{\boldsymbol{\beta}}$ and SS_e are independent. It follows that

$$\frac{\frac{\mathbf{c}^\top\widehat{\boldsymbol{\beta}} - \mathbf{c}^\top\boldsymbol{\beta}}{\sigma_e\sqrt{\mathbf{c}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbf{c}}}}{\sqrt{\frac{SS_e}{\sigma_e^2(n-p)}}} = \frac{\mathbf{c}^\top\widehat{\boldsymbol{\beta}} - \mathbf{c}^\top\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2 \mathbf{c}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbf{c}}} \sim t_{n-p}.$$

□

We can use this lemma to perform tests and construct confidence intervals for any linear combinations of regression coefficients. For example, if we take $\mathbf{c} = \boldsymbol{e}_j$, we get $\mathbf{c}^\top\widehat{\boldsymbol{\beta}} = \widehat{\beta}_j$ and $\mathbf{c}^\top\boldsymbol{\beta} = \beta_j$. Next, $\mathbf{c}^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbf{c} \equiv v_{jj}$, where v_{jj} denotes the j-th diagonal element of $(\mathbb{X}^\top\mathbb{X})^{-1}$.

4.1.1. Testing individual regression parameters

Consider the hypothesis $H_0 : \beta_j = a$ against the two-sided alternative $H_0 : \beta_j \neq a$ (usually, we take $a = 0$). Based on Lemma 4.1, we reject H_0 if

$$\frac{\left|\widehat{\beta}_j - a\right|}{\sqrt{\widehat{\sigma}_e^2 v_{jj}}} \geq t_{n-p}\big(1-\frac{\alpha}{2}\big),$$

where $t_{n-p}(1-\alpha/2)$ is the $(1-\alpha/2)$-quantile of t distribution with $n-p$ degrees of freedom. Note that $\widehat{\sigma}_e^2 v_{jj}$ is the estimated variance of $\widehat{\beta}_j$. This test has the exact level α.

4.1.2. Confidence intervals for individual regression parameters

Let β_j be the true value of the j-th regression parameter and $\widehat{\beta}_j$ be the LSE of β_j. By Lemma 4.1,

$$\mathrm{P}\left[-t_{n-p}\big(1-\frac{\alpha}{2}\big) \leq \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}_e^2 v_{jj}}} \leq t_{n-p}\big(1-\frac{\alpha}{2}\big)\right] = 1-\alpha.$$

By a simple manipulation, we get

$$\mathrm{P}\left[\widehat{\beta}_j - t_{n-p}\big(1-\frac{\alpha}{2}\big)\sqrt{\widehat{\sigma}_e^2 v_{jj}} \leq \beta_j \leq \widehat{\beta}_j + t_{n-p}\big(1-\frac{\alpha}{2}\big)\sqrt{\widehat{\sigma}_e^2 v_{jj}}\right] = 1-\alpha.$$

Thus,

$$\widehat{\beta}_j \mp t_{n-p}\big(1-\frac{\alpha}{2}\big)\sqrt{\widehat{\sigma}_e^2 v_{jj}}$$

are the boundary points of a confidence interval for β_j with coverage probability exactly $1-\alpha$.

4.1.3. Tests and confidence intervals for linear combinations of regression parameters

Choose the desired $\boldsymbol{c}$ and use Lemma 4.1 in the same way as in Sections 4.1.1 and 4.1.2.

4.1.4. Simultaneous tests of several linear combinations of regression parameters

Consider a matrix of constants $\mathbb{C}_{q\times p}$ with $q \leq p$ and $r(\mathbb{C}) = q$. By Theorem 3.5 (Gauss-Markov, ver. II), $\mathbb{C}\widehat{\boldsymbol{\beta}}$ is a best linear unbiased estimator of $\mathbb{C}\boldsymbol{\beta}$ even the data are not normal. The next lemma provides the exact distribution of $\mathbb{C}\widehat{\boldsymbol{\beta}}$ under normality.

Lemma 4.2. *Under the assumptions of the current section, for any $\mathbb{C}_{q\times p}$ with $q \le p$ and $r(\mathbb{C}) = q$,*

$$\frac{1}{q\widehat{\sigma}_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \mathbb{C}\boldsymbol{\beta})^T\left[\mathbb{C}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{C}^T\right]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \mathbb{C}\boldsymbol{\beta}) \sim F_{q,n-p}. \tag{4.1}$$

$F_{q,n-p}$ is the Fisher's F-distribution with q and $n-p$ degrees of freedom. ♢

Proof. By Lemma 3.7, part (i),

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

and hence

$$\frac{1}{\sigma_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \mathbb{C}\boldsymbol{\beta})^\mathsf{T}\left[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}\right]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \mathbb{C}\boldsymbol{\beta}) \sim \chi_q^2.$$

By Lemma 3.7, parts (iv) and (v),

$$\frac{SS_e}{\sigma_e^2} \sim \chi_{n-p}^2$$

and $\widehat{\boldsymbol{\beta}}$ and SS_e are independent. Hence

$$\frac{\frac{1}{q\sigma_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \mathbb{C}\boldsymbol{\beta})^\mathsf{T}\left[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}\right]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \mathbb{C}\boldsymbol{\beta})}{\frac{SS_e}{(n-p)\sigma_e^2}} \sim F_{q,n-p}.$$

This proves the lemma. □

This lemma can be used to perform simultaneous tests of several linear combinations of regression parameters. Put the desired coefficients into the q rows of the matrix $\mathbb{C}$ and consider testing

$$H_0 : \mathbb{C}\boldsymbol{\beta} = \boldsymbol{c} \qquad \text{against} \qquad H_1 : \mathbb{C}\boldsymbol{\beta} \neq \boldsymbol{c},$$

where $\boldsymbol{c}$ is a q-vector of constants (often zeros). The hypothesis is rejected when

$$\frac{1}{q\widehat{\sigma}_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})^\mathsf{T}[\mathbb{C}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{C}^\mathsf{T}\big]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}) \ge F_{q,n-p}(1-\alpha),$$

where $F_{q,n-p}(1-\alpha)$ is the $(1-\alpha)$-quantile of the F distribution with q and $n-p$ degrees of freedom. This test has the exact level α.

The test statistic can be expressed in a more convenient form if we realize that the hypothesis H_0 specifies a set of linear constraints on the regression parameters. We know from section 2.11, equation (2.8) that the numerator of the test statistic can we written as

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

where $SS_e(\widehat{\boldsymbol{\beta}}_C)$ is the residual sum of squares under linear constraints, that is, under H_0, and SS_e is the residual sum of squares without restrictions, i.e., when H_0 is not assumed to hold.

This consideration allows us to summarize the previous results in the following way.

Lemma 4.3. *When the hypothesis $H_0 : \mathbb{C}\boldsymbol{\beta} = \mathbf{c}$ is true, where $\mathbb{C}$ is a $q \times p$ matrix of constants with $q \leq p$ and $r(\mathbb{C}) = q$, and $\mathbf{c}$ is a q-vector of constants, then*

$$\frac{n-p}{q}\frac{SS_e^0 - SS_e}{SS_e} \sim F_{q,n-p},$$

where SS_e^0 is the residual sum of squares calculated under H_0 and SS_e is the residual sum of squares calculated without any restrictions.

The hypothesis is rejected if

$$\frac{n-p}{q}\frac{SS_e^0 - SS_e}{SS_e} \geq F_{q,n-p}(1-\alpha).$$

This test has the exact level α. ♢

The lemma provides a much easier formulation of the test, which can be further extended to any submodel testing (see the next section). The numerator of the test statistic expresses how much SS_e increased under H_0 relative to what it was when H_0 was not assumed to hold.

From Section 2.9, the numerator can be rewritten in terms of regression sums of squares as well. In the model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \varepsilon$, we decompose the total sum of squares as $SS_T = SS_R + SS_e$. Under the null hypothesis, the same total sum of squares is decomposed as $SS_T = SS_R^0 + SS_e^0$. The numerator

$$SS_e^0 - SS_e = (SS_T - SS_R^0) - (SS_T - SS_R) = SS_R - SS_R^0$$

shows how much the regression sum of squares improved after removing the restriction imposed by the null hypothesis.

4.1.5. Submodel testing

Consider two different regression models for the same response $\boldsymbol{Y}$ (with $q \geq 1$):

$$\begin{aligned} &\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ with } \mathbb{X}_{n\times p}, \boldsymbol{\beta}_{p\times 1} && \text{Model } (M) \\ \text{and } &\boldsymbol{Y} = \mathbb{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^*, \text{ with } \mathbb{Z}_{n\times(p-q)}, \boldsymbol{\gamma}_{(p-q)\times 1} && \text{Model } (M_1) \end{aligned}$$

Model M_1 is called *a submodel* of Model M if and only if $\mathscr{M}(\mathbb{Z}) \subset \mathscr{M}(\mathbb{X})$, that is, the linear space generated by the columns of $\mathbb{Z}$ is a subspace of the linear space generated by the columns of $\mathbb{X}$. The submodel explains the response through fewer covariates and fewer parameters. Whenever the submodel M_1 is true the larger model M must also be true.

Because $\mathscr{M}(\mathbb{Z}) \subset \mathscr{M}(\mathbb{X})$, each column of $\mathbb{Z}$ can be expressed as a linear combination of the columns of $\mathbb{X}$. Thus, there exists a $p \times (p-q)$ matrix $\mathbb{A}$ such that $\mathbb{Z} = \mathbb{X}\mathbb{A}$. Take any matrix $\mathbb{B}_{p\times q}$ such that $(\mathbb{A}|\mathbb{B})$ is of full rank. Then $(\mathbb{X}\mathbb{A}|\mathbb{X}\mathbb{B})$ is another basis of $\mathscr{M}(\mathbb{X})$. Therefore, $\mathbb{X}\boldsymbol{\beta}$ can be written as

$$\mathbb{X}\boldsymbol{\beta} = \mathbb{X}\mathbb{A}\boldsymbol{\gamma} + \mathbb{X}\mathbb{B}\boldsymbol{\delta} = \mathbb{Z}\boldsymbol{\gamma} + \mathbb{X}\mathbb{B}\boldsymbol{\delta}$$

Model M is equivalent to the model

$$\boldsymbol{Y} = \mathbb{Z}\boldsymbol{\gamma} + \mathbb{X}\mathbb{B}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

and the submodel M_1 is true if and only if $\boldsymbol{\delta} = \mathbf{0}$. Thus, any submodel can be obtained by linear restriction testing applied on a model that is equivalent to the larger model M.

This proves that the test specified in Lemma 4.3 can be applied to any submodel testing problem. To be specific, let

SS_e^0	be the residual sum of squares under the submodel;
SS_e	be the residual sum of squares under the larger model;
q	be the difference in the number of parameters between the larger model and the submodel.

If the submodel holds then

$$\frac{n-p}{q}\,\frac{SS_e^0 - SS_e}{SS_e} \sim F_{q,n-p}.$$

The submodel is rejected in favor of the larger model if

$$\frac{n-p}{q}\,\frac{SS_e^0 - SS_e}{SS_e} \geq F_{q,n-p}(1-\alpha).$$

Submodel testing is the most important tool for building regression models, that is, for deciding which covariates should be included in the model and in what functional form.

4.1.6. Overall regression test

Let us consider a special case of submodel (or linear restriction) testing. Take $\mathbb{C} = (\mathbf{0}|\mathbb{I}_{p-1})$ and test $H_0 : \mathbb{C}\boldsymbol{\beta} = \mathbf{0}$. This is equivalent to $\beta_2 = \cdots = \beta_p = 0$, that is, all regression coefficients except the intercept are zero. The number of tested parameters is $q = p-1$. The submodel contains only the intercept. When this hypothesis is true, the responses have the same expectation, which does not (linearly) depend on any of the covariates.

Recall the decomposition of centered sums of squares derived in Section 2.9. We have $SS_T = SS_R + SS_e$, or explicitely

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

Under H_0 (intercept only), we have $\widehat{Y}_i^0 = \overline{Y}$ for all i, $SS_R^0 = 0$, $SS_e^0 = SS_T$. The test statistic for testing this hypothesis is

$$\frac{n-p}{p-1}\frac{SS_e^0 - SS_e}{SS_e} = \frac{n-p}{p-1}\frac{SS_R}{SS_e}.$$

This can be also expressed as

$$\frac{n-p}{p-1}\frac{SS_R}{SS_T - SS_R} = \frac{n-p}{p-1}\frac{R^2}{1-R^2},$$

where R^2 is the coefficient of determination (see Section 2.10). The hypothesis that no covariates affect the expectation can be rejected if this test statistic exceeds $F_{p-1,n-p}(1-\alpha)$.

The calculation of the overall regression test is traditionally visualized in the form of so called analysis-of-variance (ANOVA) table, see Table 4.1.

4.1.7. One-way analysis of variance model

Let the observed data be (Y_i, Z_i) independent pairs, $i = 1, \ldots, n$, where Y_i is the response and $Z_i \in \{1, 2, \ldots, m\}$ classifies the subjects into one of m disjoint groups. Let the expectation of the response in the j-th group be $\mu_j = \mathsf{E}\left[Y_i \middle| Z_i = j\right]$, $j = 1, \ldots, m$. Let the conditional distribution of the response in the j-th group be $\mathsf{N}(\mu_j, \sigma_e^2)$, that is, all observations have normal distributions with potentially different means in the m groups and equal variances. This is the classical one-way analysis of variance (ANOVA) model.

The one-way ANOVA model can be formulated as a linear regression model by

$$\begin{aligned} Y_i &= \sum_{j=1}^{m} \mu_j \mathbb{1}(Z_i = j) + \varepsilon_i, \qquad \varepsilon_i \sim \mathsf{N}(0, \sigma_e^2) \\ &= \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta} + \varepsilon_i, \end{aligned}$$

where $\boldsymbol{X}_i = \boldsymbol{e}_j$ when $Z_i = j$ and $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_m)^\mathsf{T}$.

Denote the group sizes by $n_j = \sum_{i=1}^{n} \mathbb{1}(Z_i = j)$. Sort the observations so that the n_1 observations coming from group 1 are listed first, followed by the n_2 observations belonging

Source of variation	SS	d.f.	MS	F
All covariates	SS_R	$p-1$	$MS_R = \frac{SS_R}{p-1}$	$F = \frac{MS_R}{MS_e}$
Error	SS_e	$n-p$	$MS_e = \frac{SS_e}{n-p}$	
Total	SS_T	$n-1$		

Table 4.1.: Analysis of variance table for the overall regression test.

to group 2 and so on. Consider the corresponding regression matrix $\mathbb{X}$. The least squares estimator of $\boldsymbol{\beta}$ is

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

We have

$$\mathbb{X}^\mathsf{T}\mathbb{X} = \operatorname{diag}(n_1,\dots,n_m) \quad \text{and} \quad \mathbb{X}^\mathsf{T}\boldsymbol{Y} = \Big(\sum_{i=1}^n Y_i \mathbb{1}(Z_i = 1),\dots,\sum_{i=1}^n Y_i \mathbb{1}(Z_i = m)\Big).$$

Hence $\widehat{\boldsymbol{\beta}} = (\overline{Y}_1,\dots,\overline{Y}_m)$, where $\overline{Y}_j$ is the arithmetic average of the observations belonging to the j-th group. This is also the least squares estimator of the expectation μ_j in the j-th group.

The fitted values in the one-way ANOVA model are

$$\widehat{Y}_i = \sum_{j=1}^m \overline{Y}_j \mathbb{1}(Z_i = j),$$

the regression sum of squares is

$$SS_R = \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^n\sum_{j=1}^m (\overline{Y}_j - \overline{Y})^2 \mathbb{1}(Z_i = j) = \sum_{j=1}^m n_j(\overline{Y}_j - \overline{Y})^2.$$

In classical ANOVA, this is also denoted by SS_A. The residual sum of squares is

$$SS_e = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^n\sum_{j=1}^m (Y_i - \overline{Y}_j)^2 \mathbb{1}(Z_i = j).$$

Consider the hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ that all the groups have the same mean. This is equivalent to $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m$. Under this hypothesis, the data can be described by an intercept-only model and the test of this hypothesis is the overall regression test constructed in the previous section (with $p = m$). The test statistic of the overall regression test is

$$F = \frac{n-m}{m-1}\frac{SS_R}{SS_e} = \frac{SS_R/(m-1)}{SS_e/(n-m)}$$

The hypothesis is rejected if $F \geq F_{m-1,n-m}(1-\alpha)$. This the classical one-way ANOVA F-test. We have derived it as a special case of an overall regression test in a linear model.

4.1.8. Connections to maximum likelihood theory

Separate the regression parameter into $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{\beta}_1$ has $p-q$ elements and $\boldsymbol{\beta}_2$ q elements. Consider the hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. This corresponds to a submodel test and also to the test of the hypothesis $H_0 : \mathbb{C}\boldsymbol{\beta} = \mathbf{0}$ with $\mathbb{C} = (\mathbf{0}|\mathbb{I}_q)_{q\times p}$.

By Lemma 4.2 and Lemma 4.3, the test statistic for this test can be expressed as

$$\frac{n-p}{q}\frac{SS_e^0 - SS_e}{SS_e} = \frac{1}{q\widehat{\sigma}_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}\left[\mathbb{C}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{C}^{\mathsf{T}}\right]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}}) \overset{H_0}{\sim} F_{q,n-p}.$$

Let us now consider the likelihood ratio test of the same hypothesis. First, consider σ_e^2 known. The likelihood ratio statistic is

$$LR = 2[\ell(\widehat{\boldsymbol{\beta}}) - \ell(\widetilde{\boldsymbol{\beta}})],$$

where $\widetilde{\boldsymbol{\beta}}$ is the MLE (which is the same as the LSE) of $\boldsymbol{\beta}$ calculated under the submodel. Let $\widetilde{\boldsymbol{Y}}$ be the fitted values in the submodel. The maximum likelihood theory stipulates that $LR \xrightarrow{\mathsf{D}} \chi_q^2$ when the submodel is true.

From (3.1), the log-likelihoods of the larger model and the submodel are

$$\ell(\widehat{\boldsymbol{\beta}}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_e^2 - \frac{1}{2\sigma_e^2}SS_e,$$
$$\ell(\widetilde{\boldsymbol{\beta}}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma_e^2 - \frac{1}{2\sigma_e^2}SS_e^0.$$

Thus,

$$LR = \frac{1}{\sigma_e^2}(SS_e^0 - SS_e) = \frac{1}{\sigma_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}\left[\mathbb{C}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{C}^{\mathsf{T}}\right]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}}) \overset{H_0}{\sim} \chi_q^2$$

and its distribution under H_0 is exact, not only asymptotic.

With unknown σ_e^2, we modify the LR test as follows:

$$F = \frac{\frac{1}{q}LR}{\frac{(n-p)\widehat{\sigma}_e^2}{\sigma_e^2}/(n-p)} \overset{H_0}{\sim} F_{q,n-p}.$$

So, the F-test for submodel testing is equivalent to the likelihood ratio test. With some more effort we could prove that also the Wald test and Rao score test yield the same test statistic.

4.2. Asymptotic Inference Without Normality (Random Covariates)

In this section, we show that all the results derived in Section 4.1 under the assumption of normality can be extended to the general case as long as certain moment conditions are fulfilled. The results become asymptotic, though.

