Stochastics III

Ulm University
Institute of Stochastics

Lecture Notes
Prof. Dr. Volker Schmidt
Winter Term 2012/2013

Ulm, February 2013

Contents

References

[1] Büning, H., Trenkler, G. (1994)
Nichtparametrische statistische Methoden
de Gruyter, Berlin

[2] Cressie, N.A. (1993)
Statistics for Spatial Data
J. Wiley & Sons, New York

[3] Dobson, A.J. (2002)
An Introduction to Generalized Linear Models
Chapman & Hall, Boca Raton

[4] Falk, M., Marohn, F., Tewes, B. (2002)
Foundations of Statistical Analyses and Applications with SAS
Birkhäuser, Basel

[5] Hastie, T., Tibshirami, R., Friedman, J. (2001)
The Elements of Statistical Learnning
Springer, New York

[6] Koch, K.R. (1997)
Parameterschätzung und Hypothesentests in linearen Modellen
Dümmlers–Verlag, Bonn

[7] Lehmann, E.L. (1999)
Elements of Large–Sample Theory
Springer, New York

[8] Lehmann, E.L., Romano, J.P. (2005)
Testing Statistical Hypotheses
Springer, New York

[9] McCullagh, P., Nelder, J.A. (1989)
Genralized Linear Models
Chapman & Hall, London.

[10] Pruscha, H. (2000)
Vorlesungen über mathematische Statistik
Teubner–Verlag, Stuttgart.

[11] Van der Vaart, A., Wellner, J. (1996)
Weak Convergence and Empirical Processes
Springer–Verlag, New York

[12] Vapnik, V.N. (1998)
Statistical Learning Theory
J. Wiley & Sons, New York

1 Introduction and Mathematical Foundations

These lecture notes are made for students who already have a basic knowledge of mathematical statistics. Estimation and statistical test methods which have been discussed in "Stochastik I" are assumed to be known.

The present lecture notes consist of the following parts:

- multivariate normal distribution (nondegenerate and degenerate normal distribution, linear and quadratic forms)
- linear models (multiple regression, normally distributed error terms, single– and multiple–factor analysis of variance)
- generalized linear models (logistic regression, maximum–likelihood equation, weighted least squares estimator, evaluation of the goodness of fit)
- tests for distribution assumptions (Kolmogorow–Smirnow test, χ^2–goodness–of–fit test of Pearson–Fisher)
- nonparametric location tests (binomial test, iteration tests, linear rank tests)

In particular, we will use notions and results which have been introduced in the lecture notes "Elementare Wahrscheinlichkeitsrechnung und Statistik" and "Stochastik I": we will indicate references to these lecture notes by "WR" and "I" in front of the section number of the cited lemmas, theorems, corollaries and formulas.

1.1 Some Basic Notions and Results of Matrix Algebra

First, we recall some basic notions and results of matrix algebra, which are needed in these lecture notes.

1.1.1 Trace and Rank

- The *trace* $\operatorname{tr}(\mathbf{A})$ of a quadratic $n \times n$ matrix $\mathbf{A} = (a_{ij})$ is given by

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \,. \tag{1}$$

- Let $\mathbf{A}$ be an arbitrary $n \times m$ matrix. The *rank* $\operatorname{rk}(\mathbf{A})$ is the maximum number of linearly independent rows (or columns) of $\mathbf{A}$.
  - The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_\ell \in \mathbb{R}^m$ are called *linearly dependent* if there exist real numbers $c_1, \ldots, c_\ell \in \mathbb{R}$, which are not all equal to zero and $c_1\mathbf{a}_1 + \ldots + c_\ell\mathbf{a}_\ell = \mathbf{o}$.
  - Otherwise the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_\ell \in \mathbb{R}^m$ are called *linearly independent*.

From the definition of the trace of a matrix in (1) and from the definition of matrix multiplication the next lemma directly follows.

Lemma 1.1 *Let* $\mathbf{C}$ *be an arbitrary* $n \times m$ *matrix and* $\mathbf{D}$ *an arbitrary* $m \times n$ *matrix. Then* $\operatorname{tr}(\mathbf{CD}) = \operatorname{tr}(\mathbf{DC})$.

It can be proved that a quadratic matrix $\mathbf{A}$ is invertible if and only if $\mathbf{A}$ has full rank or $\det \mathbf{A} \neq 0$, respectively. The following result is also useful in this context.

Lemma 1.2 *Let* $\mathbf{A}$ *be an* $n \times m$ *matrix with* $n \geq m$ *and* $\operatorname{rk}(\mathbf{A}) = m$. *Then* $\operatorname{rk}(\mathbf{A}^\top\mathbf{A}) = m$.

Proof

- It is obvious that the rank $\mathrm{rk}(\mathbf{A}^\top\mathbf{A})$ of the $m \times m$ matrix $\mathbf{A}^\top\mathbf{A}$ cannot exceed m.
- Now, we assume that $\mathrm{rk}(\mathbf{A}^\top\mathbf{A}) < m$. Then, there exists a vector $\mathbf{c} = (c_1, \ldots, c_m)^\top \in \mathbb{R}^m$ with $\mathbf{c} \neq \mathbf{o}$ and $\mathbf{A}^\top\mathbf{A}\mathbf{c} = \mathbf{o}$.
- From this follows that $\mathbf{c}^\top\mathbf{A}^\top\mathbf{A}\mathbf{c} = \mathbf{o}$ and $(\mathbf{A}\mathbf{c})^\top(\mathbf{A}\mathbf{c}) = \mathbf{o}$, i.e., $\mathbf{A}\mathbf{c} = \mathbf{o}$.
- However, this is contradictory to the assumption that $\mathrm{rk}(\mathbf{A}) = m$. □

Furthermore, it can be proved that the following properties of trace and rank are valid.

Lemma 1.3 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be arbitrary* $n \times n$ *matrices. Then* $\mathrm{tr}(\mathbf{A} - \mathbf{B}) = \mathrm{tr}(\mathbf{A}) - \mathrm{tr}(\mathbf{B})$ *always holds. If* $\mathbf{A}$ *is idempotent and symmetric, i.e.,* $\mathbf{A} = \mathbf{A}^2$ *and* $\mathbf{A} = \mathbf{A}^\top$*, it also holds that* $\mathrm{tr}(\mathbf{A}) = \mathrm{rk}(\mathbf{A})$.

1.1.2 Eigenvalues and Eigenvectors

Definition Let $\mathbf{A}$ be an arbitrary $n \times n$ matrix. Each (complex) number $\lambda \in \mathbb{C}$ is called an *eigenvalue* of the matrix $\mathbf{A}$ if and only if there exists a vector $\mathbf{x} \in \mathbb{C}^n$ with $\mathbf{x} \neq \mathbf{o}$ and

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{o}\,. \tag{2}$$

We call $\mathbf{x}$ an *eigenvector* corresponding to λ.

Remark

- Only if λ is a solution of the so–called *characteristic equation*

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0\,, \tag{3}$$

there is a solution $\mathbf{x} \in \mathbb{C}^n$ with $\mathbf{x} \neq \mathbf{o}$ for (2). The left–hand side $P(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$ of (3) is called the *characteristic polynomial* of matrix $\mathbf{A}$.

- Let $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ be the real–valued solutions of (3). Then the characteristic polynomial can be written in the form

$$P(\lambda) = (-1)^n(\lambda - \lambda_1)^{a_1} \ldots (\lambda - \lambda_k)^{a_k} q(\lambda)\,, \tag{4}$$

where $a_1, \ldots, a_k \in \mathbb{N}$ are positive natural numbers, the so–called *algebraic multiplicities* of $\lambda_1, \ldots, \lambda_k$, and $q(\lambda)$ is a polynomial of order $n - \sum_{i=1}^{k} a_i$ which has no real solutions.

Lemma 1.4 *Let* $\mathbf{A} = (a_{ij})$ *be a symmetric* $n \times n$ *matrix with real–valued entries* a_{ij}*. Then every eigenvalue is a real number and eigenvectors* $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$ *which correspond to different eigenvalues* $\lambda_i, \lambda_j \in \mathbb{R}$ *are orthogonal to each other.*

Proof

- The determinant $\det(\mathbf{A} - \lambda\mathbf{I})$ in (3) is given by

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \sum_{\boldsymbol{\pi}} (-1)^{r(\boldsymbol{\pi})} \prod_{i:\, i \neq \pi_i} a_{i\pi_i} \prod_{i:\, i = \pi_i} (a_{i\pi_i} - \lambda)\,, \tag{5}$$

where the summation extends over all $m!$ permutations $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ of the natural numbers $1, \ldots, m$ and $r(\boldsymbol{\pi})$ is the number of pairs in $\boldsymbol{\pi}$, which are not in the natural order.

- Since the elements of $\mathbf{A}$ are real numbers, every solution $\lambda = a + \mathrm{i}\, b$ of (3) implies another solution $\overline{\lambda} = a - \mathrm{i}\, b$ of (3).
- Let $\mathbf{x} = \mathbf{a} + \mathrm{i}\, \mathbf{b}$ and $\overline{\mathbf{x}} = \mathbf{a} - \mathrm{i}\, \mathbf{b}$ be eigenvectors which correspond to λ or $\overline{\lambda}$, respectively. Then $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$ and $\mathbf{A}\overline{\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}}$ or
$$\overline{\mathbf{x}}^\top \mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^\top \lambda \mathbf{x} = \lambda \overline{\mathbf{x}}^\top \mathbf{x}$$
and
$$\overline{\mathbf{x}}^\top \mathbf{A}\mathbf{x} = \left(\mathbf{A}^\top \overline{\mathbf{x}}\right)^\top \mathbf{x} = \left(\mathbf{A}\overline{\mathbf{x}}\right)^\top \mathbf{x} = \left(\overline{\lambda}\overline{\mathbf{x}}\right)^\top \mathbf{x} = \overline{\lambda}\overline{\mathbf{x}}^\top \mathbf{x}\,.$$
- From this it follows that $\lambda \overline{\mathbf{x}}^\top \mathbf{x} = \overline{\lambda}\overline{\mathbf{x}}^\top \mathbf{x}$.
- Since $\overline{\mathbf{x}}^\top \mathbf{x} = |\mathbf{a}|^2 + |\mathbf{b}|^2 > 0$, it holds that $\lambda = \overline{\lambda}$, i.e., λ is a real number.
- In a similar way it can be proved that for different eigenvalues $\lambda_i, \lambda_j \in \mathbb{R}$ there exist eigenvectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$ with real–valued components which are orthogonal to each other.
- Since the matrix $\mathbf{A} - \lambda_i \mathbf{I}$ only contains real–valued elements, it holds that if $\mathbf{x}_i$ is an eigenvector which corresponds to λ_i, then also $\overline{\mathbf{x}}_i$ and $\mathbf{x}_i + \overline{\mathbf{x}}_i \in \mathbb{R}^n$ are eigenvectors that correspond to λ_i .
- Therefore we can (and will) assume w.l.o.g. that $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$. Furthermore, if
$$(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_i = \mathbf{o} \qquad \text{and} \qquad (\mathbf{A} - \lambda_j \mathbf{I})\mathbf{x}_j = \mathbf{o},$$
it follows that $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$ and $\mathbf{A}\mathbf{x}_j = \lambda_j \mathbf{x}_j$ as well as
$$\mathbf{x}_j^\top \mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_j^\top \mathbf{x}_i \qquad \text{and} \qquad \mathbf{x}_i^\top \mathbf{A}\mathbf{x}_j = \lambda_j \mathbf{x}_i^\top \mathbf{x}_j\,.$$
- On the other hand it is obvious that $\mathbf{x}_j^\top \mathbf{x}_i = \mathbf{x}_i^\top \mathbf{x}_j$ and with the symmetry of $\mathbf{A} = (a_{ij})$ we get the identity $\mathbf{x}_j^\top \mathbf{A}\mathbf{x}_i = \mathbf{x}_i^\top \mathbf{A}\mathbf{x}_j$ since
$$\mathbf{x}_j^\top \mathbf{A}\mathbf{x}_i = \sum_{m=1}^{n}\sum_{\ell=1}^{n} x_{\ell j} a_{\ell m} x_{mi} = \sum_{\ell=1}^{n}\sum_{m=1}^{n} x_{mi} a_{m\ell} x_{\ell j} = \mathbf{x}_i^\top \mathbf{A}\mathbf{x}_j\,.$$
- Altogether it follows that $\lambda_i \mathbf{x}_j^\top \mathbf{x}_i = \lambda_j \mathbf{x}_i^\top \mathbf{x}_j$ and $(\lambda_i - \lambda_j)\mathbf{x}_j^\top \mathbf{x}_i = 0$.
- As $\lambda_i - \lambda_j \neq 0$, it holds that $\mathbf{x}_j^\top \mathbf{x}_i = 0$. □

1.1.3 Diagonalization Method

- Now, let $\mathbf{A}$ be an invertible symmetric $n \times n$ matrix.
- In Lemma 1.4 we have shown that all eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{A}$ are real numbers (where it is possible that one number occurs more than once in this sequence).
- Since $\det \mathbf{A} \neq 0$, we get that $\lambda = 0$ is no solution of (3), i.e., all eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{A}$ are different from zero.
- Furthermore, it can be proved that there are *orthonormal* (basis) vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$, i.e.,
$$\mathbf{v}_i^\top \mathbf{v}_i = 1\,, \qquad \mathbf{v}_i^\top \mathbf{v}_j = 0\,, \qquad \forall\, i, j \in \{1, \ldots, n\} \text{ with } i \neq j\,, \tag{6}$$
such that $\mathbf{v}_i$ is an eigenvector that corresponds to λ_i; $i = 1, \ldots, n$.
- If all eigenvalues $\lambda_1, \ldots, \lambda_n$ differ from each other, then this is an immediate consequence of part 2 of Lemma 1.4.
- As a consequence, the following *diagonalization method* for invertible symmetric matrices is obtained.

Lemma 1.5

- *Let* $\mathbf{A}$ *be an invertible symmetric* $n \times n$ *matrix and let* $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ *be the* $n \times n$ *matrix that consists of the orthonormal eigenvalues* $\mathbf{v}_1, \ldots, \mathbf{v}_n$.
- *Then*

$$\mathbf{V}^\top \mathbf{A} \mathbf{V} = \mathbf{\Lambda}\,, \tag{7}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ *denotes the* $n \times n$ *diagonal matrix which consists of the eigenvalues* $\lambda_1, \ldots, \lambda_n$.