We assume that covariates are random and $(Y_i, \boldsymbol{X}_i)$ are *a random sample of independent identically distributed vectors* drawn from some $(p+1)$-variate distribution.

Let the data satisfy the linear 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 $\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$ and $\mathsf{var}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \sigma_e^2$.

We assume finite second moments of the response and the covariates and linear independence of the components of the covariate vector. We still assume that $r(\mathbb{X}) = p$[*]. No additional assumptions are imposed on the distributions of Y_i or ε_i.

Assumption.

(AS1) $\mathsf{var}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \sigma_e^2 < \infty$;

(AS2) $\mathsf{E}_X \boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T} \equiv \mathbb{V}_X < \infty$;

(AS3) $\mathbb{V}_X > 0$ (full rank regular invertible matrix).

E_X denotes the expectation over the marginal distribution of the covariates.

In this section, we need to distinguish three kinds of expectations. The notation we will use is as follows.

$$\mathsf{E}_{(Y,X)} h(Y,\boldsymbol{X}) = \int h(y,\boldsymbol{x}) f(y,\boldsymbol{x})\, d\mu(y,\boldsymbol{x})$$

is the expectation with respect to the joint distribution of $(Y,\boldsymbol{X})$. The joint density with respect to the measure μ is denoted by $f(y,\boldsymbol{x})$.

$$\mathsf{E}\, h(Y,\boldsymbol{X}) = \mathsf{E}\left[h(Y,\boldsymbol{X}) \,\middle|\, \boldsymbol{X}\right]$$

is the conditional expectation given the covariates.

$$\mathsf{E}_X h(\boldsymbol{X}) = \int h(\boldsymbol{x}) f(\boldsymbol{x})\, d\nu(\boldsymbol{x})$$

is the expectation with respect to the marginal distribution of $\boldsymbol{X}$. The marginal density with respect to the measure ν is denoted by $f(\boldsymbol{x})$.

With this notation, we have

$$\begin{aligned} \mathsf{E}_{(Y,X)} h(Y,\boldsymbol{X}) &= \mathsf{E}_X\left[\mathsf{E}\, h(Y,\boldsymbol{X})\right], \\ \mathsf{var}_{(Y,X)} h(Y,\boldsymbol{X}) &= \mathsf{E}_X \mathsf{var}\, h(Y,\boldsymbol{X}) + \mathsf{var}_X \mathsf{E}\, h(Y,\boldsymbol{X}). \end{aligned}$$

We know from Lemma 3.1 that the least squares estimator $\widehat{\boldsymbol{\beta}}$ is unbiased under these circumstances. The LSE

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

is the unique solution to the system of normal equations $\mathbb{X}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\mathsf{T}\boldsymbol{Y}$ stated in (2.1).

[*] See the note on page 18.

We can write

$$\mathbb{X}^\mathsf{T}\mathbb{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \begin{pmatrix} \boldsymbol{X}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{X}_n^\mathsf{T} \end{pmatrix} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T}$$

and $\mathbb{X}^\mathsf{T}\boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{X}_i Y_i$. Notice that by the weak law of large numbers, $\frac{1}{n}\mathbb{X}^\mathsf{T}\mathbb{X} \xrightarrow{\mathsf{P}} \mathbb{V}_X$. The normal equations can be rewritten as $\mathbb{X}^\mathsf{T}\boldsymbol{Y} - \mathbb{X}^\mathsf{T}\mathbb{X}\widehat{\boldsymbol{\beta}} = 0$. For any $\boldsymbol{\beta} \in \mathbb{R}^p$, define

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

Take a single term from the sum and denote it by

$$\boldsymbol{U}_i(\boldsymbol{\beta}) \equiv \boldsymbol{X}_i (Y_i - \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}).$$

The LSE $\widehat{\boldsymbol{\beta}}$ is the single solution to the system of equations $\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{U}_i(\boldsymbol{\beta}) = \boldsymbol{0}$. Thus, $\boldsymbol{U}(\boldsymbol{\beta})$ plays the role of the score statistic, except that we do not have a parametric model and so cannot derive the score statistic from the likelihood. This kind of an ad-hoc score statistic is sometimes called a *pseudoscore*.

The next theorem shows that under the current assumptions the LSE $\widehat{\boldsymbol{\beta}}$ defined by this particular pseudoscore is a consistent and asymptotically normal estimator of the true $\boldsymbol{\beta}$.

Theorem 4.4 (Asymptotic properties of the LSE without normality). *Under the assumptions of the current section, when $\boldsymbol{\beta}$ denotes the true regression parameters,*

(i) $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ *($\widehat{\boldsymbol{\beta}}$ is consistent);*

(ii) $\dfrac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma_e^2 \mathbb{V}_X)$*;*

(iii) $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma_e^2 \mathbb{V}_X^{-1})$. ♢

Note. Rewrite point (iii) to see the approximate distribution of $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\dot\sim \mathsf{N}_p(\boldsymbol{0}, \sigma_e^2 \mathbb{V}_X^{-1}) \\ \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\dot\sim \mathsf{N}_p(\boldsymbol{0}, \sigma_e^2 (\frac{1}{n}\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}) \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &\dot\sim \mathsf{N}_p(\boldsymbol{0}, \sigma_e^2 (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}) \\ \widehat{\boldsymbol{\beta}} &\dot\sim \mathsf{N}_p(\boldsymbol{\beta}, \sigma_e^2 (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}) \end{aligned}$$

The exact distribution of $\widehat{\boldsymbol{\beta}}$ under normality is **exactly the same** as the approximate distribution of $\widehat{\boldsymbol{\beta}}$ without assuming normality.

Proof (of Theorem 4.4).

(i)

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i Y_i\right)$$

By the weak law of large numbers and continuous transformation theorem,

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T} \xrightarrow{\mathsf{P}} \mathsf{E}_{\boldsymbol{X}}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T} = \mathbb{V}_X, \quad \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\right)^{-1} \xrightarrow{\mathsf{P}} \mathbb{V}_X^{-1}$$

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i Y_i \xrightarrow{\mathsf{P}} \mathsf{E}_{(Y,X)}\boldsymbol{X}_i Y_i = \mathsf{E}_{\boldsymbol{X}}(\boldsymbol{X}_i \mathsf{E}\, Y_i) = \mathsf{E}_{\boldsymbol{X}}(\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}) = \mathbb{V}_X\boldsymbol{\beta}.$$

Hence, $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \mathbb{V}_X^{-1}\mathbb{V}_X\boldsymbol{\beta} = \boldsymbol{\beta}$.

(ii) $\boldsymbol{U}_i(\boldsymbol{\beta}) = \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta})$ are independent identically distributed random vectors. Calculate their first and second moments at the true $\boldsymbol{\beta}$, noticing that $\boldsymbol{U}_i(\boldsymbol{\beta}) = \boldsymbol{X}_i\varepsilon_i$.

$$\mathsf{E}_{(Y,X)}\boldsymbol{U}_i(\boldsymbol{\beta}) = \mathsf{E}_{\boldsymbol{X}}\mathsf{E}\,\boldsymbol{X}_i\varepsilon_i = \boldsymbol{0},$$

$$\mathsf{var}_{(Y,X)}\boldsymbol{U}_i(\boldsymbol{\beta}) = \mathsf{E}_{\boldsymbol{X}}\mathsf{var}\,\boldsymbol{U}_i(\boldsymbol{\beta}) + \mathsf{var}_X\,\underbrace{\mathsf{E}\,\boldsymbol{U}_i(\boldsymbol{\beta})}_{=\boldsymbol{0}} = \mathsf{E}_{\boldsymbol{X}}\mathsf{var}\,(\boldsymbol{X}_i\varepsilon_i) = \mathsf{E}_{\boldsymbol{X}}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\sigma_e^2 = \mathbb{V}_X\sigma_e^2.$$

By the central limit theorem for iid random vectors,

$$\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{U}_i(\boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, \sigma_e^2\mathbb{V}_X).$$

(iii) Consider the difference between $\boldsymbol{U}(\boldsymbol{\beta})$ evaluated at the true $\boldsymbol{\beta}$ and at the LSE $\widehat{\boldsymbol{\beta}}$.

$$\boldsymbol{U}(\boldsymbol{\beta}) - \underbrace{\boldsymbol{U}(\widehat{\boldsymbol{\beta}})}_{=\boldsymbol{0}} = \sum_{i=1}^{n}\left[\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}) - \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}})\right] = \sum_{i=1}^{n}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right].$$

Next,

$$\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta}) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\right)\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}\right)}_{\xrightarrow{\mathsf{P}} \mathbb{V}_X}{}^{-1} \quad \underbrace{\frac{1}{\sqrt{n}}\boldsymbol{U}(\boldsymbol{\beta})}_{\xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, \sigma_e^2\mathbb{V}_X)}.$$

By Slutsky's Theorem,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, \sigma_e^2\underbrace{\mathbb{V}_X^{-1}\mathbb{V}_X\mathbb{V}_X^{-1}}_{=\mathbb{V}_X^{-1}}).$$

□

Note. Consistence of $\widehat{\boldsymbol{\beta}}$ can be proven directly, as shown in the proof of Theorem 4.4(i) but it also follows as a direct consequence of Theorem 4.4(iii). It is easy to see that if $\widehat{\boldsymbol{\beta}}$ did not converge in probability to $\boldsymbol{\beta}$ then $\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ cannot converge in distribution.

By Lemma 3.3, $\widehat{\sigma}_e^2 = \frac{SS_e}{n-p}$ is an unbiased estimator of the residual variance σ_e^2. Now we need to show that this estimator is consistent.

Lemma 4.5. *Under the assumptions of the current section,* $\widehat{\sigma}_e^2 \xrightarrow{P} \sigma_e^2$. ◇

Proof.

$$\widehat{\sigma}_e^2 = \underbrace{\frac{n}{n-p}}_{\to 1}\frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})^2.$$

Further,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})^2 &= \frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} + \boldsymbol{X}_i^\top \boldsymbol{\beta} - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})^2 \\ &= \frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 + \frac{1}{n}\sum_{i=1}^n [\boldsymbol{X}_i^\top(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})]^2 + \left[\frac{2}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})\boldsymbol{X}_i^\top\right](\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\end{aligned}$$

The first term converges in probability to $\mathsf{E}\,\varepsilon_i^2 = \mathsf{var}\,\varepsilon_i = \sigma_e^2$. The second term can be written as

$$\underbrace{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^\top}_{\xrightarrow{\mathsf{P}}\mathbf{0}}\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)}_{\xrightarrow{\mathsf{P}}\mathbb{V}_X}\underbrace{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}_{\xrightarrow{\mathsf{P}}\mathbf{0}};$$

therefore, it converges to zero in probability. The third term also converges to zero in probability, because $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}$ and

$$\frac{2}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})\boldsymbol{X}_i^\top \xrightarrow{\mathsf{P}} 2\mathsf{E}\,\varepsilon_i \boldsymbol{X}_i^\top = \mathbf{0}.$$

This completes the proof. □

Now we are ready to restate the key results from Section 4.1 in their asymptotic versions. Let us start with the asymptotic version of Lemma 4.1.

Lemma 4.6. *Under the assumptions of the current section, for any* $\mathbf{c} \neq \mathbf{0}$,

$$\frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2 \mathbf{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{c}}} \xrightarrow{D} N(0,1).$$ ◇

Proof. By Theorem 4.4(iii),

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &\xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0},\sigma_e^2\mathbb{V}_X^{-1})\\ \sqrt{n}(\mathbf{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}-\mathbf{c}^\mathsf{T}\boldsymbol{\beta}) &\xrightarrow{\mathsf{D}} \mathsf{N}(0,\sigma_e^2\mathbf{c}^\mathsf{T}\mathbb{V}_X^{-1}\mathbf{c})\\ \frac{\mathbf{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}-\mathbf{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\sigma_e^2\mathbf{c}^\mathsf{T}(n\mathbb{V}_X)^{-1}\mathbf{c}}} &\xrightarrow{\mathsf{D}} \mathsf{N}(0,1)\end{aligned}$$

Replace σ_e^2 by $\widehat{\sigma}_e^2$ and $\mathbb{V}_X$ by $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}$. By Lemma 4.5 and Slutsky's Theorem, this does not change the limiting distribution. □

Note. Lemma 4.1 states that the distribution of the left-hand side is exactly t_{n-p} under normality. The limiting distribution in Lemma 4.6 is standard normal. Because the distribution function of t_{n-p} converges to the distribution function of the standard normal distribution as $n \to \infty$, Lemma 4.1 can be used as an asymptotic approximation when the responses are not normally distributed. Thus, the methods for testing and for constructing confidence intervals introduced in Sections 4.1.1 and 4.1.2 can be used with non-normal responses if the sample size is large enough.

Next, we formulate an asymptotic version of Lemma 4.2.

Lemma 4.7. *Under the assumptions of the current section, for any $\mathbb{C}_{q\times p}$ with $q \le p$ and $r(\mathbb{C}) = q$,*

$$\frac{1}{\widehat{\sigma}_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta})^T\big[\mathbb{C}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{C}^T\big]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta}) \xrightarrow{D} \chi_q^2 \quad \text{as } n\to\infty.$$ ◇

Proof. By Theorem 4.4(iii),

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &\xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0},\sigma_e^2\mathbb{V}_X^{-1})\\ \sqrt{n}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta}) &\xrightarrow{\mathsf{D}} \mathsf{N}_q(\mathbf{0},\sigma_e^2\mathbb{C}\mathbb{V}_X^{-1}\mathbb{C}^\mathsf{T})\\ \frac{1}{\sigma_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta})^\mathsf{T}\big[\mathbb{C}(n\mathbb{V}_X)^{-1}\mathbb{C}^\mathsf{T}\big]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta}) &\xrightarrow{\mathsf{D}} \chi_q^2.\end{aligned}$$

Replace σ_e^2 by $\widehat{\sigma}_e^2$ and $\mathbb{V}_X$ by $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T}$. By Lemma 4.5 and Slutsky's Theorem, this does not change the limiting distribution. □

Note. Denote the left-hand side of the expression (4.1) in Lemma 4.2 by F. Lemma 4.2 claims that $F \sim F_{q,n-p}$ under normality. It follows that $qF \xrightarrow{\mathsf{D}} \chi_q^2$ (which is exactly the claim of Lemma 4.7). Thus the claim of Lemma 4.2 can be considered as an asymptotic approximation when the responses are not normal. The same is true for Lemma 4.3.

When we test a submodel against a larger model and reject the submodel if

$$\frac{n-p}{q}\,\frac{SS_e^0 - SS_e}{SS_e} \geq F_{q,n-p}(1-\alpha),$$

the test has the exact level α when the responses are normal and a level that converges to α if the responses are not normal and $n \to \infty$.

All the results derived in Sections 4.1.1 – 4.1.7 under the assumption of normality hold asymptotically even if normality is violated, as long as the assumptions (AS1)–(AS3) given on page 45 hold. Therefore, normality of Y_i or ε_i should not be considered a necessary condition for the validity of results obtained by linear regression analyses. Normality only makes asymptotic results exact.

Note. When the number of observations is large enough, we do not care whether the responses are normal or not. How large is "large enough", though? The answer depends on the complexity of the model and the degree of violation of normality. In a simple linear regression model, 25 observations may be enough for the asymptotic results to provide acceptable approximation even if the responses are strongly non-normal. In complex models with many covariates and high-order interactions, we may need hundreds or thousands of observations.

Note. Trusting regression results obtained on small datasets is dangerous in both situations. If the data are not normal, asymptotics cannot be relied upon. The only possibility to proceed with the analysis is to *assume* that the data are normal. However, on a small dataset we cannot verify this assumption and therefore we cannot trust the results either.

4.3. Asymptotic Inference Without Normality (Fixed Covariates)

In the previous section, we have shown that the results derived under the assumption of normality can be used as asymptotic approximations when the covariates are random and satisfy certain moment conditions. In that setup, it is particularly easy to prove the asymptotic results because the data form a sequence of independent and identically distributed random vectors.

However, in certain applications such as industrial experiments the covariates cannot be considered random because their values are pre-determined by the experimenter and set to the desired values.[*] In this section, we will state conditions under which the claims of the previous sections remain valid even if the covariates are constant. The proofs will be omitted, though.

[*] Imagine evaluating the effect of temperature on the performance of a certain product. The temperature is set by the investigator to equidistant values such as 5°, 10°, 15°, 20° etc. These values are definitely not random.

We assume that covariates $\boldsymbol{x}_i$ are fixed vectors of constants. The observations are $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where Y_i are independent random variables.

Suppose the data follow the linear model

$$Y_i = \boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta} + \varepsilon_i$$

where $\mathsf{E}\,\varepsilon_i = 0$ and $\mathsf{var}\,\varepsilon_i = \sigma_e^2 < \infty$. Let the regression matrix be of full rank. The distribution of Y_i is otherwise arbitrary.

Redefine $\mathbb{V}_X = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$. Huber (1973) formulated the following condition for the validity of asymptotic results.

Condition (Huber's condition). The largest diagonal element of $\mathbb{H} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ converges to zero as $n \to \infty$.

Proposition 4.8 (Huber 1973). *Under the assumptions of the current section, when $\boldsymbol{\beta}$ denotes the true regression parameters,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma_e^2 \mathbb{V}_X^{-1})$$

if and only if Huber's condition holds. ◇

Huber's condition is sufficient to prove consistency of $\widehat{\boldsymbol{\beta}}$ and necessary and sufficient to prove normality. The proof relies on Feller-Lindeberg central limit theorem and is omitted.

Arnold (1980) showed that Lemmas 4.6 and 4.7 also hold if Huber's condition is fulfilled. Thus, analysis of the linear regression model with fixed covariates proceeds exactly in the same way as with random covariates.

5. Predictions

5.1. Predictions and Their Pitfalls

Obtaining predictions for an individual observation based on the observed values of the covariates is one of the common goals of regression analysis. Actually, regression analysis may have a number of possible objectives — each of them requires somewhat different approach to development and evaluation of the model. Among the possible objectives are, e.g.,

1. Separate the signal $\boldsymbol{X}_i^\top \boldsymbol{\beta}$ from the noise ε_i.
2. Predict the *expectation* of a future observation with known covariates.
3. Predict the *value* of a future observation with known covariates.
4. Determine the functional shape of $m(\boldsymbol{x}) = \mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right]$.
5. Find out which covariates affect the expectation of the response and evaluate their influence.
6. Evaluate the influence of a single specific covariate on the expectation of the response.
7. *Et cetera.*

Objectives 1–3 are related to making predictions of the response. A point prediction can be easily obtained by taking the fitted value $\widehat{Y}$, which is the best linear unbiased estimator of the true expectation. However, one must be extremely careful not to make predictions for covariate values that are outside the scope of the covariates that were used to build the model. The validity of the model can be verified only for covariate values that were present in the data. Predicting responses outside the scope of the covariates is called *extrapolation*. Extrapolation represents one the most frequent abuses of regression models in practice.