Proof

- Equation (2) in the definition of eigenvalues and eigenvectors implies that $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for each $i = 1, \ldots, n$.
- This means that $\mathbf{AV} = (\lambda_1 \mathbf{v}_1, \ldots, \lambda_n \mathbf{v}_n)$ and with (6) it follows that $\mathbf{V}^\top \mathbf{AV} = \mathbf{V}^\top (\lambda_1 \mathbf{v}_1, \ldots, \lambda_n \mathbf{v}_n) = \mathbf{\Lambda}$. □

1.1.4 Symmetric and Definite Matrices; Factorization

Lemma 1.6 *Let* $\mathbf{A}$ *be a symmetric and positive definite* $n \times n$ *matrix, i.e.,* $\mathbf{A} = \mathbf{A}^\top$ *and* $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ *for each vector* $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ *with* $\mathbf{x} \neq \mathbf{o}$. *Then* $\mathbf{A}$ *is invertible and there is an invertible* $n \times n$ *matrix* $\mathbf{H}$, *such that*

$$\mathbf{A} = \mathbf{H}\mathbf{H}^\top\,. \tag{8}$$

Proof We only prove the second part of Lemma 1.6.

- Lemma 1.5 implies that $\mathbf{V}^\top \mathbf{AV} = \mathbf{\Lambda}$ and

$$\mathbf{A} = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda} \mathbf{V}^{-1}\,, \tag{9}$$

  - where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is the $n \times n$ matrix which consists of the orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$,
  - and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the $n \times n$ diagonal matrix which consists of the (positive) eigenvalues $\lambda_1, \ldots, \lambda_n$.
- Now, let $\mathbf{\Lambda}^{1/2}$ be the $n \times n$ diagonal matrix $\mathbf{\Lambda}^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ and let

$$\mathbf{H} = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top\,. \tag{10}$$

- It is obvious that the matrix $\mathbf{H}$, given in (10), is invertible. Because of $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$ it also holds that

$$\begin{aligned} \mathbf{H}\mathbf{H}^\top &= (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top \Big((\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top \Big)^\top = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{V}^\top \mathbf{V} \mathbf{\Lambda}^{1/2} \mathbf{V}^{-1} \\ &= (\mathbf{V}^\top)^{-1} \mathbf{\Lambda}^{1/2} \mathbf{\Lambda}^{1/2} \mathbf{V}^{-1} = (\mathbf{V}^\top)^{-1} \mathbf{\Lambda} \mathbf{V}^{-1} = \mathbf{A}\,, \end{aligned}$$

  where the last equality follows from (9). □

Remark

- Each invertible $n \times n$ matrix $\mathbf{H}$ with $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$ is called a *square root* of $\mathbf{A}$ and is denoted by $\mathbf{A}^{1/2}$.
- Using the *Cholesky decomposition* for symmetric and positive definite matrices, one can show that there exists a (uniquely determined) lower triangular matrix $\mathbf{H}$ with $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$.

The following property of symmetric matrices is a generalization of Lemma 1.6.

Lemma 1.7 *Let* $\mathbf{A}$ *be a symmetric and positive semidefinite* $n \times n$ *matrix, i.e., it holds that* $\mathbf{A} = \mathbf{A}^\top$ *and* $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ *for each vector* $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$*. Now, let* $\mathrm{rk}(\mathbf{A}) = r \ \big(\leq n\big)$*. Then there exists an* $n \times r$ *matrix* $\mathbf{H}$ *with* $\mathrm{rk}(\mathbf{H}) = r$*, such that* $\mathbf{A} = \mathbf{H}\mathbf{H}^\top$*.*

The *proof* of Lemma 1.7 is similar to the proof of Lemma 1.6.

Lemma 1.8

- *Let* $m, r \in \mathbb{N}$ *be arbitrary natural numbers with* $1 \leq r \leq m$*. Let* $\mathbf{A}$ *be a symmetric and positive definite* $m \times m$ *matrix and let* $\mathbf{B}$ *be an* $r \times m$ *matrix with full rank* $\mathrm{rk}(\mathbf{B}) = r$*.*
- *Then also the matrices* $\mathbf{B}\mathbf{A}\mathbf{B}^\top$ *and* $\mathbf{A}^{-1}$ *are positive definite.*

Proof

- Because of the full rank of $\mathbf{B}^\top$ it holds that $\mathbf{B}^\top \mathbf{x} \neq \mathbf{o}$ for each $\mathbf{x} \in \mathbb{R}^r$ with $\mathbf{x} \neq \mathbf{o}$.
- Since $\mathbf{A}$ is positive definite, it also holds that

  $$\mathbf{x}^\top \big(\mathbf{B}\mathbf{A}\mathbf{B}^\top\big)\mathbf{x} = (\mathbf{B}^\top \mathbf{x})^\top \mathbf{A}(\mathbf{B}^\top \mathbf{x}) > 0$$

  for each $\mathbf{x} \in \mathbb{R}^r$ with $\mathbf{x} \neq \mathbf{o}$, i.e., $\mathbf{B}\mathbf{A}\mathbf{B}^\top$ is positive definite.
- Therefore, we get for $\mathbf{B} = \mathbf{A}^{-1}$ that

  $$\mathbf{A}^{-1} = \mathbf{A}^{-1}\big(\mathbf{A}\mathbf{A}^{-1}\big) = \mathbf{A}^{-1}\mathbf{A}\big(\mathbf{A}^{-1}\big)^\top$$

  is positive definite. □

1.2 Multivariate Normal Distribution

In this section we recall the notion of a multivariate normal distribution and discuss some fundamental properties of this family of distributions.

1.2.1 Definition and Fundamental Properties

- Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be independent and (identically) normally distributed random variables, i.e.,

  $$X_i \sim \mathrm{N}(\mu, \sigma^2)\,, \qquad \forall i = 1, \ldots, n\,, \tag{11}$$

  where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.
- The assumption of normality in (11) and the independence of the sampling variables $X_1, \ldots, X_n$ mean in vector notation that the distribution of the random sample $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is given by

  $$\mathbf{X} \sim \mathrm{N}\big(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\big)\,, \tag{12}$$

  where $\boldsymbol{\mu} = (\mu, \ldots, \mu)^\top$ and $\mathrm{N}\big(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\big)$ denotes the n–dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2 \mathbf{I}_n$.
- *Recall* (cf. Section WR-4.3.4): In general, the n–dimensional normal distribution is defined as follows.
  - Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ be an arbitrary vector and let $\mathbf{K}$ be a symmetric and positive definite $n \times n$-matrix.

– Let $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ be an absolutely continuous random vector, where the joint density of $\mathbf{Z}$ is given by

$$f(\mathbf{z}) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \frac{1}{\sqrt{\det \mathbf{K}}} \exp\Big(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{z}-\boldsymbol{\mu})\Big) \tag{13}$$

for each $\mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{R}^n$.

– Then the random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ is called (nondegenerate) *normally distributed*.

– Notation: $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$

Now, we show that the function given in (13) is an (n–dimensional) probability density.

Theorem 1.1 *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ be an arbitrary vector and let $\mathbf{K}$ be a symmetric and positive definite $n \times n$-matrix. Then it holds that*

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\, dx_1 \ldots dx_n = (2\pi)^{n/2} (\det \mathbf{K})^{1/2}\,. \tag{14}$$

Proof

- Since $\mathbf{K}$ is symmetric and positive definite (and therefore invertible), Lemma 1.5 implies that there exists an $n \times n$ matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ consisting of the orthogonal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbf{K}$, such that

$$\mathbf{V}^\top \mathbf{K} \mathbf{V} = \boldsymbol{\Lambda}\,, \tag{15}$$

  where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the $n \times n$ diagonal matrix which is built up of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{K}$.
- Since $\mathbf{K}$ is positive definite, it holds that $\mathbf{v}_i^\top \mathbf{K} \mathbf{v}_i = \lambda_i > 0$ for each $i = 1, \ldots, n$, i.e., all eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{K}$ are positive.
- Because of $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$, it holds $\mathbf{V}^\top = \mathbf{V}^{-1}$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$, respectively.
- Due to the fact that $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ and due to (15), it follows that

$$\big(\mathbf{V}^\top \mathbf{K} \mathbf{V}\big)^{-1} = \mathbf{V}^\top \mathbf{K}^{-1} \mathbf{V} = \mathrm{diag}\big(\lambda_1^{-1}, \ldots, \lambda_n^{-1}\big)\,.$$

- The mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ with $\mathbf{y} = \varphi(\mathbf{x}) = \mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu})$, i.e., $\mathbf{x} - \boldsymbol{\mu} = \mathbf{V}\mathbf{y}$, maps $\mathbb{R}^n$ bijectively onto itself and for the Jacobian determinant of $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ it holds that

$$\det\Big(\frac{\partial \varphi_i}{\partial x_j}(x_1, \ldots, x_n)\Big) = \det \mathbf{V} = \pm 1\,,$$

  where the last equality follows from the fact that $1 = \det(\mathbf{V}^\top \mathbf{V}) = (\det \mathbf{V})^2$.
- Therefore, the integral on the left–hand side of (14) can be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\, dx_1 \ldots dx_n \\
&= \int_{\mathbb{R}^n} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\, d(x_1, \ldots, x_n) = \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{y_i^2}{\lambda_i}\right) d(y_1, \ldots, y_n) \\
&= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{y_i^2}{\lambda_i}\right) dy_1 \ldots dy_n = \prod_{i=1}^n (2\pi\lambda_i)^{1/2}\,.
\end{aligned}$$

- This implies (14) since

$$\prod_{i=1}^n \lambda_i = \det \boldsymbol{\Lambda} = \det\big(\mathbf{V}^\top \mathbf{K} \mathbf{V}\big) = \det\big(\mathbf{V}^\top \mathbf{V}\big) \det \mathbf{K} = \det \mathbf{K}\,. \qquad \square$$

1.2.2 Characteristics of the Multivariate Normal Distribution

- Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ be an arbitrary vector and let $\mathbf{K} = (k_{ij})$ be a symmetric and positive definite $n \times n$ matrix.
- First, we determine the characteristic function of a normally distributed random vector.
- *Recall*: The *characteristic function* $\varphi : \mathbb{R}^n \to \mathbb{C}$ of an arbitrary n–dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top : \Omega \to \mathbb{R}^n$ is given by

$$\varphi(\mathbf{t}) = \mathbb{E}\,\exp\bigl(\mathrm{i}\,\mathbf{t}^\top \mathbf{X}\bigr) = \mathbb{E}\,\exp\left(\mathrm{i}\,\sum_{\ell=1}^{n} t_\ell X_\ell\right), \qquad \forall\,\mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n\,. \tag{16}$$

Theorem 1.2

- *Let the random vector* $\mathbf{X} = (X_1, \ldots, X_n)^\top : \Omega \to \mathbb{R}^n$ *be normally distributed with* $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$.
- *Then the characteristic function* $\varphi : \mathbb{R}^n \to \mathbb{C}$ *of* $\mathbf{X}$ *fulfills*

$$\varphi(\mathbf{t}) = \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\,\mathbf{t}^\top \mathbf{K}\mathbf{t}\Bigr), \qquad \forall\,\mathbf{t} \in \mathbb{R}^n\,. \tag{17}$$

Proof

- Equations (13) and (16) imply

$$\begin{aligned}
\varphi(\mathbf{t}) &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\left(\mathrm{i}\,\sum_{\ell=1}^{n} t_\ell x_\ell\right) f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\
&= \frac{1}{(2\pi)^{n/2}\,(\det \mathbf{K})^{1/2}} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top \mathbf{x} - \frac{1}{2}\,(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Bigr)\, dx_1 \ldots dx_n \\
&= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det \mathbf{K})^{1/2}} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\Bigl(\mathrm{i}\,\mathbf{t}^\top \mathbf{y} - \frac{1}{2}\,\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}\Bigr)\, dy_1 \ldots dy_n\,,
\end{aligned}$$

  where the last equality holds due to the substitution $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$, for which the matrix of the partial derivatives is the identity matrix and therefore the Jacobian determinant is equal to 1.
- Similar to the proof of Theorem 1.1 it follows with the help of the substitutions $\mathbf{y} = \mathbf{V}\mathbf{x}$ and $\mathbf{t} = \mathbf{V}\mathbf{s}$ that

$$\begin{aligned}
\varphi(\mathbf{t}) &= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det \mathbf{K})^{1/2}} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\Bigl(\mathrm{i}\,\mathbf{s}^\top \mathbf{x} - \frac{1}{2}\,\mathbf{x}^\top \mathbf{V}^\top \mathbf{K}^{-1}\mathbf{V}\mathbf{x}\Bigr)\, dx_1 \ldots dx_n \\
&= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det \mathbf{K})^{1/2}} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\Bigl(\sum_{\ell=1}^{n}\Bigl(\mathrm{i}\,s_\ell x_\ell - \frac{x_\ell^2}{2\lambda_\ell}\Bigr)\Bigr)\, dx_1 \ldots dx_n
\end{aligned}$$

  and thus

$$\begin{aligned}
\varphi(\mathbf{t}) &= \frac{\exp(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu})}{(2\pi)^{n/2}\,(\det \mathbf{K})^{1/2}} \prod_{\ell=1}^{n} \int_{-\infty}^{\infty} \exp\Bigl(\mathrm{i}\,s_\ell x_\ell - \frac{x_\ell^2}{2\lambda_\ell}\Bigr)\, dx_\ell \\
&= \exp(\mathrm{i}\,\mathbf{t}^\top \boldsymbol{\mu}) \prod_{\ell=1}^{n} \frac{1}{\sqrt{2\pi\lambda_\ell}} \int_{-\infty}^{\infty} \exp\Bigl(\mathrm{i}\,s_\ell x_\ell - \frac{x_\ell^2}{2\lambda_\ell}\Bigr)\, dx_\ell\,,
\end{aligned}$$

  where the matrix $\mathbf{V}$ consists of the orthonormal eigenvectors of $\mathbf{K}$ and $\lambda_1, \ldots, \lambda_n > 0$ are the eigenvalues of $\mathbf{K}$ with $\det \mathbf{K} = \lambda_1 \cdot \ldots \cdot \lambda_n$.