Example 5.1. Consider the situation illustrated in Figure 5.1. There is one covariate X and the true conditional expectation is $\mathsf{E}\left[Y \,\middle|\, X = x\right] = m(x) = x + 2$ for $x \leq 3$ and

$$m(x) = \frac{\exp\{2.5x - 7.5\}}{1 + \exp\{2.5x + 7.5\}} + 4.5 \quad \text{for } x \geq 3.$$

The conditional expectation is linear up to $x = 3$, then the speed of increase slows down and the expectation approaches the limit $m(x) = 5.5$ as x increases to infinity.[*] The observed covariates range from 0 to 3. The regression model captures well the linear part of $m(x)$ over the interval $(0, 3)$ but it cannot recognize that the relationship changes for $x > 3$. If the linear model is extrapolated to obtain predictions for $x > 3$, the predictions will be seriously biased upwards. △

[*] We can call this function a logistic pipe curve.

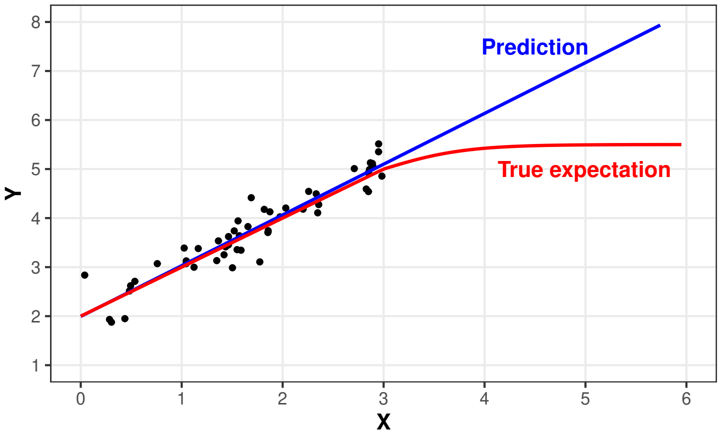


Figure 5.1.: Incorrect extrapolation of a logistic pipe curve beyond the range of data.

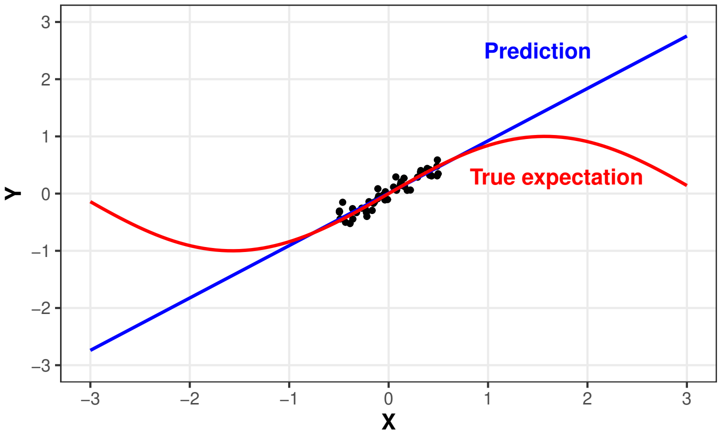


Figure 5.2.: Incorrect extrapolation of a sinus curve beyond the range of data.

Example 5.2. Figure 5.2 shows the results of extrapolation when the true expectation follows the function $m(x) = \sin x$ and the regression model is linear, with the covariate observed within the interval $(-0.5, 0.5)$. Inside that interval, the fitted regression line approximates the sinus function relatively well. Outside that interval, however, the predictions obtained from the regression line are worthless. △

The only case when extrapolation is allowed is the situation when the true shape of the function $m(x)$ is known (there is some physical law that determines that shape without any uncertainty) and we are certain that the estimate of $m(x)$ applies to covariate values that are beyond the scope of the data.

Example 5.3. The problem of extrapolation may be difficult to spot when multiple covari-

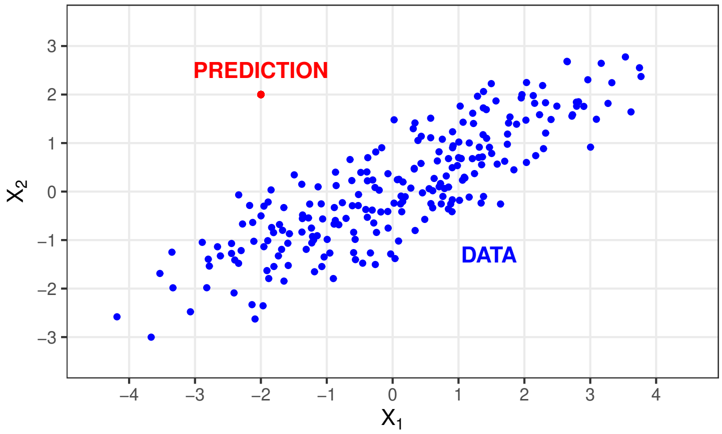


Figure 5.3.: Extrapolation beyond the scope of data: two covariates X_1 and X_2, the prediction is made within the range of both but outside the area where observations are available.

ates are present in the model. Figure 5.3 shows observed values of two covariates X_1 and X_2 (the response is not shown). The first covariate has values in the interval $(-4.2, 4)$, the other covariate lies within $(-3, 3)$. We intend to make a prediction at $x_1 = -2$ and $x_2 = 2$. Even though both covariates are within the ranges represented in the data, this particular combination is out of the scope of the observed pairs and thus the prediction at this point suffers from the extrapolation issue. △

Note. When presenting the results of regression models it is important to include detailed description of the covariate values that were used to fit the model. Otherwise the estimated regression parameters could be misused to obtain extrapolated predictions at covariate values which are far from the observations represented in the data.

5.2. Confidence Intervals for Conditional Expectations

Consider a vector of covariates $\boldsymbol{X} = \boldsymbol{x}$, which is in the scope of data (so that we avoid the extrapolation mistake). We want to estimate $\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right]$ for this particular covariate vector. The point estimate is $\widehat{Y} = \boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}$, and it is the BLUE by Gauss-Markov Theorem 3.6.

Let us construct a confidence interval for the unknown $\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$ to capture the uncertainty in the prediction appropriately. By Lemma 4.1,

$$U = \frac{\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2 \boldsymbol{x}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\boldsymbol{x}}} \sim t_{n-p}$$

under normality. By Lemma 4.6, the same statement holds approximately for non-normal data. Now,

$$\mathsf{P}\Big[-t_{n-p}(1-\alpha/2) < U < t_{n-p}(1-\alpha/2)\Big] = 1-\alpha$$

and, after an easy manipulation,

$$\mathsf{P}\Big[\boldsymbol{x}^{\top}\widehat{\boldsymbol{\beta}} - t_{n-p}(1-\alpha/2)\widehat{\sigma}_e\sqrt{\boldsymbol{x}^{\top}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{x}} < \boldsymbol{x}^{\top}\boldsymbol{\beta} < \\ < \boldsymbol{x}^{\top}\widehat{\boldsymbol{\beta}} + t_{n-p}(1-\alpha/2)\widehat{\sigma}_e\sqrt{\boldsymbol{x}^{\top}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{x}}\Big] = 1-\alpha$$

For non-normal data, this claim holds asymptotically, as $n \to \infty$.

Thus, the confidence interval for $\mathsf{E}\big[Y\,\big|\boldsymbol{X} = \boldsymbol{x}\big] = \boldsymbol{x}^{\top}\boldsymbol{\beta}$ with coverage probability $1-\alpha$ (exact or asymptotic) is

$$\boldsymbol{x}^{\top}\widehat{\boldsymbol{\beta}} \mp t_{n-p}(1-\alpha/2)\widehat{\sigma}_e\sqrt{\boldsymbol{x}^{\top}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{x}}.$$

Notice that for an observation that was used to fit the model, when $\boldsymbol{x} = \boldsymbol{X}_i$ for some i, $\boldsymbol{X}_i^{\top}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{X}_i = h_{ii}$, the i-th diagonal element of the projection matrix $\mathbb{H}$.

5.3. Prediction Intervals for Future Responses

Consider a future observation with a known vector of covariates $\boldsymbol{X} = \boldsymbol{x}$, which is in the scope of data. We want to construct an interval that includes the future observed response of such an observation with a desired probability, i.e., find C_L and C_U such that

$$\mathsf{P}\left[C_L < Y < C_U\right] = 1-\alpha.$$

Because we want the interval to cover a realization of a random variable rather than an unknown fixed quantity, we call the interval a *prediction interval* rather than a confidence interval.

The construction of the interval must be based on the distribution of a single observation Y and that cannot be approximated by asymptotic results. There fore we *must assume* in this section that the observation follows the normal distribution, in particular

$$Y \sim \mathsf{N}(\boldsymbol{x}^{\top}\boldsymbol{\beta}, \sigma_e^2).$$

Write Y as $Y = \boldsymbol{x}^{\top}\boldsymbol{\beta} + \varepsilon$ where $\varepsilon \sim \mathsf{N}(0, \sigma_e^2)$. We assume that the new observation Y is independent of the observations $Y_1, \dots, Y_n$ contained in the dataset used to estimate the parameters. Calculate the fitted value $\widehat{Y} = \boldsymbol{x}^{\top}\widehat{\boldsymbol{\beta}}$. By Lemma 3.7, part (i),

$$\widehat{Y} \sim \mathsf{N}(\boldsymbol{x}^{\top}\boldsymbol{\beta}, \sigma_e^2\boldsymbol{x}^{\top}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{x})$$

and, by independence and normality of Y and $\widehat{Y}$,

$$\widehat{Y} - Y \sim \mathsf{N}(0, \sigma_e^2 + \sigma_e^2\boldsymbol{x}^{\top}(\mathbb{X}^{\top}\mathbb{X})^{-1}\boldsymbol{x}).$$

Hence,

$$\frac{\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}} - Y}{\widehat{\sigma}_e\sqrt{1+\boldsymbol{x}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\boldsymbol{x}}} \sim t_{n-p}$$

and the resulting prediction interval for Y with coverage probability $1-\alpha$ is

$$\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}} \mp t_{n-p}(1-\alpha/2)\widehat{\sigma}_e\sqrt{1+\boldsymbol{x}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\boldsymbol{x}}. \tag{5.1}$$

It differs from the interval derived in the previous section by adding 1 under the square root and by the necessity to assume normality.

6. Diagnostic Methods Based on Residuals

In this section we introduce and illustrate residual-based methods for checking model assumptions and assessing the validity of the model.

Lemma 3.2 assures that the i-th residual $u_i = Y_i - \widehat{Y}_i = Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}$ has zero mean and variance $\operatorname{var} u_i = \sigma_e^2(1-h_{ii})$, where h_{ii} is the i-th diagonal element of the projection matrix $\mathbb{H} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$. Because raw residuals do not have the same variance, we will use so called *standardized residuals*.

Definition 6.1.

$$u_i^* = \frac{u_i}{\widehat{\sigma}_e\sqrt{1-h_{ii}}}$$

are called *standardized residuals*. ∇

If the model is valid and all assumptions are fulfilled, standardized residuals have approximately zero mean and unit variance. This can be verified by plotting standardized residuals in various ways. Common examples are:

- Scatterplots of residuals against various continuous variables, smoothed by some nonparametric smoother to facilitate recognition of patterns.
- Boxplots of residuals for specific subgroups of observations.
- Histograms and Q-Q plots of residuals.

Scatterplot of residuals against order of observation

This type of plot puts standardized residuals on the vertical axis and observation order i on the horizontal axis. It is useful when the ordering of observations has some real meaning, for example, if it captures the time sequence in which the observations are recorded. If the assumptions are satisfied, the plot shows just a random cloud of points centered around the line $y = 0$. If the plot suggests some effect of order on the residuals there is a suspicion that the observation order has some unaccounted effect on the response and the data are not truly independent.

The top panel of Figure 6.1 shows the situation when the assumptions are fulfilled. The data are generated from the model $Y_i = 1 + X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathsf{N}(0, 0.25)$, $i = 1, \ldots, 100$. The fitted model was $\mathsf{E}\, Y_i = \beta_1 + \beta_2 X_i$. The model is correct and there is no recognizable pattern

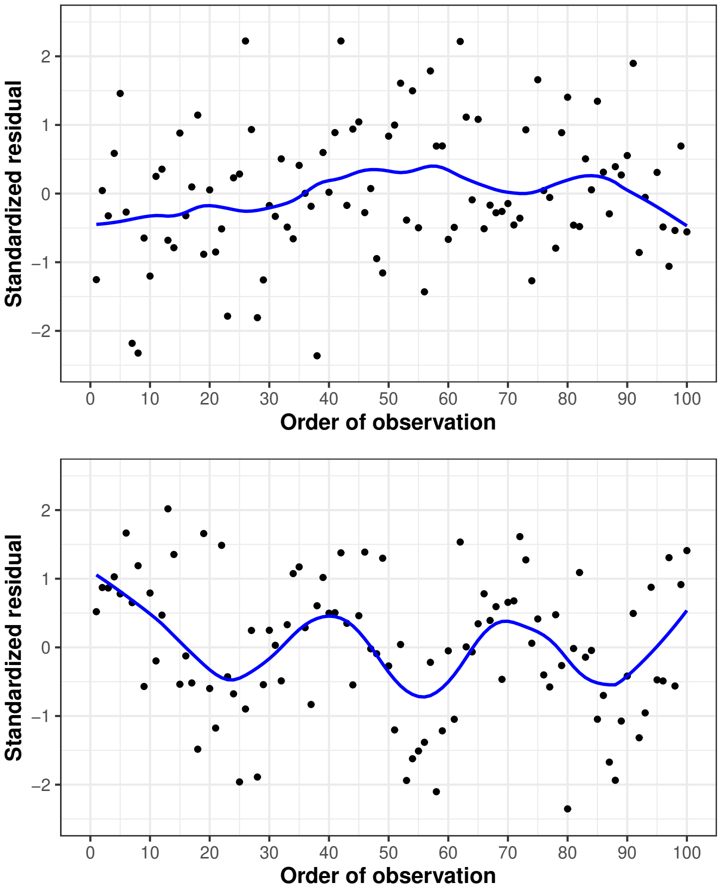


Figure 6.1.: Standardized residuals against observation order. Top panel: all assumptions are satisfied. Bottom panel: an uncaptured periodic effect. Both plots were smoothed by lowess smoother with window over 1/4 of the data range (blue).

in the top panel. The bottom panel shows residuals from the same linear model for data generated from the model $Y_i = 1 + X_i + \sin(i/5) + \varepsilon_i$. A periodic effect of the observation order was added to the responses but the analysis ignored that and proceeded in the same way as above. The omitted periodic effect is clearly demonstrated by the smoothed scatterplot in the bottom panel.

Scatterplot of residuals against fitted values

This version plots fitted values $\widehat{Y}_i$ on the horizontal axis. It provides kind of general assessment of the validity of the model. If the assumptions are satisfied, the plot shows just a random cloud of points centered around the line $y = 0$. If the plot suggests some pattern across the fitted values, there is a suspicion that the effect of some covariate on the response is modeled inappropriately.

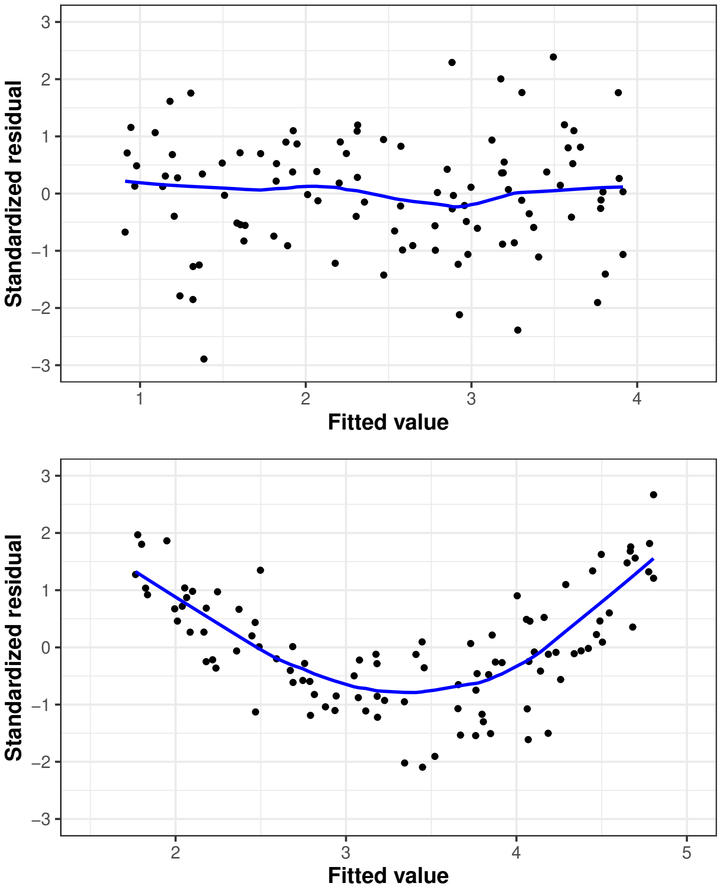


Figure 6.2.: Standardized residuals against fitted values. Top panel: all assumptions are satisfied. Bottom panel: omitted quadratic effect. Both plots were smoothed by lowess smoother with window over 1/2 of the data range (blue).

The top panel of Figure 6.2 shows the situation when the assumptions are fulfilled. The data are again generated from the model $Y_i = 1 + X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathsf{N}(0, 0.25)$, $i = 1, \ldots, 100$. The fitted model was $\mathsf{E}\, Y_i = \beta_1 + \beta_2 X_i$. The model is correct and there is no recognizable pattern in the top panel. The bottom panel shows residuals from the same linear model for data generated from the model $Y_i = 1 + X_i + (X_i - 1.5)^2 + \varepsilon_i$. A quadratic effect of the covariate was added to the responses but the analysis ignored that and proceeded in the same way as above. The omitted quadratic effect is clearly demonstrated by the smoothed scatterplot in the bottom panel.

When the model includes several covariates, the plot of residuals against fitted values may not reflect clearly that one of the covariates was modeled in an inappropriate way or to determine which covariate it was. For checking more complex models it is much better to plot the residuals against individual covariates rather than against fitted values.

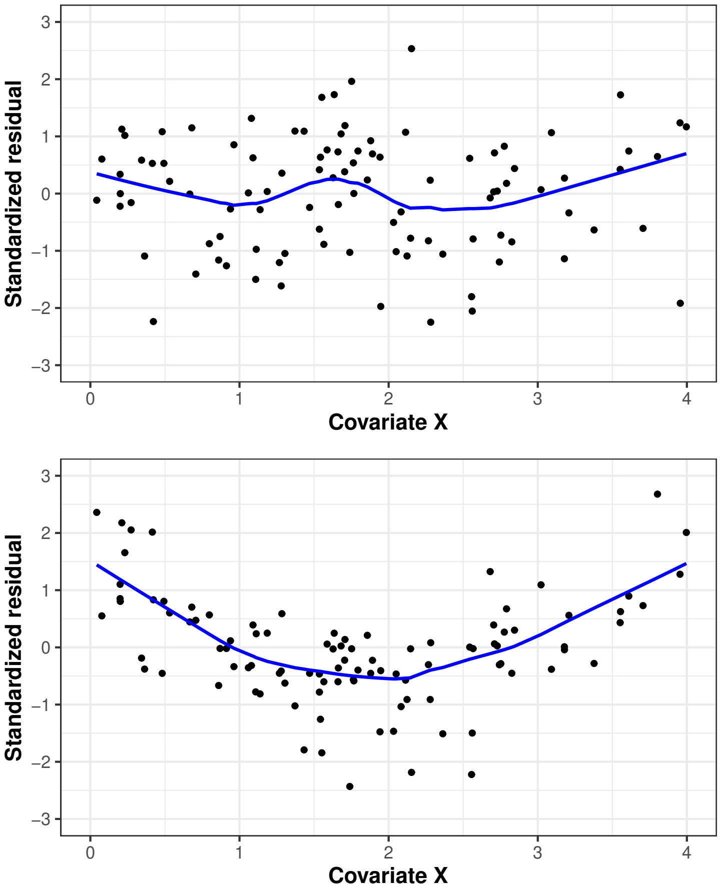


Figure 6.3.: Standardized residuals against a covariate. Top panel: all assumptions are satisfied. Bottom panel: omitted quadratic effect. Smoothed by lowess smoother with window over 1/2 of the data range (blue).