- Now, it is sufficient to consider that $\varphi_\ell : \mathbb{R} \to \mathbb{C}$ with

$$\varphi_\ell(s) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\lambda_\ell}} \exp\Big(\mathrm{i}\, sx - \frac{x^2}{2\lambda_\ell}\Big)\, dx$$

  is the characteristic function of the (one dimensional) $\mathrm{N}(0, \lambda_\ell)$–distribution.
- In Section WR-5.3.3 we already have seen that $\varphi_\ell(s) = \exp\big(-\lambda_\ell s^2/2\big)$.
- Hence, we get

$$\begin{aligned} \varphi(\mathbf{t}) &= \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}) \prod_{\ell=1}^{n} \exp\Big(-\frac{\lambda_\ell s_\ell^2}{2}\Big) = \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})\, \exp\Big(-\frac{\sum_{\ell=1}^{n} \lambda_\ell s_\ell^2}{2}\Big) \\ &= \exp(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu})\, \exp\Big(-\frac{\mathbf{t}^\top\mathbf{K}\mathbf{t}}{2}\Big)\,. \end{aligned}$$

□

Using (17) for the characteristic function we are able to determine expectation and covariance matrix of a normally distributed random vector.

Corollary 1.1 *If* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, *it holds for arbitrary* $i, j = 1, \ldots, n$ *that*

$$\mathbb{E}\, X_i = \mu_i\,, \qquad \textit{and} \qquad \mathrm{Cov}\,(X_i, X_j) = k_{ij}\,. \tag{18}$$

Proof

- From (17) it follows that

$$\frac{\partial\,\varphi(\mathbf{t})}{\partial t_i} = \Big(\mathrm{i}\,\mu_i - \sum_{\ell=1}^{n} k_{i\ell} t_\ell\Big)\varphi(\mathbf{t}) \tag{19}$$

  and

$$\frac{\partial^2\varphi(\mathbf{t})}{\partial t_i \partial t_j} = -\,k_{ij}\varphi(\mathbf{t}) + \Big(\mathrm{i}\,\mu_i - \sum_{\ell=1}^{n} k_{i\ell} t_\ell\Big)\Big(\mathrm{i}\,\mu_j - \sum_{\ell=1}^{n} k_{j\ell} t_\ell\Big)\varphi(\mathbf{t})\,. \tag{20}$$

- It is easy to see that

$$\mathbb{E}\, X_i = \mathrm{i}^{-1}\, \frac{\partial\,\varphi(\mathbf{t})}{\partial t_i}\,\Big|_{\,\mathbf{t}=\mathbf{o}}\,.$$

  Because of $\varphi(\mathbf{o}) = 1$ and (19), it follows that $\mathbb{E}\, X_i = \mu_i$.
- Furthermore,

$$\mathbb{E}\,(X_i X_j) = -\,\frac{\partial^2\varphi(\mathbf{t})}{\partial t_i \partial t_j}\,\Big|_{\,\mathbf{t}=\mathbf{o}}\,.$$

  This equation and (20) imply $\mathrm{Cov}\,(X_i, X_j) = k_{ij}$. □

Remark

- In Theorem WR-4.14 we have shown that the covariance matrix $\mathbf{K} = \mathbf{K}_\mathbf{X}$ of an arbitrary random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is always symmetric and positive semidefinite.
- In (13), where the density of the nondegenerate multivariate normal distribution is defined, it is additionally required that the covariance matrix $\mathbf{K}$ is positive definite.
- Here, $\mathbf{K}$ being positive definite is not only sufficient but also necessary to ensure that the matrix $\mathbf{K}$ is invertible, i.e., $\det\mathbf{K} \neq 0$ or $\mathbf{K}$ has full rank.

1.2.3 Marginal Distributions and Independence of Subvectors; Convolution Properties

- In this section it is shown how to derive further interesting properties of the multivariate normal distribution using Theorem 1.2.
- For this purpose we need a *vectorial version* of the uniqueness theorem for characteristic functions (cf. Corollary WR-5.5), which we will state without proof.

Lemma 1.9 *Let* $\mathbf{X}, \mathbf{Y} : \Omega \to \mathbb{R}^n$ *be arbitrary random vectors;* $\mathbf{X} = (X_1, \ldots, X_n)^\top$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$. *Then it holds that*

$$\mathbf{X} \overset{\mathrm{d}}{=} \mathbf{Y} \qquad \textit{if and only if} \qquad \varphi_{\mathbf{X}}(\mathbf{t}) = \varphi_{\mathbf{Y}}(\mathbf{t}) \qquad \forall\, \mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n\,, \tag{21}$$

where

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\,\exp\Big(\mathrm{i} \sum_{j=1}^{n} t_j X_j\Big) \qquad \textit{and} \qquad \varphi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}\,\exp\Big(\mathrm{i} \sum_{j=1}^{n} t_j Y_j\Big)$$

are the characteristic functions of $\mathbf{X}$ *and* $\mathbf{Y}$, *respectively.*

First, we show that arbitrary subvectors of normally distributed random vectors are also normally distributed.

- We assume $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$ to be an arbitrary vector and $\mathbf{K} = (k_{ij})$ to be a symmetric and positive definite $n \times n$-matrix.
- It is obvious that the random vector $(X_{\pi_1}, \ldots, X_{\pi_n})^\top$ is normally distributed for each permutation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ of the natural numbers $1, \ldots, n$ if $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is normally distributed.
- Therefore, we can w.l.o.g restrict the examination of the distribution of subvectors of normally distributed random vectors to the examination of the first components.

Corollary 1.2 *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, *where* $\mathbf{K}$ *is positive definite. Then it holds that*

$$(X_1, \ldots, X_m)^\top \sim \mathrm{N}(\boldsymbol{\mu}_m, \mathbf{K}_m) \qquad \forall\, m = 1, \ldots, n\,,$$

where $\boldsymbol{\mu}_m = (\mu_1, \ldots, \mu_m)^\top$ *and* $\mathbf{K}_m$ *denotes the* $m \times m$ *matrix which consists of the first* m *rows and columns of* $\mathbf{K}$.

Proof

- Let $\varphi : \mathbb{R}^n \to \mathbb{C}$ be the characteristic function of $(X_1, \ldots, X_n)^\top$.
- Now, the characteristic function $\varphi_m : \mathbb{R}^m \to \mathbb{C}$ of $(X_1, \ldots, X_m)^\top$ fulfills

$$\varphi_m(\mathbf{t}_m) = \varphi\Big((\mathbf{t}_m, \underbrace{0, \ldots, 0}_{n-m})\Big)\,, \qquad \forall\, \mathbf{t}_m = (t_1, \ldots, t_m))^\top \in \mathbb{R}^m\,.$$

- This result and (17) imply that

$$\varphi_m(\mathbf{t}_m) = \exp\Big(\mathrm{i}\,\mathbf{t}_m^\top \boldsymbol{\mu}_m - \frac{1}{2}\,\mathbf{t}_m^\top \mathbf{K}_m \mathbf{t}_m\Big)\,, \qquad \forall\, \mathbf{t}_m \in \mathbb{R}^m\,.$$

- Since $\mathbf{K}$ is symmetric and positive definite, we know that also the $m \times m$ matrix $\mathbf{K}_m$ is symmetric and positive definite. From this fact and from Theorem 1.2 it follows that the characteristic function of the subvector $(X_1, \ldots, X_m)^\top$ is identical with the characteristic function of the $\mathrm{N}(\boldsymbol{\mu}_m, \mathbf{K}_m)$–distribution.
- The statement follows because of the one-to-one correspondence of characteristic functions and distributions of random vectors (cf. Lemma 1.9). □

There is a simple criterion for two subvectors $(X_1, \ldots, X_m)^\top$ and $(X_{m+1}, \ldots, X_n)^\top$, with $1 \leq m < n$, of the normally distributed random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$ being independent.

Corollary 1.3 *Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a normally distributed random vector with $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$; $\mathbf{K} = (k_{ij})$. The subvectors $(X_1, \ldots, X_m)^\top$ and $(X_{m+1}, \ldots, X_n)^\top$ are independent if and only if $k_{ij} = 0$ for arbitrary $i \in \{1, \ldots, m\}$ and $j \in \{m+1, \ldots, n\}$.*

Proof

- If the subvectors $(X_1, \ldots, X_m)^\top$ and $(X_{m+1}, \ldots, X_n)^\top$ are independent, then the (one–dimensional) random variables X_i and X_j are independent for arbitrary $i \in \{1, \ldots, m\}$ and $j \in \{m+1, \ldots, n\}$.
- Thus, it holds that $\mathrm{Cov}\,(X_i, X_j) = 0$ and Corollary 1.1 implies that $k_{ij} = 0$.
- Let us now assume that $k_{ij} = 0$ for arbitrary $i \in \{1, \ldots, m\}$ and $j \in \{m+1, \ldots, n\}$.
- Then Theorem 1.2 implies that the characteristic function $\varphi(\mathbf{t})$ of $\mathbf{X} = (X_1, \ldots, X_n)^\top$ has the following factorization.
- For each $\mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n$ it holds that

$$\begin{aligned}\varphi(\mathbf{t}) &= \exp\Big(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\,\mathbf{t}^\top\mathbf{K}\mathbf{t}\Big) = \exp\Big(\mathrm{i}\sum_{i=1}^{n} t_i\mu_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} t_i k_{ij} t_j\Big) \\ &= \exp\Big(\mathrm{i}\sum_{i=1}^{m} t_i\mu_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} t_i k_{ij} t_j\Big)\,\exp\Big(\mathrm{i}\sum_{i=m+1}^{n} t_i\mu_i - \frac{1}{2}\sum_{i=m+1}^{n}\sum_{j=m+1}^{n} t_i k_{ij} t_j\Big)\,,\end{aligned}$$

  where the factors of the last term are the characteristic functions of $(X_1, \ldots, X_m)^\top$ and $(X_{m+1}, \ldots, X_n)^\top$.
- The statement follows because of the one-to-one correspondence of characteristic functions and distributions of random vectors (cf. Lemma 1.9). □

Remark

- Finally, we show that the family of multivariate normal distributions is closed under convolution. In the following we call this property briefly "convolution stability" of the multivariate normal distribution. In Corollary WR-3.2 we already have proved the convolution stability of one–dimensional normal distributions.
- The following formula for the characteristic function of sums of independent random vectors is useful in this context. The proof is analog to the proof of the one–dimensional case (cf. Theorem WR-5.18).

Lemma 1.10 *Let $\mathbf{Z}_1, \mathbf{Z}_2 : \Omega \to \mathbb{R}^n$ be independent random vectors. The characteristic function $\varphi_{\mathbf{Z}_1+\mathbf{Z}_2} : \mathbb{R}^n \to \mathbb{C}$ of the sum $\mathbf{Z}_1 + \mathbf{Z}_2$ can then be written as*

$$\varphi_{\mathbf{Z}_1+\mathbf{Z}_2}(\mathbf{t}) = \varphi_{\mathbf{Z}_1}(\mathbf{t})\;\varphi_{\mathbf{Z}_2}(\mathbf{t})\,, \qquad \forall\,\mathbf{t} \in \mathbb{R}^n\,, \tag{22}$$

where $\varphi_{\mathbf{Z}_i}$ denotes the characteristic function of $\mathbf{Z}_i$; $i = 1, 2$.

The following statement is called *convolution stability* of the multivariate normal distribution.

Corollary 1.4 *Let $\mathbf{Z}_1, \mathbf{Z}_2 : \Omega \to \mathbb{R}^n$ be independent random vectors with $\mathbf{Z}_i \sim \mathrm{N}(\boldsymbol{\mu}_i, \mathbf{K}_i)$ for $i = 1, 2$. Then it holds that $\mathbf{Z}_1 + \mathbf{Z}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \mathbf{K}_1 + \mathbf{K}_2)$.*

Proof

- Equations (17) and (22) imply that

$$\begin{aligned}\varphi_{\mathbf{Z}_1+\mathbf{Z}_2}(\mathbf{t}) &= \varphi_{\mathbf{Z}_1}(\mathbf{t})\,\varphi_{\mathbf{Z}_2}(\mathbf{t}) \\ &= \exp\Big(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}_1 - \frac{1}{2}\,\mathbf{t}^\top\mathbf{K}_1\mathbf{t}\Big)\,\exp\Big(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu}_2 - \frac{1}{2}\,\mathbf{t}^\top\mathbf{K}_2\mathbf{t}\Big) \\ &= \exp\Big(\mathrm{i}\,\mathbf{t}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) - \frac{1}{2}\,\mathbf{t}^\top(\mathbf{K}_1+\mathbf{K}_2)\mathbf{t}\Big)\,.\end{aligned}$$

- This result and the uniqueness theorem for characteristic functions (cf. Lemma 1.9) imply the statement. □

1.2.4 Linear Transformation of Normally Distributed Random Vectors

Now, we show that the linear transformation of a normally distributed random vector again is a normally distributed random vector.