Scatterplot of residuals against continuous covariates

This plot uses values of a particular covariate on the horizontal axis. It checks whether the covariate was included in the model in an appropriate form. If it was and other assumptions are also satisfied, the plot shows just a random cloud of points centered around the line $y = 0$. If the plot suggests some pattern depending on the covariate, there is a suspicion that the effect of this covariate on the response is modeled inappropriately.

The top panel of Figure 6.3 shows the situation when the assumptions are fulfilled. The data are generated from the model $Y_i = 1 + Z_i + 0.5X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathsf{N}(0, 1)$, $i = 1, \ldots, 100$. The covariate Z_i is binary, the covariate X_i is continuous and correlated with Z_i. The fitted model was correct: $\mathsf{E}\, Y_i = \beta_1 + \beta_2 Z_i + \beta_3 X_i$. There is no recognizable pattern in the top panel, which plots the standardized residuals against the values of X_i. The bottom panel shows residuals from the same model but for data generated from the model $Y_i = 1 + Z_i +$

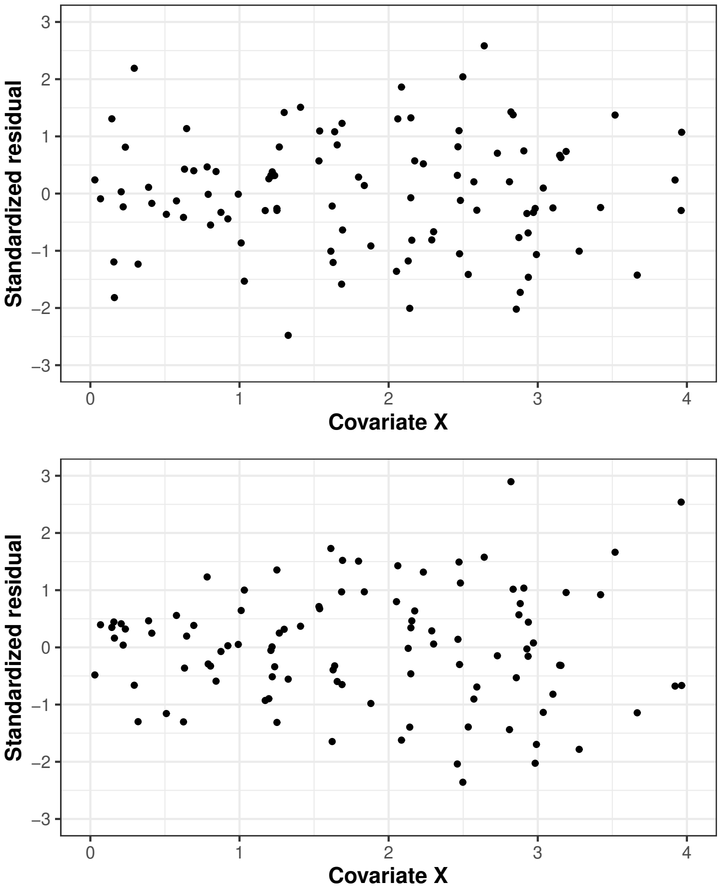


Figure 6.4.: Standardized residuals against a covariate. Top panel: all assumptions are satisfied. Bottom panel: mild increase of residual variance with covariate X.

$0.5X_i + (X_i - 1.5)^2 + \varepsilon_i$. A quadratic effect of the covariate X_i was added to the response but the analysis ignored that and proceeded with an incorrect model. The omitted quadratic effect is clearly demonstrated on the smoothed scatterplot in the bottom panel.

The plot of the residuals against a covariate is able to capture an inappropriately modeled effect of that covariate even if the model contains other covariates. Even better way to determine a suitable functional format for a continuous covariate provides the plot of partial residuals, which will be discussed later in this section.

This plot can be also used to check whether the variance is constant or whether it changes with the values of the covariate. If the variance is constant and other assumptions are also satisfied, the plot shows points that are about equally spread across the range of values of the covariate. If the residuals seem more variable in some regions than in others, there is a suspicion that the assumption of equal variance is not true.

The top panel of Figure 6.4 shows the situation when the assumptions are fulfilled. The

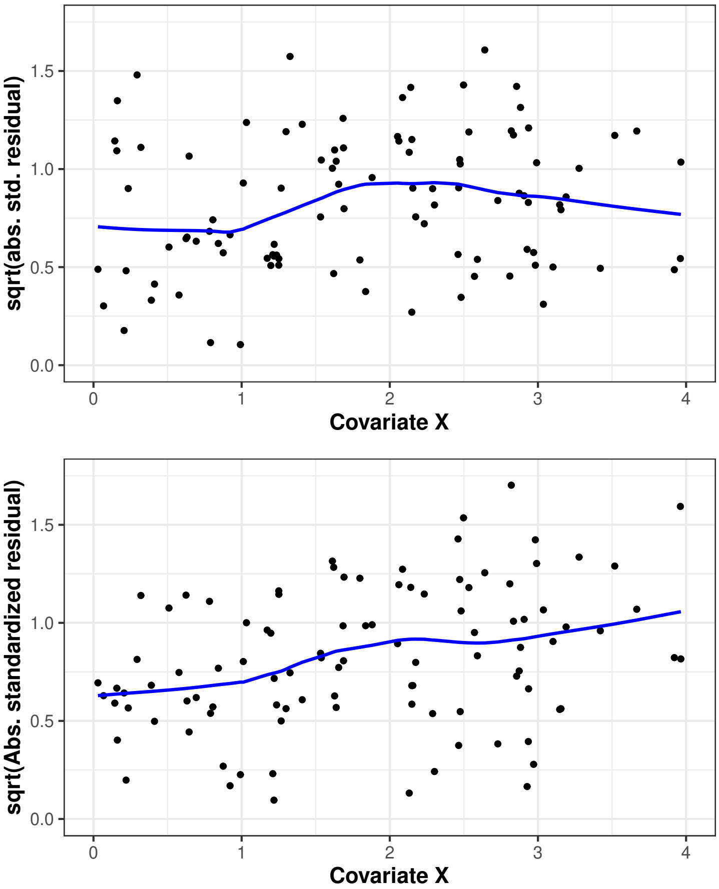


Figure 6.5.: Square root of absolute standardized residuals against a covariate. Top panel: all assumptions are satisfied. Bottom panel: mild increase of residual variance with covariate X.

data are generated from the model $Y_i = 1 + Z_i + 0.5X_i + (X_i - 1.5)^2 + \varepsilon_i$ with $\varepsilon_i \sim \mathsf{N}(0, 1)$, $i = 1, \ldots, 100$ (as before). This time, the correct model with quadratic term was used to perform the analysis and no assumptions are violated. We can see in the top panel, that the points are about equally spread in the vertical direction at all values of X_i. The bottom panel shows residuals from the same model but with unequal variance depending on X_i — $\mathsf{var}\, \varepsilon_i = (X_i + 1)^2/9$. The variance increases with X_i and the spread of the residuals seems to slightly increase with the value of X_i.

From this type of plot, it is relatively difficult to see whether the variance changes or not, and it is not possible to seek assistance from a smoother.

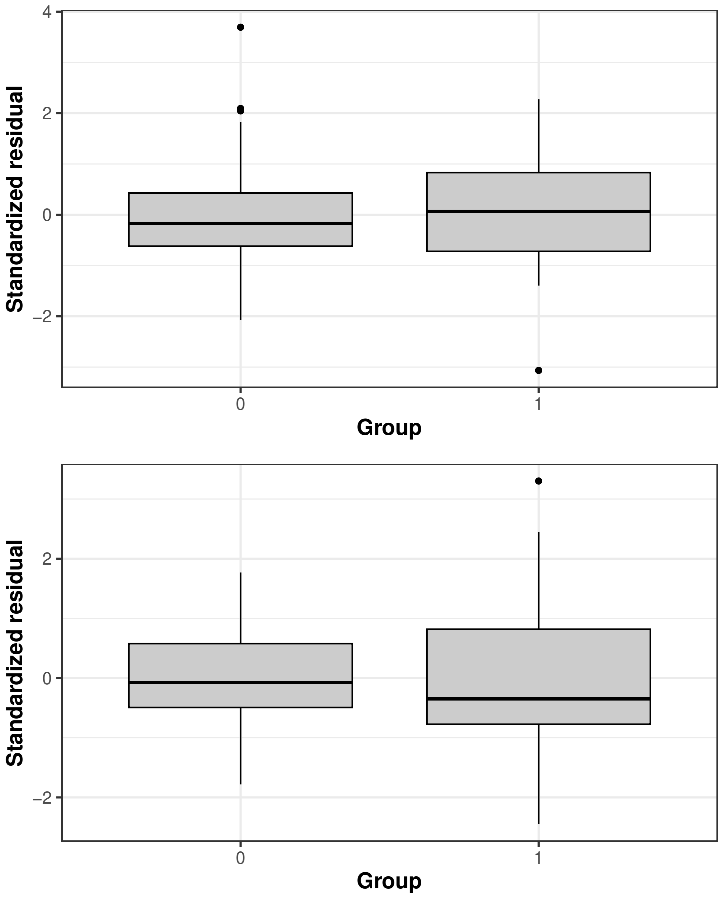


Figure 6.6.: Boxplots of standardized residuals by a factor covariate. Top panel: all assumptions are satisfied. Bottom panel: unequal variances of error terms and an omitted effect of another covariate.

Scatterplots of square root of absolute standardized residuals against continuous covariates

A better way to check the assumption of constant variability is to plot square root of absolute standardized residuals against a covariate. Such a plot can be smoothed to facilitate the interpretation. Figure 6.5 shows this type of plot for the same situations as in Figure 6.4. In the top panel, where the assumptions are satisfied, we see some increase in a certain range of covariate values but there is no overall trend. In the bottom panel, where there is an actual increase in variability, we clearly see increasing trend in the absolute standardized residuals.

Boxplots of standardized residuals by a categorical covariate

For categorical (factor) covariates that classify the observations into disjoint subgroups, boxplots of standardized residuals can be used to assess model assumptions. If the assumptions are satisfied, the medians of residuals in each group (the bars inside the boxes) will be close to zero and the height of the boxes (interquartile range) will be similar in all the subgroups (Figure 6.6, top panel). In the bottom panel of the same figure, the heights of the boxes visibly differ because the residual variance in Group 1 was larger than that in Group 0 and the box for Group 1 is not centered around zero because an important covariate correlated with group membership was omitted from the model.

Boxplots of standardized residuals by a factorized continuous covariate

Continuous covariates can be factorized in order to assess model assumptions using boxplots of standardized residuals. In the top panel of Figure 6.7, all the model assumptions were satisfied. In the bottom panel, the heights of the boxes somewhat increase from the left to the right because the residual variance increased with the covariate X and the boxes are located around a quadratic curve because the model failed to include an important quadratic effect of this covariate.

Histograms of standardized residuals

Histograms are helpful to visualize the distribution of residuals and to detect the presence of observations with unusually large absolute residuals. Figure 6.8 shows results of such a visualization when the residuals are normally distributed (top panel), when the residuals have a relatively heavy tailed distribution (middle panel) and when the distribution of residuals is skewed to the right (bottom panel).

Quantile plots of standardized residuals

Quantile plots (Q-Q plots) are generally more helpful tools to assess normality than histograms. These plots contain ordered standardized residuals on the vertical axis and plot them against theoretical expected values of corresponding order statistics calculated under normality (on the horizontal axis). More precisely, the Q-Q plot is a scatterplot of pairs $(u^*_{(i)}, z_i)$, where $u^*_{(i)}$ is the i-th smallest standardized residual and

$$z_i = \Phi^{-1}\left(\tfrac{i}{n+1}\right)$$

approximates $\mathsf{E}\, Z_{(i)}$, the expectation of the i-th order statistic in a random sample from $\mathsf{N}(0,1)$ of the size n. If the distribution or error terms is normal, the points displayed on the Q-Q plot approximately follow a line (the top panel of Figure 6.9). If the distribution of errors has heavier tails than the normal distribution, the Q-Q plot displays an S-shaped curve as shown in the middle panel of Figure 6.9. When the errors come from a skewed distribution, the Q-Q plot shows a bow-shaped curve (bottom panel of Figure 6.9).

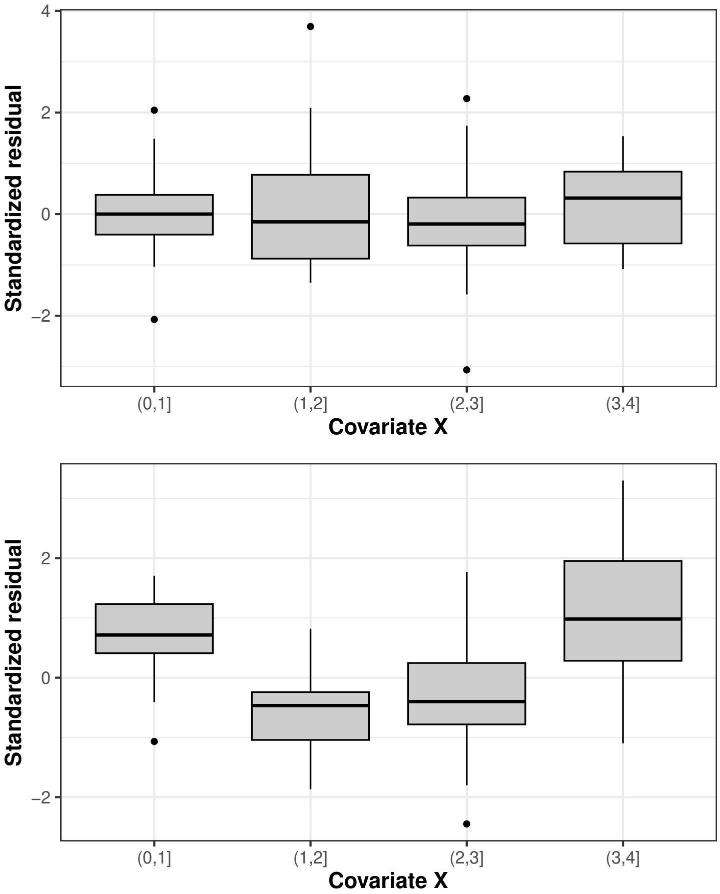


Figure 6.7.: Boxplots of standardized residuals by a factorized continuous covariate. Top panel: all assumptions are satisfied. Bottom panel: omitted quadratic effect and mildly increasing variance with X.

Assessing the correct functional form of a covariate

In linear regression, it is important to consider whether the effect of a particular continuous covariate can be modeled in a linear way or whether a more complex functional form (e.g., quadratic) is required. Scatterplots of standardized residuals against the values of that covariate can help to assess whether the covariate was included in an appropriate way (see Figure 6.3). A more convenient way to find the correct transformation of the covariate is to plot so called *partial residuals*.

Partial residuals are differences between the observed response Y_i and the fitted value $\widehat{Y}_i$ from which the estimated effect of the covariate was entirely removed. E.g., if a continuous covariate X_{i4} is included in the current model in the linear form, we obtain the partial residuals for that covariate by

$$Y_i - (\widehat{Y}_i - \widehat{\beta}_4 X_{i4}) = u_i + \widehat{\beta}_4 X_{i4}.$$

A smoothed scatterplot of partial residuals plotted against the values of the covariate of interest directly suggests an appropriate functional form for that covariate. If the linear form is sufficient the partial residuals seem to follow a line (see the top panel of Figure 6.10). In the bottom panel of the same figure, the true effect of the covariate is quadratic and partial residuals testify to that by following a parabolic function.

Testing model assumptions

A number of formal statistical tests have been developed to verify the validity of various model assumptions: equal variances, independence of error terms, normality of error terms. These tests always take as the null hypothesis the situation when the assumptions are fulfilled and reject the null hypothesis when the data contain sufficient evidence that a particular assumption is violated. The larger the data set is the more likely it is that the test will discover even minute and entirely unimportant violations. With small data sets, the test will often fail to find even very large and harmful violations. The impact of violated model assumptions on the results of the model depends on the extent of the violation – the test results depend on both the extent of violation and the sample size.

Therefore we recommend to assess model assumptions by various descriptive methods (residual plots, descriptive statistics) rather than by formal tests. Tests of normality in linear regression are particularly discouraged: with small sample sizes, normality is important for the validity of statistical inference, however the test of normality is not able to confirm that the data are truly normal. Even substantial departures from normality will go unnoticed in small data sets. On the other hand, with large sample sizes, the test will reject normality more frequently, however this assumption is entirely unimportant because asymptotic results provide reliable approximations even if the data are strongly non-normal. Formal tests of normality in linear regression are thus entirely worthless.

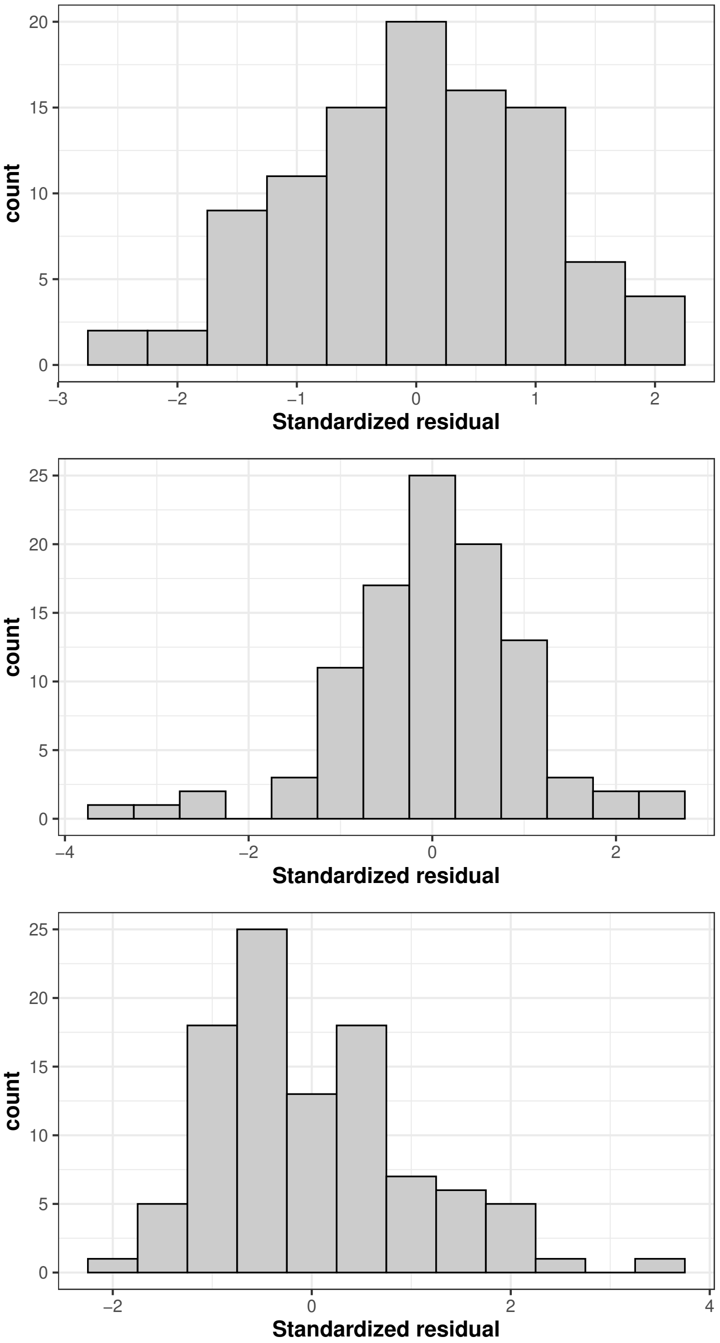


Figure 6.8.: Histograms of standardized residuals. Top panel: normal distribution of errors. Middle panel: heavy-tailed distribution of errors (t_4). Bottom panel: right-skewed distribution of errors (negative Gumbel).

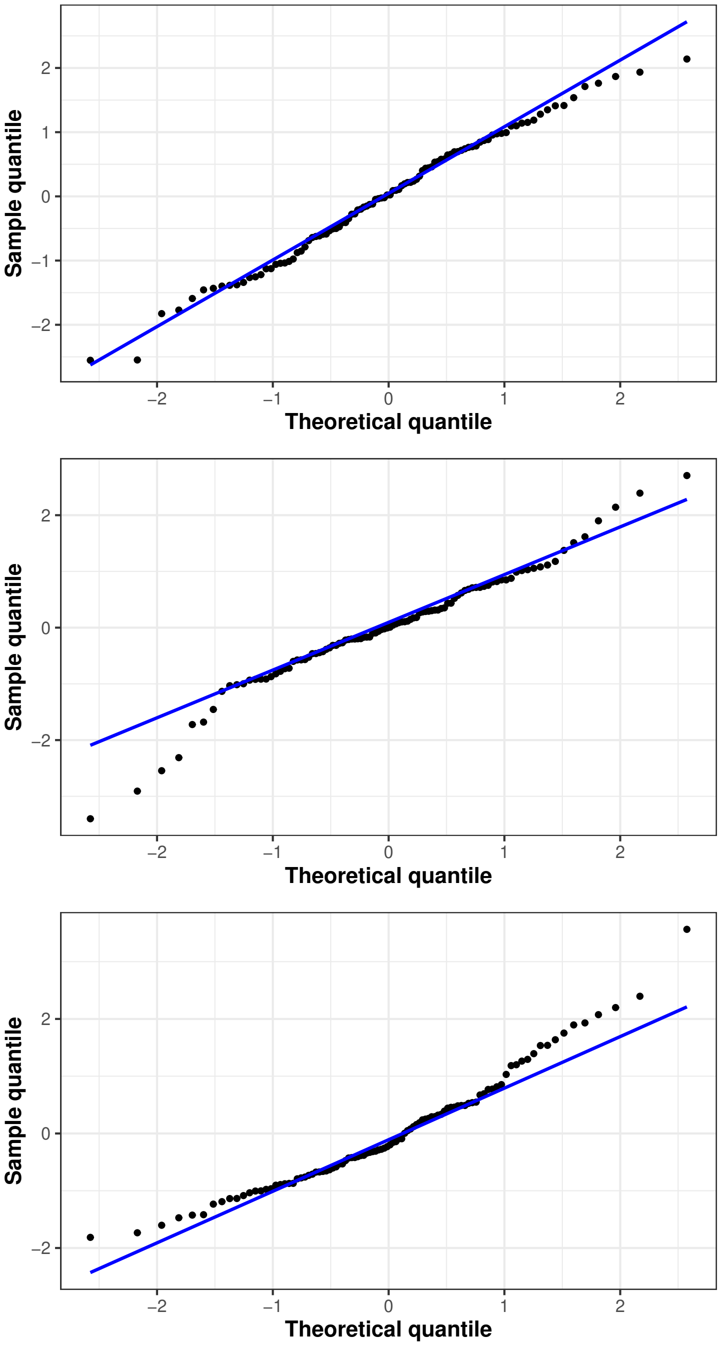


Figure 6.9.: Q-Q plots of standardized residuals. Top panel: normal distribution of errors. Middle panel: heavy-tailed distribution of errors (t_4). Bottom panel: right-skewed distribution of errors (negative Gumbel).

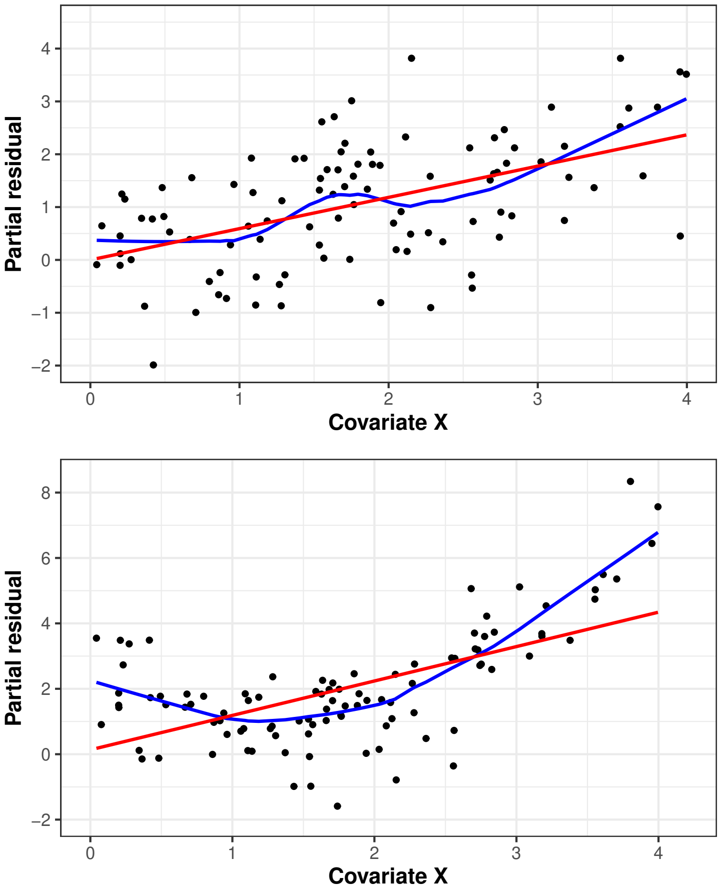


Figure 6.10.: Scatterplots of partial residuals against a covariate. Top panel: all assumptions are satisfied. Bottom panel: omitted quadratic effect. Smoothed by lowess smoother with window over 1/2 of the data range (blue).

7. Transformations of the Response

7.1. Introduction

Responses that do not follow the normal distribution typically violate the assumption of equal variances as well. For example, consider a response Y that can be close to zero but cannot be negative. If we use a linear model without intercept to express the dependence of $\mathsf{E}\,Y$ on a covariate $X > 0$

$$Y = \beta X + \varepsilon$$

it is clear that for very small X the variability in ε cannot be large because Y is positive and ε can only take small negative values. On the other hand, when X is large, the variability in ε can be much larger. The variance of ε thus tends to increase with increasing X. This is what frequently happens with this type of data.

The fact that the response is not normal is not of much concern if the sample size is large enough. However, non-constant variance is always of concern, no matter how large the data set is.

In these situations, a transformation of the response may help to satisfy the model assumptions. Suppose there exists a strictly monotone smooth function g such that

$$g(Y_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $\mathsf{E}\,\varepsilon_i = 0$ and $\mathsf{var}\,\varepsilon_i = \sigma_e^2$. Thus, we impose the linear model on $g(Y_i)$ rather than on Y_i itself.

We have $\mathsf{E}\,g(Y_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$. The parameters of this model express the effect of the covariate vector on the expectation of $g(Y_i)$. In general, the model does not allow us to estimate the effects of the covariates on $\mathsf{E}\,Y_i$. The implied model for Y_i is actually $Y_i = g^{-1}(\boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i)$ and hence $\mathsf{E}\,Y_i = \mathsf{E}\,g^{-1}(\boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i)$. We only know that $\mathsf{E}\,\varepsilon_i = 0$ and evaluation of $\mathsf{E}\,Y_i$ from this is not possible – unless g is a logarithm (see the next section).

A backward transformation of $\mathsf{E}\,g(Y_i)$ does not help. If g is convex and strictly increasing, we have by Jensen inequality $\mathsf{E}\,g(Y_i) > g(\mathsf{E}\,Y_i)$ and by application of the inverse on both sides, $\mathsf{E}\,Y_i < g^{-1}(\mathsf{E}\,g(Y_i))$.

If the goal of the transformed model is to make predictions about the response, the situation is less dire. Suppose that the density of ε_i in the transformed model is symmetric around 0. Then the mean and the median of $g(Y_i)$ are both equal to $\boldsymbol{X}_i^\top \boldsymbol{\beta}$. Unlike the mean, the median can be back-transformed and we get

$$\mathrm{med}(Y_i) = g^{-1}(\mathrm{med}(g(Y_i))) = g^{-1}(\boldsymbol{X}_i^\top \boldsymbol{\beta}). \tag{7.1}$$

Thus, if we manage to transform the response so that the errors have a symmetric distribution we can predict the median of Y_i by $g^{-1}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$ for any strictly monotone transformation g.

7.2. Interpretation of Linear Model With Log-transformed Response

Let $Y_i > 0$ almost surely and consider the model

$$\log Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

where $\mathsf{E}\,\varepsilon_i = 0$ and $\mathsf{var}\,\varepsilon_i = \sigma_e^2$. Thus, the linear regression model holds after the logarithmic transformation of the response. Then

$$Y_i = \mathrm{e}^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} \mathrm{e}^{\varepsilon_i} = \mathrm{e}^{\beta_1} \cdot \mathrm{e}^{\beta_2 X_{i2}} \cdots \mathrm{e}^{\beta_p X_{ip}} \cdot \mathrm{e}^{\varepsilon_i}.$$

Assume that $\mathrm{e}^{\varepsilon_i}$ has finite variance. We obtained a model where covariate effects multiply each other and the error also acts multiplicatively. The first two moments of the untransformed response can be calculated as follows:

$$\mathsf{E}\,Y_i = \mathrm{e}^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}\,\mathsf{E}\,\mathrm{e}^{\varepsilon_i},$$
$$\mathsf{var}\,Y_i = \left(\mathrm{e}^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}\right)^2 \mathsf{var}\,\mathrm{e}^{\varepsilon_i}.$$

If ε_i are independent and identically distributed, then $\mathsf{var}\,\mathrm{e}^{\varepsilon_i}$ is a constant and the variance of the response is an increasing function of $\boldsymbol{X}_i^\top \boldsymbol{\beta}$.

Denote $\gamma = \mathsf{E}\,\mathrm{e}^{\varepsilon_i}$. We know that $\mathsf{E}\,\varepsilon_i = 0$ and from Jensen's inequality (e^x is convex), $\gamma \equiv \mathsf{E}\,\mathrm{e}^{\varepsilon_i} > \mathrm{e}^{\mathsf{E}\,\varepsilon_i} = 1$.

Let us determine the interpretation of $\boldsymbol{\beta}$ in the untransformed model. We have

$$\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] = \mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{\beta}} \gamma \quad \text{and}$$
$$\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x} + \boldsymbol{e}_j\right] = \mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{\beta} + \beta_j} \gamma \quad j = 2, \ldots, p.$$

Taking the ratio of the two equations, the unknown γ cancels and we get

$$\frac{\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x} + \boldsymbol{e}_j\right]}{\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right]} = \frac{\mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{\beta} + \beta_j} \gamma}{\mathrm{e}^{\boldsymbol{x}^\top \boldsymbol{\beta}} \gamma} = \mathrm{e}^{\beta_j}.$$

In the linear model with a log-transformed response, e^{β_j} expresses the proportional increase in $\mathsf{E}\,Y$ after increasing the j-th covariate by one unit while keeping all other covariates fixed.

The log-transformed linear model estimates the multiplicative effects of the covariates on the expectation of the response and these are expressed by the exponentiated regression coefficients.

What about the intercept term? Set $\boldsymbol{X} = (1,0,\dots,0)^\mathsf{T}$. Then $\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{e}_1\right] = \gamma \mathrm{e}^{\beta_1} = \mathrm{e}^{\beta_1 + \log\gamma}$. But γ is not specified, we only know that $\gamma > 1$. Thus, the intercept cannot tell anything about the expectation of the response when all covariates are zero unless we assume a specific distribution of the response. For example, when $\varepsilon \sim \mathsf{N}(0,\sigma_e^2)$ (that is, Y is log-normal) then $\gamma = \mathsf{E}\,\mathrm{e}^{\varepsilon} = \mathrm{e}^{\sigma_e^2/2}$ and $\mathsf{E}\left[Y \,\middle|\, \boldsymbol{X} = \boldsymbol{e}_1\right] = \mathrm{e}^{\beta_1 + \sigma_e^2/2}$.

The logarithmic transformation of the response is a valuable and frequently employed tool when the responses are strictly positive, attain values that are close to zero and have long right tails with increasing variability as the mean increases. The logarithmic transformation of such a response improves equality of variances and brings the distribution of the transformed response closer to normality.

Note.

(i) When the response attains non-negative values with $\mathsf{P}[Y = 0] > 0$, the log transform can be used after shifting the responses by a constant $c > 0$. Thus, the transformed response is $\log(Y + c)$. Question for thought: What is the interpretation of $\boldsymbol{\beta}$ after such transformation?

(ii) If the response is log-transformed, the covariates usually require a transformation as well.

7.3. The Box-Cox Transformation

Let $Y_i > 0$ almost surely. Consider the following family of transformations of the response depending on a parameter $\lambda \in \mathbb{R}$:

$$
\begin{aligned}
g_\lambda(y) &= \frac{y^\lambda - 1}{\lambda} && \text{for } y \neq 0,\\
g_\lambda(y) &= \log y && \text{for } y = 0.
\end{aligned}
$$

The transformations are defined in this way so that $\lim_{\lambda\to 0} g_\lambda(y) = g_0(y)$, that is, they are continuous in λ. Otherwise, subtracting a constant and dividing the response by a constant is irrelevant. Among the transformations included in the Box-Cox family are all power function transformations such as Y^p, $\sqrt[p]{Y}$, $\frac{1}{Y^p}$, $\frac{1}{\sqrt[p]{Y}}$, $p = 2,3,\dots$.

Assuming that there exists $\lambda \in \mathbb{R}$ such that $g_\lambda(Y)$ is normally distributed, such λ can be estimated via maximum likelihood methods based on the normal density. This is done numerically, the estimated λ does not have a closed-form expression.

However, there need not exist any λ that makes the response normal after the Box-Cox transformation and, even if it existed, the model does not have interpretable parameters.

However, this approach can be valuable when the goal of the regression analysis is to make predictions, we are happy with predicting the median instead of the expectation (see (7.1)) and the Box-Cox transformation is successful in making the responses normal.

When $g_\lambda(Y)$ is normal, we can use the methods of Section 5.3 to make predictions of individual future responses. By (5.1) we calculate a prediction interval (C_L, C_U) satisfying

$$\mathrm{P}[C_L < g_\lambda(Y) < C_U] = 1 - \alpha$$

and because $g_\lambda(Y)$ is a monotone transformation of Y, we get

$$\mathrm{P}\left[g_\lambda^{-1}(C_L) < Y < g_\lambda^{-1}(C_U)\right] = 1 - \alpha$$

(this is for an increasing $g_\lambda(Y)$, otherwise we switch the bounds).

Calculating a prediction interval for $\mathsf{E}\,Y$ of a future observation via Box-Cox transformation is not possible, however.

8. Parametrization of a Single Covariate

8.1. Parametrization of Categorical Covariates (Factors)

Categorical covariates (factors, qualitative variables)* are variables that indicate group membership. Their values, typically integers, do not have any numerical meaning.

Let the studied population be divided into m disjoint groups. Consider a covariate $Z_i \in \{1,\dots,m\}$ where $Z_i = j$ means that the i-th observation belongs to group j. The numerical coding of the groups is frequently arbitrary, in fact any permutation of integers $\{1,\dots,m\}$ could be used to distinguish the groups from each other. Therefore the linear model $\mathsf{E}\left[Y_i \,\middle|\, Z_i\right] = \beta_1 + \beta_2 Z_i$ does not make sense for a categorical covariate. Instead, we need to allow each group to have its own expectation which is not related to the expectations of other groups. So, let $\mu_j = \mathsf{E}\left[Y_i \,\middle|\, Z_i = j\right]$ denote the expectation of the response in the j-th group.

This situation has been already discussed in the context of one-way ANOVA models, see Section 4.1.7. There, the model was specified as

$$\mathsf{E}\left[Y_i \,\middle|\, Z_i\right] = \beta_1 \mathbb{1}(Z_i = 1) + \dots + \beta_m \mathbb{1}(Z_i = m) \tag{8.1}$$

The regression matrix $\mathbb{X}$ includes the indicators $\mathbb{1}(Z_i = j)$ in the j-th column. The columns of the regression matrix sum to one, so $\mathbf{1}_n \in \mathscr{M}(\mathbb{X})$ and the model includes an intercept. The parameters correspond to the group means: $\beta_j = \mu_j$ for $j = 1,\dots,m$. This is called the *cell-means parametrization* of a factor covariate. This parametrization is suitable for calculation the LSE of group expectations, fitted values and sums of squares. However, we cannot use it for parametrization of multiple factors included in the same model because the columns sum to one for each of the factors and the model does not have full rank.

We need a model that includes the intercept explicitly ($X_{i1} = 1$ for all i) and models the effects of the factor by $m-1$ additional parameters. We start from

$$\mathsf{E}\left[Y_i \,\middle|\, Z_i\right] = \beta_1 + \gamma_1 \mathbb{1}(Z_i = 1) + \dots + \gamma_m \mathbb{1}(Z_i = m).$$

Here, we only recoded the parameters of the previous model as $\gamma_1,\dots,\gamma_m$ and added an intercept denoted by β_1. Of course, this model does not have full rank. To deal with this, we introduce a single linear constraint on the parameters $\gamma_1,\dots,\gamma_m$. We can choose different constraints and each leads to a different parametrization of the model and different interpretation of the regression coefficients. In the rest of this section, we give examples of the most common parametrizations of factor variables.

* Česky *faktor, kategoriální veličina*

All the parametrizations generate models that are equivalent to the cell-means model (8.1) in the sense of Section 2.5; all have the same fitted values, residuals, residual sums of squares and regression sums of squares. They differ in the meaning and interpretation of regression parameters and their LSE's.

8.1.1. Reference group parametrization

Take the restriction $\gamma_1 = 0$. This leads to the model

$$\begin{aligned} \mathsf{E}\left[Y_i \,\middle|\, Z_i\right] &= \beta_1 + \gamma_2 \mathbb{1}(Z_2 = 1) + \cdots + \gamma_m \mathbb{1}(Z_i = m) \\ &= \beta_1 + \beta_2 \mathbb{1}(Z_2 = 1) + \cdots + \beta_m \mathbb{1}(Z_i = m) = \boldsymbol{X}_i^\top \boldsymbol{\beta}. \end{aligned}$$

The following table displays the group expectations and the rows of the regression matrix used for each group.