Theorem 1.3

- *Let* $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ *be an* n*–dimensional normally distributed random vector with mean vector* $\boldsymbol{\mu} \in \mathbb{R}^n$ *and (positive definite) covariance matrix* $\mathbf{K}$*.*
- *Moreover, let* $m \leq n$*, let* $\mathbf{A}$ *be an arbitrary* $m \times n$ *matrix having full rank* $\mathrm{rk}(\mathbf{A}) = m$ *and let* $\mathbf{c} \in \mathbb{R}^m$ *be an arbitrary* m*–dimensional vector.*
- *Then it holds that* $\mathbf{Z} = \mathbf{A}\mathbf{Y} + \mathbf{c}$ *is an (*m*–dimensional) normally distributed random vector with*

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{c},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top)\,. \tag{23}$$

Proof

- For each $\mathbf{a} \in \mathbb{R}^m$ it holds that

$$\varphi_{\mathbf{Z}}(\mathbf{t}) = \exp(\mathrm{i}\,\mathbf{t}^\top\mathbf{a})\varphi_{\mathbf{Z}-\mathbf{a}}(\mathbf{t})\,, \qquad \forall\,\mathbf{t} \in \mathbb{R}^m\,.$$

- From (17) derived in Theorem 1.2 and from the uniqueness theorem for the characteristic function of normally distributed random vectors it follows that

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{c},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top) \quad \text{if and only if} \quad \mathbf{Z} - (\mathbf{A}\boldsymbol{\mu} + \mathbf{c}) \sim \mathrm{N}(\mathbf{o},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top)\,.$$

- Therefore, we will w.l.o.g. assume that $\mathbf{Y} \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$ and $\mathbf{c} = \mathbf{o}$.
- Then the characteristic function $\varphi_{\mathbf{Z}}(\mathbf{t})$ of $\mathbf{Z} = \mathbf{A}\mathbf{Y}$ fulfills

$$\begin{aligned}\varphi_{\mathbf{Z}}(\mathbf{t}) &= \mathbb{E}\,e^{\mathrm{i}\,\mathbf{t}^\top\mathbf{Z}} \\ &= \mathbb{E}\,e^{\mathrm{i}\,\mathbf{t}^\top\mathbf{A}\mathbf{Y}} = \mathbb{E}\,e^{\mathrm{i}\,(\mathbf{A}^\top\mathbf{t})^\top\mathbf{Y}} \\ &= \varphi_{\mathbf{Y}}(\mathbf{A}^\top\mathbf{t})\,,\end{aligned}$$

  for each $\mathbf{t} \in \mathbb{R}^m$, where $\varphi_{\mathbf{Y}}(\mathbf{A}^\top\mathbf{t})$ denotes the value of the characteristic function of the normally distributed random vector $\mathbf{Y}$ at $\mathbf{A}^\top\mathbf{t} \in \mathbb{R}^n$.

- Now, formula (17) for the characteristic function of normally distributed random vectors implies

$$\begin{aligned}\varphi_{\mathbf{Z}}(\mathbf{t}) &= \varphi_{\mathbf{Y}}(\mathbf{A}^\top \mathbf{t}) \\ &= \exp\Big(-\frac{1}{2}(\mathbf{A}^\top \mathbf{t})^\top \mathbf{K}(\mathbf{A}^\top \mathbf{t})\Big) \\ &= \exp\Big(-\frac{1}{2}\mathbf{t}^\top(\mathbf{A}\mathbf{K}\mathbf{A}^\top)\mathbf{t}\Big)\,.\end{aligned}$$

- In other words: The characteristic function of $\mathbf{Z}$ is equal to the characteristic function of $\mathrm{N}(\mathbf{o},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top)$.
- The uniqueness theorem for characteristic functions of random vectors implies $\mathbf{Z} \sim \mathrm{N}(\mathbf{o},\, \mathbf{A}\mathbf{K}\mathbf{A}^\top)$. □

By using Theorem 1.3 it follows in particular that it is possible to create normally distributed random vectors by a linear transformation of vectors whose components are independent $\mathrm{N}(0,1)$–distributed random variables.

Corollary 1.5

- *Let $Y_1, \ldots, Y_n : \Omega \to \mathbb{R}$ be independent random variables with $Y_i \sim \mathrm{N}(0,1)$ for each $i = 1, \ldots, n$, i.e., $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \sim \mathrm{N}(\mathbf{o}, \mathbf{I})$.*
- *Let $\mathbf{K}$ be a symmetric and positive definite $n \times n$ matrix and let $\boldsymbol{\mu} \in \mathbb{R}^n$.*
- *Then the random vector $\mathbf{Z} = \mathbf{K}^{1/2}\mathbf{Y} + \boldsymbol{\mu}$ satisfies $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu},\, \mathbf{K})$, where $\mathbf{K}^{1/2}$ is the square root of $\mathbf{K}$.*

Proof

- With the help of Theorem 1.3 it follows that

$$\mathbf{Z} \sim \mathrm{N}\big(\boldsymbol{\mu},\, \mathbf{K}^{1/2}\big(\mathbf{K}^{1/2}\big)^\top\big)\,.$$

- Now, this result and Lemma 1.6 imply the statement. □

1.2.5 Degenerate Multivariate Normal Distribution

In the following, we will give a generalization of the notation of (nondegenerate) multivariate normal distributions, which was introduced in Section 1.2.1.

- A factorization property of covariance matrices which has already been mentioned in Lemma 1.7 is useful in this context.
- *Recall*: Let $\mathbf{K}$ be a symmetric and positive semidefinite $n \times n$ matrix with $\mathrm{rk}(\mathbf{K}) = r \leq n$. Then there is an $n \times r$ matrix $\mathbf{B}$ with $\mathrm{rk}(\mathbf{B}) = r$, such that

$$\mathbf{K} = \mathbf{B}\mathbf{B}^\top\,. \tag{24}$$

Definition

- Let $\mathbf{Y}$ be an n–dimensional random vector with mean vector $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{Y}$ and covariance matrix $\mathbf{K} = \mathrm{Cov}\,(\mathbf{Y})$, such that $\mathrm{rk}(\mathbf{K}) = r$ with $r \leq n$.
- Then $\mathbf{Y}$ is called normally distributed if $\mathbf{Y} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \mathbf{B}\mathbf{Z}$, where $\mathbf{B}$ is an $n \times r$ matrix with $\mathrm{rk}(\mathbf{B}) = r$ fulfilling (24) and where $\mathbf{Z}$ is an r–dimensional random vector with $\mathbf{Z} \sim \mathrm{N}(\mathbf{o}, \mathbf{I}_r)$.

- We say that $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ follows a *degenerate normal distribution* if $\mathrm{rk}(\mathbf{K}) < n$. (Notation: $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$)

Remark

- If $\mathrm{rk}(\mathbf{K}) = r < n$, then the random vector $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ is *not* absolutely continuous
  - because the values of $\mathbf{Y} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \mathbf{BZ}$ are almost surely (with probability 1) elements of the r–dimensional subset $\{\boldsymbol{\mu} + \mathbf{Bx} : \mathbf{x} \in \mathbb{R}^r\}$ of $\mathbb{R}^n$,
  - i.e., the distribution of $\mathbf{Y}$ has no density with respect to the n–dimensional Lebesgue measure.
  - An example for this is the random vector $\mathbf{Y} = (Z, Z)^\top = \mathbf{B}Z$ with $Z \sim \mathrm{N}(0, \sigma^2)$ and $\mathbf{B} = (1, 1)^\top$, which only takes values on the diagonal $\{(z_1, z_2) \in \mathbb{R}^2 : z_1 = z_2\}$.
- The distribution of the random vector $\boldsymbol{\mu} + \mathbf{BZ}$ does *not* depend on the choice of matrix $\mathbf{B}$ of the factorization (24).
- This is an immediate consequence of both of the following criteria for (degenerate and nondegenerate) multivariate normal distributions.