Group	$\mathsf{E}\, Y_i$	$\boldsymbol{X}_i^\top$				
1	$\mu_1 = \beta_1$	1	0	0	$\cdots$	0
2	$\mu_2 = \beta_1 + \beta_2$	1	1	0	$\cdots$	0
3	$\mu_3 = \beta_1 + \beta_3$	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\ddots$	
m	$\mu_m = \beta_1 + \beta_m$	1	0	0	$\cdots$	1

The second up to the last column of the matrix shown on the right of the above table defines an $m \times (m-1)$ matrix of so called *dummy variables* used to implement this parametrization in a regression matrix (the first column represents the intercept).

Let us determine the interpretation of the components of $\boldsymbol{\beta}$ using the above table. We have

$$\begin{aligned} \beta_1 &= \mu_1 = \mathsf{E}\left[Y_i \,\middle|\, Z_i = 1\right] \\ \beta_2 &= \mu_2 - \mu_1 = \mathsf{E}\left[Y_i \,\middle|\, Z_i = 2\right] - \mathsf{E}\left[Y_i \,\middle|\, Z_i = 1\right] \\ &\vdots \\ \beta_m &= \mu_m - \mu_1 = \mathsf{E}\left[Y_i \,\middle|\, Z_i = m\right] - \mathsf{E}\left[Y_i \,\middle|\, Z_i = 1\right] \end{aligned}$$

Thus, the first group serves as a reference group and the expectations in the other groups are all compared to the first group.

Note.

(i) This parametrization is used by R as the default coding of factors (there, it is called "treatment contrasts").

(ii) Any group can be selected as the reference group. It is undesirable, however, to choose a very small group as the reference group.

8.1.2. Zero-sum parametrization

This parametrization arises by the restriction $\sum_{j=1}^m \gamma_j = 0$.

With this restriction, we can express γ_m as $\gamma_m = -\gamma_1 - \cdots - \gamma_{m-1}$ and plug it into the model.

$$\begin{aligned}\mathsf{E}\big[Y_i \big| Z_i\big] &= \beta_1 + \gamma_1 \mathbb{1}(Z_i = 1) + \cdots + \gamma_{m-1}\mathbb{1}(Z_i = m-1) - (\gamma_1 + \cdots + \gamma_{m-1})\mathbb{1}(Z_i = m)\\ &= \beta_1 + \underbrace{\gamma_1}_{\beta_2} \underbrace{[\mathbb{1}(Z_i = 1) - \mathbb{1}(Z_i = m)]}_{X_{i2}} + \cdots + \underbrace{\gamma_{m-1}}_{\beta_m} \underbrace{[\mathbb{1}(Z_i = m-1) - \mathbb{1}(Z_i = m)]}_{X_{im}} = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}.\end{aligned}$$

Now we can determine the group expectations and the rows of the regression matrix used for each group and display them in a table.

Group	$\mathsf{E}\,Y_i$	$\boldsymbol{X}_i^\mathsf{T}$				
1	$\mu_1 = \beta_1 + \beta_2$	1	1	0	$\cdots$	0
2	$\mu_2 = \beta_1 + \beta_3$	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\ddots$	
$m-1$	$\mu_{m-1} = \beta_1 + \beta_m$	1	0	0	$\cdots$	1
m	$\mu_m = \beta_1 - \sum_{j=2}^m \beta_j$	1	-1	-1	$\cdots$	-1

Again, the second up to the last column of the matrix shown on the right of the above table defines an $m \times (m-1)$ matrix of so called *dummy variables* used to implement the zero-mean parametrization in a regression matrix (the first column represents the intercept).

From the above table, we have $\sum_{j=1}^m \mu_j = m\beta_1$, hence $\beta_1 = \frac{1}{m}\sum_{j=1}^m \mu_j \equiv \overline{\mu}$. Thus, the intercept corresponds to the arithmetic average of the expectations of the m groups. The interpretation of the other components of $\boldsymbol{\beta}$ is as follows.

$$\begin{aligned}\beta_2 &= \gamma_1 = \mu_1 - \beta_1 = \mu_1 - \overline{\mu}\\ \beta_3 &= \gamma_2 = \mu_2 - \beta_1 = \mu_2 - \overline{\mu}\\ &\vdots\\ \beta_m &= \gamma_{m-1} = \mu_{m-1} - \beta_1 = \mu_{m-1} - \overline{\mu}\end{aligned}$$

Finally, we can calculate

$$\gamma_m = -\sum_{j=1}^{m-1}\gamma_j = (m-1)\overline{\mu} - \sum_{j=1}^{m-1}\mu_j = \sum_{j=1}^{m}\mu_j - \overline{\mu} - \sum_{j=1}^{m-1}\mu_j = \mu_m - \overline{\mu}.$$

Thus, the parameters γ_j compare the expectation in the j-th group to the arithmetic average of the expectations of all groups.

Note. Zero-sum parametrization is specified in R by assigning this type of contrasts (dummy variables) to the factor `z` through the function `C(z,contr.sum)` and using this in the regression model instead of the original factor `z`.

8.1.3. Weighted zero-sum parametrization

This parametrization arises by the restriction $\sum_{j=1}^{m} n_j \gamma_j = 0$ where $n_j = \sum_{i=1}^{n} \mathbb{1}(Z_i = j)$ is the size of the j-th group. This restriction leads to $\beta_1 = \frac{1}{n} \sum_{j=1}^{m} n_j \mu_j$, which can be estimated by the overall mean of all observations and is itself an estimate of the unconditional mean of the response. The other parameters have the interpretation $\gamma_j = \beta_{j+1} = \mu_j - \beta_1$ and represent the differences between the conditional mean of the response in the j-th group and the unconditional mean of the response.

This parametrization is not recommended despite its seemingly attractive interpretation. The reason is that the meaning of its parameters depends on the data (through n_j) and thus it changes when the data set is modified. This is undesirable. We strongly prefer parametrizations that have the same interpretation in any data set.

8.1.4. Helmert parametrization

This is another parametrization offered by R. It leads to parameters that can be interpreted as follows: $\beta_1 = \mu_1$, $\beta_2 = \mu_2 - \mu_1$, $\beta_3 = \mu_3 - \frac{\mu_1+\mu_2}{2}$, $\beta_4 = \mu_4 - \frac{\mu_1+\mu_2+\mu_3}{3}$ etc. The parameter β_j for $j > 1$ expresses the difference between the conditional mean in the j-th group and the arithmetic average of the conditional means in the preceding groups.

This parametrization is rarely useful, it is suitable only for certain specific industrial applications.

8.1.5. Orthogonal polynomials for ordered factors

We say that $Z \in \{1, \dots, m\}$ is an *ordered factor* (or ordered categorical variable) if it indicates membership in one of m groups and the groups are in some sense ordered, that is, group 1 precedes group 2, which precedes group 3, etc.

Examples of ordered factors are school grades (1 = excellent, 2 = very good, ..., 5 = very poor), opinion scales (1 = strongly agree, 2 = somewhat agree, ..., 5 = strongly disagree), or grouped continuous variables such as age groups. The values of ordered factors have some numerical interpretation and it is desirable to take this into account in the analysis. In particular, it makes sense to consider a polynomial model expressing the mean in the j-th group as follows:

$$\mathsf{E}\left[Y_i \middle| Z_i\right] = \beta_1 + \gamma_1 Z_i + \gamma_2 Z_i^2 + \dots + \gamma_{m-1} Z_i^{m-1}.$$

This model is equivalent to the cell-means model (8.1) because any constellation of values of the m means can be exactly fitted by a polynomial of degree $m-1$. However, the interpretation of the regression coefficients is difficult and their least squares estimators are strongly correlated because certain columns of the resulting regression matrix are correlated.

Instead, we consider an equivalent polynomial model

$$\mathsf{E}\big[Y_i \,\big|Z_i\big] = \beta_1 + \gamma_1 g_1(Z_i) + \gamma_2 g_2(Z_i) + \cdots + \gamma_{m-1} g_{m-1}(Z_i),$$

where $g_j(Z_i)$ are mutually orthogonal polynomials in Z_i of degree j, $j = 1, \dots, m-1$. The form of the regression matrix can be described by the following table.

Group	$\boldsymbol{X}_i^\mathsf{T}$				
1	1	$g_1(1)$	$g_2(1)$	$\cdots$	$g_{m-1}(1)$
2	1	$g_1(2)$	$g_2(2)$	$\cdots$	$g_{m-1}(2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$m-1$	1	$g_1(m-1)$	$g_2(m-1)$	$\cdots$	$g_{m-1}(m-1)$
m	1	$g_1(m)$	$g_2(m)$	$\cdots$	$g_{m-1}(m)$
		$\underbrace{\hphantom{g_1(m-1)}}_{\boldsymbol{g}_1}$	$\underbrace{\hphantom{g_2(m-1)}}_{\boldsymbol{g}_2}$		$\underbrace{\hphantom{g_{m-1}(m-1)}}_{\boldsymbol{g}_{m-1}}$

The vectors $\boldsymbol{g}_j$ are polynomials of degree j evaluated at the group codes $1, 2, \dots, m$ satisfying $\mathbf{1}^\mathsf{T}\boldsymbol{g}_j = 0$ for $j \in \{1, \dots, m-1\}$ and $\boldsymbol{g}_j^\mathsf{T}\boldsymbol{g}_k = 0$ for $j \neq k \in \{1, \dots, m-1\}$. Such polynomials are not uniquely determined. For the most frequent situations when $m = 3$ or $m = 4$, we can use

$$m = 3: \quad (\boldsymbol{g}_1, \boldsymbol{g}_2) = \begin{pmatrix} -1 & 1 \\ 0 & -2 \\ 1 & 1 \end{pmatrix}, \qquad m = 4: \quad (\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3) = \begin{pmatrix} -3 & 1 & -1 \\ -1 & -1 & 3 \\ 1 & -1 & -3 \\ 3 & 1 & 1 \end{pmatrix}.$$

The functional dependence of $\mathsf{E}\big[Y_i \,\big|Z_i\big]$ on Z_i can be determined by testing the significance of the individual parameters $\gamma_1, \dots, \gamma_{m-1}$. If there is interest in the linear part, the coefficient γ_1 describes the slope of the linear relationship and a test of linearity can be performed by testing $H_0: \gamma_2 = \cdots = \gamma_{m-1} = 0$. The LSE of γ_1 has exactly the same value and standard error as if the model only contained the linear term without any higher order terms (see Chapter **??** for a justification of this claim).

Note. This is the default parametrization in R when a factor is declared as ordered by calling the function `factor(z,ordered=TRUE)` or `ordered(z)`. For any factor, this parametrization can be also requested by calling the function `C(z,contr.poly)` and using this in the regression model instead of the original factor `z`. In R, it is called "polynomial contrasts".

Note. If the ordered factor arose by grouping a numerical covariate into intervals (for example age groups defined as 0–10 years, 11–20 years, etc.), it is better to code the groups by midpoints of those intervals rather than by integers $1, 2, \dots, m$, especially if the intervals do not have the same width. Such covariates can be also treated by the methods described in the next section, though.

8.2. Parametrization of Quantitative (Numerical) Covariates

Quantitative (numerical) covariates are measurements of some numerical quantity that often has some associated physical units (age, height, price, temperature, percentage, number of children). Thus, their values have numerical interpretation. Quantity covariates may be realizations of a continuous random variable with many distinct values and infrequent ties due to rounding, may attain a limited number of pre-specified values (temperature in industrial experiments) or may be purely discrete. A discrete quantitative variable can be also treated as a factor or an ordered factor by the methods discussed in the previous section.

The simplest way how to include a quantitative covariate in the model is in a linear form $m(Z) \equiv \mathsf{E}\left[Y \,\middle|\, Z\right] = \beta_1 + \beta_2 Z$. (In this section, we will use the notation $m(Z)$ to express the functional form of the dependence of $\mathsf{E}\, Y$ on Z.) However, the linear form of $m(Z)$ may not be appropriate.

To come up with more complex functional forms, we can transform Z by a pre-selected set of basis functions $g_1, \dots, g_m$ and fit the model

$$m(Z) = \beta_1 + \gamma_1 g_1(Z_i) + \gamma_2 g_2(Z_i) + \dots + \gamma_m g_m(Z_i).$$

This form can capture any function that can be expressed as a linear combination of the selected basis functions.

How to choose the basis functions? We can learn from the results of descriptive analysis of residuals (see Chapter 6) or we can use submodel testing to compare more complex models with simpler models and remove the unnecessary terms. In the rest of this section, we will describe various possibilities for the choice of the basis functions.

8.2.1. Factorization of a numerical covariate

We can transform a numerical variable Z into a factor variable by specifying groups determined by ranges of values of Z. Take the points

$$t_0 = -\infty < t_1 < \dots < t_{m-1} < t_m = \infty,$$

define intervals $A_j = \langle t_{j-1}, t_j)$ for $j = 1, \dots, m$ and set

$$Z_i^* = j \Leftrightarrow Z_i \in A_j.$$

Then treat Z_i^* as a factor variable (or an ordered factor variable) with m levels using the methods of Section 8.1. This is equivalent to approximating $m(Z)$ by a piecewise constant function on the prespecified intervals. We can visualize the fitted piecewise constant function and investigate it to choose a suitable continuous functional form for further investigation or keep the piecewise constant approximation in the final model.

Note. We can test whether the true effect of the covariate is linear by fitting the model that includes the linear form as well as the factorized form

$$m(Z) = \beta_1 + \beta_2 Z + \gamma_2 \mathbb{1}(Z \in A_2) + \cdots + \gamma_m \mathbb{1}(Z \in A_m)$$

and testing the hypothesis $H_0 : \gamma_2 = \cdots = \gamma_m = 0$ by comparing this model with $m(Z) = \beta_1 + \beta_2 Z$.

Note. When choosing the cutoff points $t_1, \dots, t_{m-1}$, we should make sure that there is enough observations in each interval A_j so that we can estimate the expectation on that interval with sufficient precision.

Advantages: Simple flexible method, easy to implement, with straightforward interpretation of model parameters.

Disadvantages: The results depend on the choice of the cutoff points t_j, the intervals must be wide enough to include enough observations, the model may include too many parameters, factorization entails a partial loss of information contained in the original form of the covariate and is inefficient.

8.2.2. Polynomial basis

Choose g_j as a polynomials of degree j for $j = 1, \dots, m$. Then $m(Z)$ has the form of a polynomial of degree m. For example,

$$m(Z) = \beta_1 + \gamma_1 Z + \gamma_2 Z^2 + \cdots + \gamma_m Z_m.$$

Because this choice of basis functions creates a regression matrix with highly correlated columns (see the discussion in Section 8.1.5) one could create the basis from orthogonal polynomials instead, e.g. using Legendre polynomials after transforming all the values of the covariate into the closed interval $\langle -1, 1 \rangle$.

Note. In R, an orthogonal polynomial basis of degree m can be calculated by calling the function `poly(z,degree=m)`, which can be used directly on the right-hand side of the model formula.

According to the Weierstrass Theorem, any continuous function defined on a finite closed interval can be approximated by a polynomial with an arbitrary precision. In practice, it is not really so. Fitting a polynomial of a very high degree requires a lot of parameters and their estimates are imprecise. The resulting polynomial tends to be wiggly and may fit the data poorly at the edges as well as over flat areas.* A visualization of the fitted polynomial may give a misleading impression about the true shape of the function $m(Z)$.

Advantages: Simple flexible method, easy to implement, theoretically may provide a good approximation to $m(Z)$.

* This is called *Runge's phenomenon*

Disadvantages: Difficult interpretation of model parameters, poor fit at the edges of the data and over flat areas.

8.2.3. Piecewise polynomial bases, B-splines

Splines represent another convenient and flexible method for modeling complex relationships between a continuous covariate and the mean of the response.

Definition 8.1.

- A spline of degree k is a piecewise polynomial function of degree k with $k-1$ continuous derivatives everywhere.
- A spline basis is a set of functions $B_{j,k}(x)$ such that any spline of degree k can be expressed as a linear combination of $B_{j,k}(x)$. ∇

The most convenient spline basis is so called B-spline basis that satisfies the conditions

$$B_{j,k}(x) \geq 0 \ \forall x \quad \text{and} \quad \sum_j B_{j,k}(x) = 1 \ \forall x.$$

Consider a continuous variable Z with observed values $Z_1, \dots, Z_n$. Choose points $t_1 < t_2 < \dots < t_m < t_{m+1}$ such that $t_1 < Z_i < t_{m+1}$ $\forall i$. These points are called *knots*[*]. The knots divide the range of data values $Z_1, \dots, Z_n$ into m disjoint intervals $I_j \equiv \langle t_j, t_{j+1})$. They should be chosen so that each of these intervals contains a sufficient number of observations. The knots t_1 and t_{m+1} are called *boundary knots*, the other $m-1$ knots are called *inner knots*.

Given the set of knots t_j and intervals I_j, define the B-spline basis functions of degree $k = 0$ as

$$B_{j,0}(x) = \mathbb{1}(x \in I_j) = \begin{cases} 1 & x \in \langle t_j, t_{j+1}) = I_j, \\ 0 & \text{otherwise}, \end{cases}$$

for $j = 1, \dots, m$. Linear combinations of these zero-degree basis functions are discontinuous piecewise constant functions. The basis functions are non-negative and sum to one, as required. They can be included in the linear model in the form

$$m(Z) = \gamma_1 B_{1,0}(Z) + \gamma_2 B_{2,0}(Z) + \dots + \gamma_m B_{m,0}(Z)$$

or, preferably, with an explicit intercept term after dropping the first basis function

$$m(Z) = \beta_1 + \gamma_2 B_{2,0}(Z) + \dots + \gamma_m B_{m,0}(Z)$$

These basis function create exactly the same approximation by a piecewise constant function as the factorization approach discussed in Section 8.2.1.

[*] Česky *uzly*

$$\begin{tikzcd}
 & 0 \arrow[dr] & \\
0 \arrow[dr] & & B_{1,2} \\
 & B_{1,1} \arrow[ur] \arrow[dr] & \\
B_{1,0} \arrow[ur] \arrow[dr] & & B_{2,2} \\
 & B_{2,1} \arrow[ur] \arrow[dr] & \\
B_{2,0} \arrow[ur] & & B_{3,2} \\
\vdots & & \vdots \\
B_{m-1,0} \arrow[dr] & & B_{m,2} \\
 & B_{m,1} \arrow[ur] \arrow[dr] & \\
B_{m,0} \arrow[ur] \arrow[dr] & & B_{m+1,2} \\
 & B_{m+1,1} \arrow[ur] \arrow[dr] & \\
0 \arrow[ur] & & B_{m+2,2} \\
 & 0 \arrow[ur] &
\end{tikzcd}$$

Figure 8.1.: Recurrent procedure for generation of B-spline bases.

The higher degree B-spline basis functions are constructed by a recurrent relationship called *the Cox–de Boor recursion formula*:

$$B_{j,k}(x) = w_{j-1,k}(x)B_{j-1,k-1}(x) + [1 - w_{j,k}(x)]B_{j,k-1}(x)$$

for $j = 1, \ldots, m+k$ and $k = 1, 2, \ldots$, where

$$w_{j,k}(x) = \begin{cases} \frac{x-t_j}{t_{j+k}-t_j} & \text{for } x \in \langle t_j, t_{j+k}), j \neq 1 \text{ and } j \neq m+k, \\ 0 & \text{for } x \notin \langle t_j, t_{j+k}) \text{ or } j = 1, \\ 1 & \text{for } j = m+k. \end{cases}$$

This creates higher degree basis functions by linear combinations of two neighboring lower-degree basis functions, with coefficients that linearly depend on the argument. A graphical illustration of the recurrent procedure is provided by Figure 8.1.

For degree zero, each basis function spans only one interval. For degree one, each basis functions spans two neighboring intervals (except at the boundary intervals). For degree k, each basis functions spans $k+1$ neighboring intervals (except for the first k and the last k basis functions). With m intervals, we have $m-1$ inner knots between them, $m+k$ basis functions and $m+k-1$ model parameters (the basis functions are linearly dependent, so the first of them is dropped).

Using the Cox-de Boor formula, the B-spline basis functions of degree 1 can be expressed as follows:

$$
\begin{aligned}
B_{1,1}(x) &= 0 + \left[1 - \frac{x - t_1}{t_2 - t_1}\right]\mathbb{1}(x \in \langle t_1, t_2)) = \frac{t_2 - x}{t_2 - t_1}\mathbb{1}(x \in \langle t_1, t_2)) \\
B_{2,1}(x) &= \frac{x - t_1}{t_2 - t_1}\mathbb{1}(x \in \langle t_1, t_2)) + \frac{t_3 - x}{t_3 - t_2}\mathbb{1}(x \in \langle t_2, t_3))
\end{aligned}
$$

et cetera. Figures 8.2, 8.3, and 8.4 illustrate the linear, quadratic and cubic B-spline bases.

When selecting the knot locations, it is important to make sure that there are enough observations in the interval between two neighboring knots. One could place the inner knots equidistantly, or place them at certain empirical quantiles of the observed values of the covariate, or make them denser at regions where one expects more abrupt changes of the fitted function and sparser elsewhere. Outer knots should be placed independently of the data (not at the minimum and maximum of the observed covariate values).

One can use a submodel F-test to verify the degree of the spline as well as the significance of one or more selected knots: lowering a degree of the spline or removing knots creates a submodel.

Advantages:

- very flexible tool, many functions can be well approximated by splines with suitably selected knots
- each basis spans at most $k+1$ intervals; individual observations have only local influence on the fitted curve, not a global influence over the whole range of the data
- splines are easy to use in `R` using the function `bs` in the library `spline`

Disadvantages:

- the estimated parameters are hard to interpret (except in case of linear splines)
- the fitted function must be plotted in order to understand the meaning of the results
- it may be difficult to calculate predictions correctly in an external data set (one must make sure that the spline bases are calculated in exactly the same way as in the original data)

Figures 8.5 and 8.6 show examples of nonlinear mean functions fitted by linear regression lines, high-order polynomials and cubic splines.

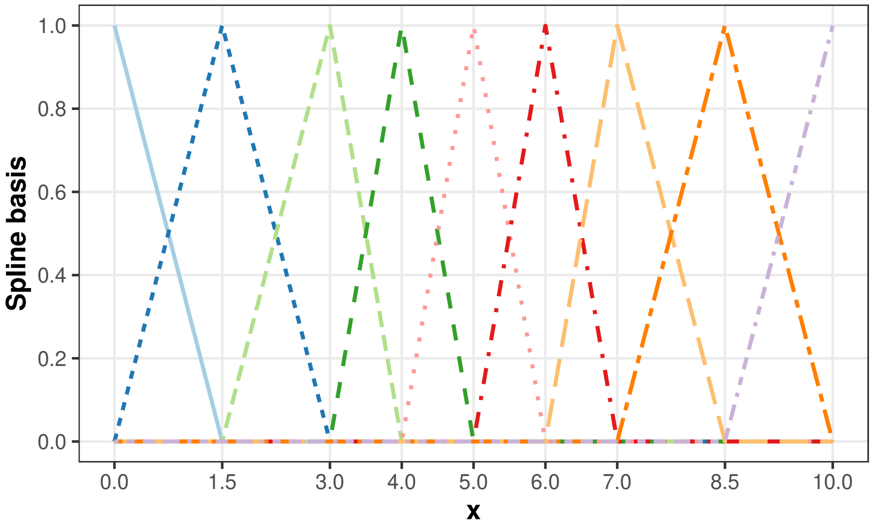


Figure 8.2.: B-spline bases of degree 1 over the interval $\langle 0, 10\rangle$.

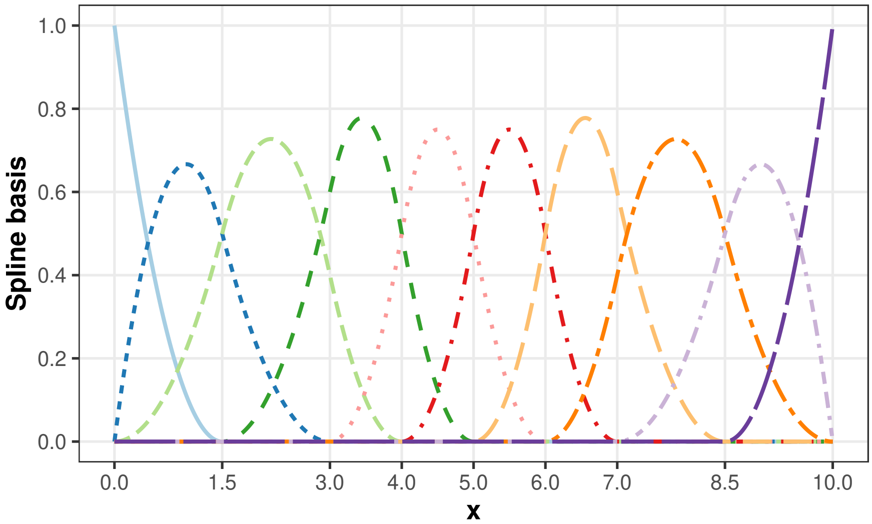


Figure 8.3.: B-spline bases of degree 2 over the interval $\langle 0, 10\rangle$.

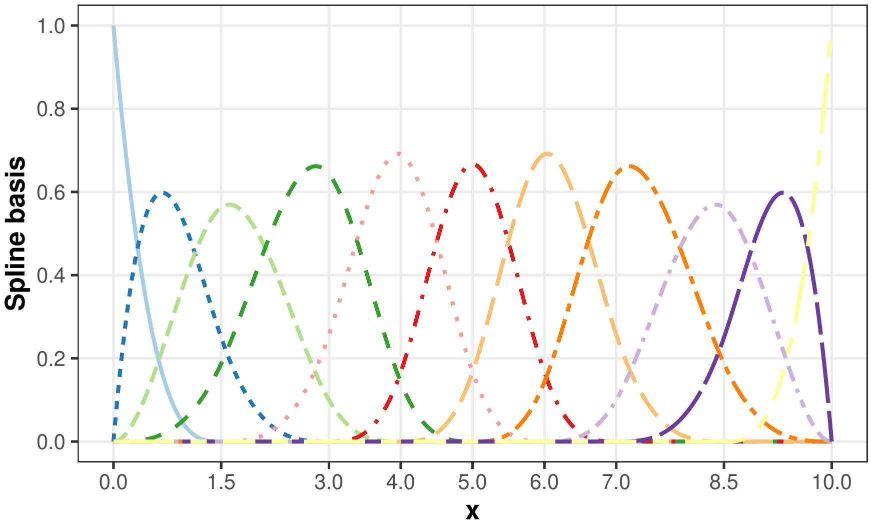


Figure 8.4.: B-spline bases of degree 3 over the interval $\langle 0, 10\rangle$.

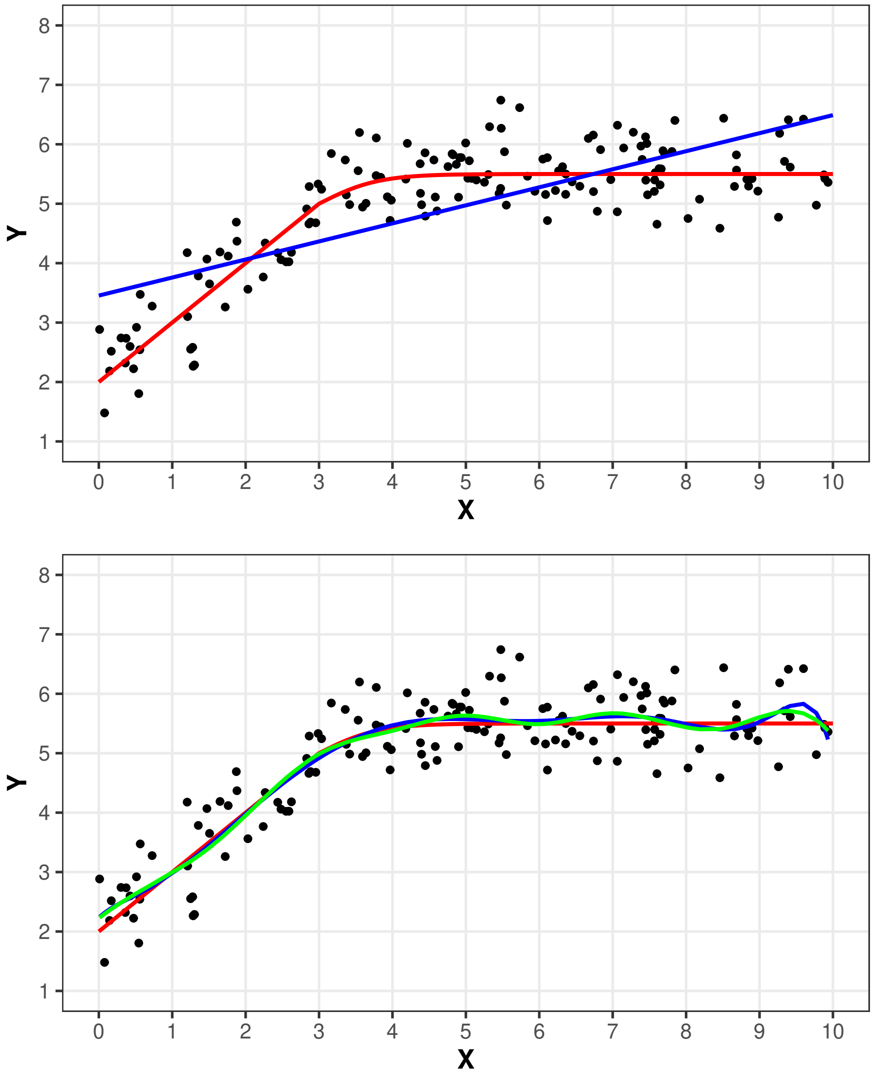


Figure 8.5.: Logistic pipe function (red curve = true mean) fitted by a regression line (top panel), by a 10-th degree polynomial (blue curve in bottom panel), and by a 3rd degree spline with 7 inner knots (green curve in bottom panel).

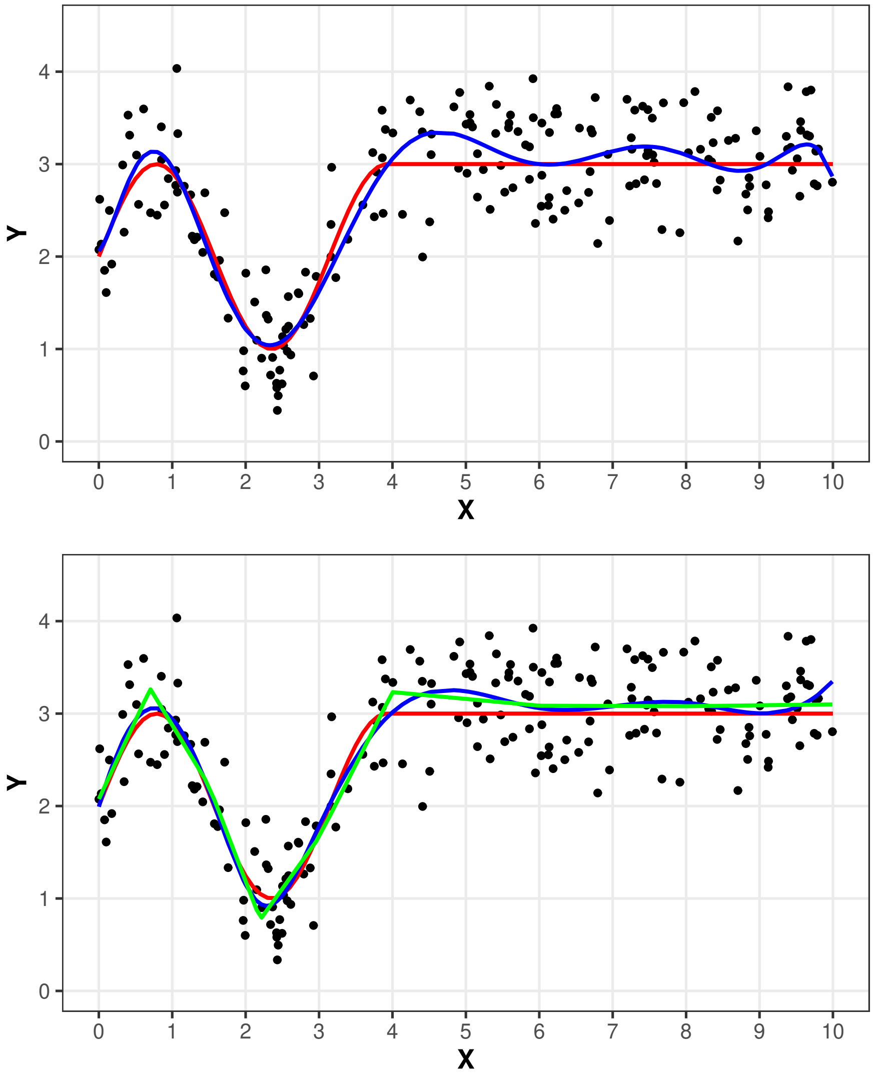


Figure 8.6.: Sine-constant function (red curve = true mean) fitted by 10th degree polynomial (blue curve in top panel) and by 1st degree (green curve in bottom panel) and 3rd degree (blue curve in bottom panel) spline with 7 inner knots.

8.2.4. Sine-cosine basis

Suppose $m(z) \equiv \mathsf{E}\left[Y \,\middle|\, Z = z\right]$ is a periodic function of a bounded variation in the argument z, with a period d. Such a function can be expressed using Fourier series in a sine-cosine form as follows

$$m(z) = a_0 + \sum_{j=1}^{M} a_j \cos\varphi_j \cos\left(\frac{2\pi j}{d} z\right) + \sum_{j=1}^{M} a_j \sin\varphi_j \sin\left(\frac{2\pi j}{d} z\right),$$

where M may be infinite, a_j represents the amplitude of the j-th term and φ_j represents the phase shift of the j-th term.

This expression can be approximates in a linear regression model by choosing the maximum period d and a finite number of terms M and transforming the covariate Z in the following way:

$$m(Z) = \mathsf{E}\left[Y \,\middle|\, Z\right] = \beta_1 + \sum_{j=1}^{M} \gamma_{1j} \cos\left(\frac{2\pi j}{d} Z\right) + \sum_{j=1}^{M} \gamma_{2j} \sin\left(\frac{2\pi j}{d} Z\right).$$

The parameters γ_{1j} and γ_{2j} are estimated as regression parameters using least squares.

9. Multiple Comparisons and Simultaneous Confidence Intervals

Recall Section 4.1.4, where we assumed normality and considered testing the hypothesis

$$H_0 : \mathbb{C}\boldsymbol{\beta} = \mathbf{0} \qquad \text{against} \qquad H_1 : \mathbb{C}\boldsymbol{\beta} \neq \mathbf{0},$$

where $\mathbb{C}_{q\times p}$ with $q \leq p$ and $r(\mathbb{C}) = q$ is a matrix of constants. This is tested by a submodel F-test. By Lemma 4.7, the same test can be applied asymptotically when the responses are not normal. The hypothesis H_0 is equivalent to the set of q partial hypotheses $H_{0j} : \boldsymbol{c}_j^\top \boldsymbol{\beta} = 0$ for $j = 1, \ldots, q$, where $\mathbf{c}_j$ is the j-th row of the matrix $\mathbb{C}$.

Now imagine the composite hypothesis H_0 is not true. This means that at least one of the partial hypotheses is violated. We want to test the partial hypotheses in order to find out which of them can be rejected. As we will see in the following section, this cannot be done by performing the q partial hypotheses with the usual level because the overall rate of type I error would be much larger than desired.

9.1. Bonferroni Procedure

Let us state the problem quite generally. Suppose Θ is the parameter space and consider subsets $\Theta_k \subset \Theta$ for $k = 1, \ldots, M$ such that $\bigcap_{k=1}^{M} \Theta_k \neq \emptyset$. Formulate hypotheses about the true parameter $\boldsymbol{\theta}$ falling into those subsets:

$$H_{0k} : \boldsymbol{\theta} \in \Theta_k, \quad k = 1, \ldots, M.$$

These hypotheses are to be tested with tests specified by test statistics T_k and critical regions C_k. Hence, the hypothesis H_{0k} is rejected if and only if $T_k \in C_k$. Suppose the critical regions are selected so that the level of each test is α^*, that is

$$\mathrm{P}\big(T_k \in C_k \,\big|\, H_{0k} \text{ holds}\big) = \alpha^*$$

(replace the equality by a limit if the test is asymptotic). In this way, we can control the probability of type I error for each test separately. But suppose that we need to control these errors *simultaneously* for all the tests. We need a procedure that assures

$$\mathrm{P}\big(\{T_1 \in C_1\} \cup \{T_2 \in C_2\} \cup \cdots \cup \{T_M \in C_M\} \,\big|\, H_{01}, \ldots, H_{0M} \text{ hold}\big) \leq \alpha$$

so that the probability of committing at least one type I error is bounded from above by α.

For any random events $A_1,\ldots,A_M$,

$$\mathrm{P}\left[A_1 \cup A_2 \cup \cdots \cup A_M\right] \le \sum_{k=1}^{M} \mathrm{P}\left[A_k\right]$$

and equality holds if and only if A_j and A_k are disjoint for any $j \neq k$. This is called *Bonferroni inequality* (Bonferroni 1936). Applying this, we get

$$\begin{aligned}\mathrm{P}\Big(\bigcup_{k=1}^{M}\{T_k \in C_k\}\,\Big|H_{01},\ldots,H_{0M} \text{ hold}\Big)\\ \le \sum_{k=1}^{M}\mathrm{P}\Big(T_k \in C_k\,\Big|H_{01},\ldots,H_{0M} \text{ hold}\Big) = \sum_{k=1}^{M}\mathrm{P}\Big(T_k \in C_k\,\Big|H_{0k} \text{ holds}\Big) = M\alpha^*.\end{aligned}$$

Carlo Emilio Bonferroni *(1892 – 1960) was an Italian mathematician who worked on probability theory. Bonferroni is best known for the Bonferroni inequalities (a generalization of the union bound), and for the Bonferroni correction in statistics, which he did not invent, but which is related to the Bonferroni inequalities.*
Source: https://en.wikipedia.org/wiki/Carlo_Emilio_Bonferroni

If we conduct M tests on the level α^*, the overall level (probability of committing at least one type I error) may be as large as $M\alpha^*$. *The more tests are performed the more likely it is that at least one valid hypothesis is rejected.* To protect against this, it is sufficient to take $\alpha^* = \alpha/M$ for the level of each individual test. The overall level of all tests is then guaranteed to be at most the desired α. This is called *the Bonferroni method*.

The same principle applies to confidence intervals. Suppose $U_1,\ldots,U_M$ are confidence intervals for parameters $\theta_1,\ldots,\theta_M$, each of them has coverage probability $1-\alpha^*$ (or converging to $1-\alpha^*$ if the interval is asymptotic). This is equivalent to $\mathrm{P}\left[U_k \not\ni \theta_k\right] = \alpha^*$.

We want to control the errors simultaneously requiring that the probability that at least one interval does not cover the true parameter is at most α, i.e.,

$$\mathrm{P}\left[\bigcup_{k=1}^{M}\{U_k \not\ni \theta_k\}\right] \le \alpha$$

and, equivalently,

$$\mathrm{P}\left[\bigcap_{k=1}^{M}\{U_k \ni \theta_k\}\right] \ge 1-\alpha.$$

By Bonferroni inequality, we can achieve this by setting $\alpha^* = \alpha/M$, that is, by setting the coverage probabilities of individual intervals to $1-\alpha/M$.

Bonferroni method in linear regression

Let us return to the problem of testing

$$H_0 : \mathbb{C}\boldsymbol{\beta} = \mathbf{0} \qquad \text{against} \qquad H_1 : \mathbb{C}\boldsymbol{\beta} \neq \mathbf{0},$$

and the partial hypotheses $H_{0j} : \mathbf{c}_j^{\mathsf{T}}\boldsymbol{\beta} = 0$ in linear regression.

By Lemma 4.2 and Lemma 4.7, H_0 is rejected when

$$\frac{1}{q\widehat{\sigma}_e^2}\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbb{C}^{\mathsf{T}}\big[\mathbb{C}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{C}^{\mathsf{T}}\big]^{-1}\mathbb{C}\widehat{\boldsymbol{\beta}} \geq F_{q,n-p}(1-\alpha).$$

This test has the level α (exact or asymptotic) but if it rejects we do not know which of the hypotheses H_{0j} should be rejected and which should not.

Instead, we can test each $H_{0j} : \mathbf{c}_j^{\mathsf{T}}\boldsymbol{\beta} = 0$ separately by a t-test at the level α/q (recall Lemma 4.1 and Lemma 4.6). We reject H_{0j} when

$$\frac{\left|\mathbf{c}_j^{\mathsf{T}}\widehat{\boldsymbol{\beta}}\right|}{\sqrt{\widehat{\sigma}_e^2\mathbf{c}_j^{\mathsf{T}}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbf{c}_j}} \geq t_{n-p}\left(1-\frac{\alpha}{2q}\right).$$

Of course, the Bonferroni method has the true overall level bounded from above by α, and sometimes the true overall level is much less than α. It is quite possible that H_0 is rejected by the overall F-test but none of the H_{0j} is rejected with the adjusted level α/M.

Now consider individual confidence intervals for the true values of $\mathbf{c}_j^{\mathsf{T}}\boldsymbol{\beta}$. By Lemma 4.1 and Lemma 4.6, these confidence intervals have the form

$$\mathbf{c}_j^{\mathsf{T}}\widehat{\boldsymbol{\beta}} \mp t_{n-p}\big(1-\frac{\alpha}{2}\big)\widehat{\sigma}_e\sqrt{\mathbf{c}_j^{\mathsf{T}}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbf{c}_j}$$

and their coverage probability is $1-\alpha$ (exactly or asymptotically).

From Lemma 4.2 and Lemma 4.7, we can get a confidence region for $\mathbb{C}\boldsymbol{\beta}$ with coverage probability $1-\alpha$. It includes all values $\mathbb{C}\boldsymbol{\beta} \in \mathbb{R}^q$ such that

$$\frac{1}{q\widehat{\sigma}_e^2}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta})^{\mathsf{T}}\big[\mathbb{C}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{C}^{\mathsf{T}}\big]^{-1}(\mathbb{C}\widehat{\boldsymbol{\beta}}-\mathbb{C}\boldsymbol{\beta}) \leq F_{q,n-p}(1-\alpha).$$

However, this is a hyper-ellipsoid in $\mathbb{R}^q$, which is difficult to describe or evaluate. We would much prefer to have confidence intervals for the individual linear combinations $\mathbf{c}_j^{\mathsf{T}}\boldsymbol{\beta}$, keeping their simultaneous coverage at $\geq 1-\alpha$. Such intervals can be obtained using Bonferroni method. They have the form

$$\mathbf{c}_j^{\mathsf{T}}\widehat{\boldsymbol{\beta}} \mp t_{n-p}\big(1-\frac{\alpha}{2q}\big)\widehat{\sigma}_e\sqrt{\mathbf{c}_j^{\mathsf{T}}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbf{c}_j}$$

and their simultaneous coverage probability is $\geq 1-\alpha$.

Pairwise comparison of means in one-way ANOVA using Bonferroni method

Consider a factor variable $G_i \in \{1, \ldots, m\}$ specifying membership in one of m groups for each observation. Let the expectations of the response in the groups be $\mu_j \equiv \mathsf{E}\left[Y_i \,\middle|\, G_i = j\right]$. The question is: which of the groups differ in the expectations from each other and which are the same? This is one of the most common questions in regression analyses of factor covariates. Suppose for simplicity that there is no other predictor in the model.

We use the one-way ANOVA model with reference group parametrization:

$$\mathsf{E}\left[Y_i \,\middle|\, G_i\right] = \beta_1 + \gamma_2 \mathbb{1}(G_i = 2) + \cdots + \gamma_m \mathbb{1}(G_i = m).$$

The group expectations can be expressed as $\mu_1 = \beta_1$, $\mu_2 = \beta_1 + \gamma_2$, $\ldots$, $\mu_m = \beta_1 + \gamma_m$.

Consider the overall hypothesis $H_0 : \mu_1 = \cdots = \mu_m$ and the $\binom{m}{2}$ partial hypotheses comparing the individual expectations with each other: $H_{jk} : \mu_j = \mu_k$ for $j < k$. Denote by n_j the number of observations in the j-th group. In one-way ANOVA model, we have

$$\begin{aligned}\widehat{\mu}_j &= \frac{1}{n_j} \sum_{i=1}^{n} Y_i \mathbb{1}(G_i = 1) \\ \mathsf{var}\, \widehat{\mu}_j &= \frac{\sigma_e^2}{n_j}\end{aligned}$$

Also, $\widehat{\mu}_j$ and $\widehat{\mu}_k$ are independent for $j \neq k$.

The hypothesis H_0 is violated if and only if any of the hypotheses H_{jk} is violated. If we reject H_0 by the one-way ANOVA F-test, we want to know which of the H_{jk} can be rejected. Using the Bonferroni method, we test each of these partial hypotheses by a t-test at the level $\alpha^* = \alpha / \binom{m}{2} = \frac{2\alpha}{m(m-1)}$ and we reject $H_{jk} : \mu_j = \mu_k$ when

$$\frac{\left|\widehat{\mu}_j - \widehat{\mu}_k\right|}{\widehat{\sigma}_e\left(\frac{1}{n_j} + \frac{1}{n_k}\right)} \geq t_{n-p}\left(1 - \frac{\alpha}{m(m-1)}\right).$$

It is not difficult to see that for larger m this procedure is very inefficient: it will be rarely able to detect differences between the individual group means unless the differences are very large.

Bonferroni method: summary

The Bonferroni method is extremely general and has no restrictive assumptions. It can be applied to any testing problem and any set of confidence intervals no matter what the model is and no matter how the tests and confidence intervals are constructed.

On the other hand, the Bonferroni method is very conservative. The resulting tests have low power and the confidence intervals are unnecessarily wide. The larger is the number of partial tests or confidence intervals M the poorer is the performance of this procedure.

In the next two sections we introduce two alternative methods that are specifically designed for linear regression.

9.2. Tukey Method

The Tukey method Tukey (1949) is specifically designed to address the problem of pairwise comparisons of the expectations in one-way ANOVA.*

***John Tukey** (1915 – 2000) was an American mathematician and statistician, best known for the development of the fast Fourier Transform (FFT) algorithm and the box plot. The Tukey HSD test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma all bear his name. He is also credited with coining the term* bit *and the first published use of the word* software.
Source: https://en.wikipedia.org/wiki/John_Tukey

Tukey method: introduction

Consider M independent variables $T_j \sim \mathrm{N}(\mu_j, \sigma^2)$, $j = 1, \ldots, M$. Let $\widehat{\sigma}^2$ be an estimator of σ^2 such that

(i) $\frac{f\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_f$ for some natural number f and
(ii) $\widehat{\sigma}^2$ is independent of $(T_1, \ldots, T_M)$.

Normalize the observations and look at their range. Define

$$Q = \frac{\max\limits_{j=1,\ldots,M}\left(\frac{T_j-\mu_j}{\sigma}\right) - \min\limits_{j=1,\ldots,M}\left(\frac{T_j-\mu_j}{\sigma}\right)}{\widehat{\sigma}/\sigma} = \frac{\max\limits_{j=1,\ldots,M}(T_j-\mu_j) - \min\limits_{j=1,\ldots,M}(T_j-\mu_j)}{\widehat{\sigma}}$$

Because $\frac{T_j-\mu_j}{\sigma}$ are independent variables distributed as $\mathrm{N}(0,1)$ and $\frac{\widehat{\sigma}}{\sigma}$ is a transformation of a random variable with a χ^2_f distribution, which is independent of the variables in the numerator, the following lemma holds.

Lemma 9.1 (Tukey). *The statistic Q has a distribution that depends on M and f but not on $\mu_1, \ldots, \mu_n$ and σ^2.* ♢

The distribution function of the statistic Q can be calculated. The distribution is called *Tukey distribution*. Denote its quantile function by $q_{f,M}(u)$.

From this, we have

$$\mathrm{P}\left[\max_j(T_j-\mu_j) - \min_j(T_j-\mu_j) < \widehat{\sigma} q_{f,M}(1-\alpha)\right] = 1-\alpha.$$

The difference between the largest and the smallest number from a finite set is smaller than a constant if and only if each of the distances between any pair of them is smaller than the same constant. Hence,

$$\mathrm{P}\left[\left|(T_j-\mu_j)-(T_k-\mu_k)\right| < \widehat{\sigma} q_{f,M}(1-\alpha) \quad \forall j,k\right] = 1-\alpha$$

[*] This method is also called Tukey's honest significant differences or Tukey HSD

and

$$\mathrm{P}\big[T_j - T_k - \widehat{\sigma} q_{f,M}(1-\alpha) < \mu_j - \mu_k < T_j - T_k + \widehat{\sigma} q_{f,M}(1-\alpha) \quad \forall j,k\big] = 1-\alpha$$

We have obtained simultaneous confidence intervals for the differences in the expectations and the probability that *all of them* contain the true values is jointly $1-\alpha$.

The confidence intervals can be easily converted into tests. The hypothesis $H_{jk} : \mu_j = \mu_k$ is rejected when

$$\frac{\left|T_j - T_k\right|}{\widehat{\sigma}} > q_{f,M}(1-\alpha).$$

The overall level of all $\binom{M}{2}$ tests is α.

Tukey method: extension

The Tukey method assumes that the observations have all the same variance, which is very restrictive. So, relax the assumption to $T_j \sim \mathsf{N}(\mu_j, v_j\sigma^2)$, where v_j is a known positive constant. We can repeat the derivation of the Tukey method until we get the equality

$$\mathrm{P}\left[\left|\frac{T_j - \mu_j}{\sqrt{v_j}} - \frac{T_k - \mu_k}{\sqrt{v_j}}\right| < \widehat{\sigma} q_{f,M}(1-\alpha) \quad \forall j,k\right] = 1-\alpha.$$

Hayter (1984) has proven that this equality under the current assumptions implies

$$\mathrm{P}\left[\left|(T_j - \mu_j) - (T_k - \mu_k)\right| < \widehat{\sigma}\sqrt{\frac{v_j + v_k}{2}} q_{f,M}(1-\alpha) \quad \forall j,k\right] \geq 1-\alpha.$$

What appears inside is actually the square root of the average estimated variance in the two groups: $\sqrt{\frac{v_j\widehat{\sigma}^2 + v_k\widehat{\sigma}^2}{2}}$. Thus, we can use the Tukey method even with unequal variances replacing the equal estimated variance by the average of estimated variances in the two groups, otherwise the procedure is unchanged. However, the Tukey method becomes conservative under such circumstances.

Tukey method: application in linear regression

In linear regression, we could take $T_j = \boldsymbol{c}_j^\top \widehat{\boldsymbol{\beta}}$ and try to apply the Tukey method on this. We have $T_j \sim \mathsf{N}(\boldsymbol{c}_j^\top \boldsymbol{\beta}, \sigma_e^2 \boldsymbol{c}_j^\top (\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{c}_j)$. So we take $\mu_j \equiv \boldsymbol{c}_j^\top \boldsymbol{\beta}$ and $v_j \equiv \boldsymbol{c}_j^\top (\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{c}_j$. However, the Tukey method also requires independence between the individual T_j's. This is very difficult to achieve because the T_j's are different linear combinations of the same random vector $\widehat{\boldsymbol{\beta}}$. The only case when this assumption clearly holds is the case of simple one-way ANOVA with μ_j equal to the expectation of the j-th group and $T_j = \widehat{\mu}_j$ is the average of the observations in the j-th group. The constants v_j are defined as $v_j = 1/n_j$.

To conclude, in a simple one-way ANOVA with m groups and no additional covariates, we can use the Tukey method to test the hypotheses $H_{jk} : \mu_j = \mu_k$. We reject H_{jk} when

$$\frac{\left|\widehat{\mu}_j - \widehat{\mu}_k\right|}{\widehat{\sigma}_e\sqrt{\frac{1}{2}\left(\frac{1}{n_j}+\frac{1}{n_k}\right)}} \geq q_{n-m,m}(1-\alpha).$$

These test have an overall level at most α. We can also construct simultaneous confidence intervals for the differences between the group means. The interval for $\mu_j - \mu_k$ is

$$\widehat{\mu}_j - \widehat{\mu}_k \mp \widehat{\sigma}_e\sqrt{\frac{1}{2}\left(\frac{1}{n_j}+\frac{1}{n_k}\right)q_{n-m,m}(1-\alpha)},$$

and the overall coverage of these intervals is $\geq 1-\alpha$.

The Tukey method is theoretically appealing but it has so strong assumptions that it can rarely used in linear regression models. The only case when this method can be applied is for comparison of group means in one-way ANOVA.

9.3. Scheffé Method

Henry Scheffé *(1907 – 1977) was an American statistician. He is known for the Lehmann–Scheffé theorem and Scheffé's method. Scheffé was president of the Institute of Mathematical Statistics in 1954, and also served as vice president of the American Statistical Association from 1954 to 1956.*
Source: https://en.wikipedia.org/wiki/Henry_Scheffé

Scheffé (1953) proposed another method for simultaneous tests and confidence intervals in linear regression. It is based on the following result from matrix algebra.

Lemma 9.2. *Let $\mathbb{A}$ be a symmetric positive definite matrix. Then for any column vector $\boldsymbol{x}$ of a suitable dimension,*

$$\boldsymbol{x}^T\mathbb{A}\boldsymbol{x} \leq 1 \quad \textit{if and only if} \quad (\boldsymbol{c}^T\boldsymbol{x})^2 \leq \boldsymbol{c}^T\mathbb{A}^{-1}\boldsymbol{c} \quad \forall \boldsymbol{c}. \qquad \diamondsuit$$

Proof. (i) Let $\boldsymbol{x}^\top\mathbb{A}\boldsymbol{x} \leq 1$. Consider the Cauchy-Schwartz inequality

$$(\boldsymbol{a}^\top\boldsymbol{b})^2 \leq ||a||^2||b||^2.$$

Apply it with $\boldsymbol{a} = \mathbb{A}^{1/2}\boldsymbol{x}$ and $\boldsymbol{b} = \mathbb{A}^{-1/2}\boldsymbol{c}$. We get

$$(\boldsymbol{x}^\top\boldsymbol{c})^2 \leq (\boldsymbol{x}^\top\mathbb{A}\boldsymbol{x})(\boldsymbol{c}^\top\mathbb{A}^{-1}\boldsymbol{c}) \leq \boldsymbol{c}^\top\mathbb{A}^{-1}\boldsymbol{c}$$

and this inequality holds for any vector $\boldsymbol{c}$.

(i) Let $(\boldsymbol{c}^\top \boldsymbol{x})^2 \le \boldsymbol{c}^\top \mathbb{A}^{-1}\boldsymbol{c} \quad \forall \boldsymbol{c}$. Choose $\boldsymbol{c} = \mathbb{A}\boldsymbol{x}$. We get $(\boldsymbol{x}^\top \mathbb{A}\boldsymbol{x})^2 \le \boldsymbol{x}^\top \mathbb{A}\mathbb{A}^{-1}\mathbb{A}\boldsymbol{x} = \boldsymbol{x}^\top \mathbb{A}\boldsymbol{x}$. When $\boldsymbol{x} = \boldsymbol{0}$ then indeed $0 = \boldsymbol{x}^\top \mathbb{A}\boldsymbol{x} \le 1$. Otherwise, divide the inequality by $\boldsymbol{x}^\top \mathbb{A}\boldsymbol{x} > 0$. □

Theorem 9.3 (Scheffé). *Let $\boldsymbol{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \sigma^2 \mathbb{V})$, where $\mathbb{V} > 0$ is known. Let $\widehat{\sigma}^2$ be an estimator of σ^2 satisfying $\frac{f\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_f$ for some natural number f. Let $\widehat{\sigma}^2$ be independent of $\boldsymbol{X}$. Take $\mathscr{B}$ an m-dimensional linear subspace of $\mathbb{R}^n$ with $m \le n$. Then*

$$\mathrm{P}\left[\left|\boldsymbol{a}^T\boldsymbol{X} - \boldsymbol{a}^T\boldsymbol{\mu}\right| \le \sqrt{m\widehat{\sigma}^2 F_{m,f}(1-\alpha)\,\boldsymbol{a}^T\mathbb{V}\boldsymbol{a}} \quad \forall \boldsymbol{a} \in \mathscr{B}\right] = 1 - \alpha. \qquad \diamondsuit$$

9.4. Simultaneous Confidence Bounds

CITE HUBER AND ARNOLD in CHAP 4!!!!!

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

List of Tables

Bibliography

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

Arnold, S. F. (1980). Asymptotic validity of f tests for the ordinary linear model and the multiple correlation model, *Journal of the American Statistical Association* **75**(372): 890–894.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita, Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze. Firenze: Libr. Internaz. Seeber.

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.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative, *The Annals of Statistics* **12**(1): 61–75.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo, *The Annals of Statistics* **1**(5): 799–821.

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.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**(1/2): 87–104.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance, *Biometrics* **5**(2): 99–114.

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*

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