Theorem 1.4

- *Let $\mathbf{Y}$ be an n–dimensional random vector with mean vector $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{Y}$ and covariance matrix $\mathbf{K} = \mathrm{Cov}\,(\mathbf{Y})$, such that $\mathrm{rk}(\mathbf{K}) = r$ with $r \leq n$.*
- *The random vector $\mathbf{Y}$ is normally distributed if and only if one of the following conditions is fulfilled:*
  1. *The characteristic function $\varphi(\mathbf{t}) = \mathbb{E}\,\exp\Big(\mathrm{i}\sum_{j=1}^n t_j Y_j\Big)$ of $\mathbf{Y}$ is given by*
  $$\varphi(\mathbf{t}) = \exp\Big(\mathrm{i}\,\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\,\mathbf{t}^\top\mathbf{K}\mathbf{t}\Big)\,, \qquad \forall\,\mathbf{t} = (t_1, \ldots, t_n)^\top \in \mathbb{R}^n\,. \tag{25}$$
  2. *The linear function $\mathbf{c}^\top\mathbf{Y}$ of $\mathbf{Y}$ is normally distributed for each $\mathbf{c} \in \mathbb{R}^n$ with $\mathbf{c} \neq \mathbf{o}$ and*
  $$\mathbf{c}^\top\mathbf{Y} \sim \mathrm{N}(\mathbf{c}^\top\boldsymbol{\mu}, \mathbf{c}^\top\mathbf{K}\mathbf{c})\,.$$

The *proof* of Theorem 1.4 is omitted (and left as an exercise).

1.3 Linear and Quadratic Forms of Normally Distributed Random Vectors

1.3.1 Definition, Expectation and Covariance

Definition

- Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ be arbitrary n–dimensional random vectors and let $\mathbf{A}$ be a symmetric $n \times n$ matrix with real–valued entries.
- Then the (real–valued) random variable $\mathbf{Y}^\top\mathbf{AY} : \Omega \to \mathbb{R}$ is called a *quadratic form* of $\mathbf{Y}$ (with respect to $\mathbf{A}$).
- The random variable $\mathbf{Y}^\top\mathbf{AZ} : \Omega \to \mathbb{R}$ is called a *bilinear form* of $\mathbf{Y}$ and $\mathbf{Z}$ (with respect to $\mathbf{A}$).

First, we derive the expectation of quadratic or bilinear forms.

Theorem 1.5 *Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ be arbitrary n–dimensional random vectors and let $\mathbf{A}$ be a symmetric $n \times n$ matrix with real–valued entries. Furthermore, let the mean vectors $\boldsymbol{\mu}_\mathbf{Y} = \mathbb{E}\,\mathbf{Y}$ and $\boldsymbol{\mu}_\mathbf{Z} = \mathbb{E}\,\mathbf{Z}$ as well as the covariance matrices $\mathbf{K}_{\mathbf{YY}} = \big(\mathrm{Cov}\,(Y_i, Y_j)\big)$ and $\mathbf{K}_{\mathbf{ZY}} = \big(\mathrm{Cov}\,(Z_i, Y_j)\big)$ be well–defined. Then it holds that*

$$\mathbb{E}\,\big(\mathbf{Y}^\top\mathbf{AY}\big) = \mathrm{tr}(\mathbf{AK}_{\mathbf{YY}}) + \boldsymbol{\mu}_\mathbf{Y}^\top\mathbf{A}\boldsymbol{\mu}_\mathbf{Y} \qquad \textit{and} \qquad \mathbb{E}\,\big(\mathbf{Y}^\top\mathbf{AZ}\big) = \mathrm{tr}(\mathbf{AK}_{\mathbf{ZY}}) + \boldsymbol{\mu}_\mathbf{Y}^\top\mathbf{A}\boldsymbol{\mu}_\mathbf{Z}\,. \tag{26}$$

Proof

- We only prove the second formula in (26) since the first formula follows as a special case for $\mathbf{Z} = \mathbf{Y}$.
- It obviously holds that $\mathbf{Y}^\top\mathbf{A}\mathbf{Z} = \operatorname{tr}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Z}\big)$. Moreover, from Lemma 1.1 it follows that $\operatorname{tr}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Z}\big) = \operatorname{tr}\big(\mathbf{A}\mathbf{Z}\mathbf{Y}^\top\big)$.
- Altogether we get

$$\begin{aligned}\mathbb{E}\,\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Z}\big) &= \mathbb{E}\,\operatorname{tr}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Z}\big) = \mathbb{E}\,\operatorname{tr}\big(\mathbf{A}\mathbf{Z}\mathbf{Y}^\top\big) = \operatorname{tr}\big(\mathbf{A}\mathbb{E}\,(\mathbf{Z}\mathbf{Y}^\top)\big)\\ &= \operatorname{tr}\big(\mathbf{A}(\mathbf{K}_{\mathbf{Z}\mathbf{Y}} + \boldsymbol{\mu}_{\mathbf{Z}}\boldsymbol{\mu}_{\mathbf{Y}}^\top)\big) = \operatorname{tr}(\mathbf{A}\mathbf{K}_{\mathbf{Z}\mathbf{Y}}) + \boldsymbol{\mu}_{\mathbf{Y}}^\top\mathbf{A}\boldsymbol{\mu}_{\mathbf{Z}}\,.\end{aligned}$$

□

In a similar way it is possible to derive a formula for the covariance of quadratic forms of normally distributed random vectors. The following formulas for the third and fourth mixed moments of the components of centered normally distributed random vectors are useful in this context.

Lemma 1.11 *Let $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$ be a normally distributed random vector with mean vector $\boldsymbol{\mu} = \mathbf{o}$ and with an arbitrary covariance matrix $\mathbf{K} = (k_{ij})$. Then it holds that*

$$\mathbb{E}\,(Z_iZ_jZ_\ell) = 0 \quad \textit{and} \quad \mathbb{E}\,(Z_iZ_jZ_\ell Z_m) = k_{ij}k_{\ell m} + k_{i\ell}k_{jm} + k_{j\ell}k_{im} \qquad \forall\, i,j,\ell,m \in \{1,\ldots,n\}\,. \tag{27}$$

The *proof* of Lemma 1.11 is omitted. It is an immediate consequence of Theorems 1.2 and 1.4, cf. the proof of Corollary 1.1.

Theorem 1.6

- *Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ be an n–dimensional random vector with $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ and let $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b_{ij})$ be arbitrary symmetric $n \times n$ matrices.*
- *Then*

$$\operatorname{Cov}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\, \mathbf{Y}^\top\mathbf{B}\mathbf{Y}\big) = 2\operatorname{tr}(\mathbf{A}\mathbf{K}\mathbf{B}\mathbf{K}) + 4\boldsymbol{\mu}^\top\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}\,. \tag{28}$$

- *In particular, it holds that* $\operatorname{Var}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}\big) = 2\operatorname{tr}\big((\mathbf{A}\mathbf{K})^2\big) + 4\boldsymbol{\mu}^\top\mathbf{A}\mathbf{K}\mathbf{A}\boldsymbol{\mu}$.

Proof

- From the definition of the covariance and from Theorem 1.5 it follows that

$$\begin{aligned}\operatorname{Cov}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\, \mathbf{Y}^\top\mathbf{B}\mathbf{Y}\big) &= \mathbb{E}\,\big((\mathbf{Y}^\top\mathbf{A}\mathbf{Y} - \mathbb{E}\,(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}))(\mathbf{Y}^\top\mathbf{B}\mathbf{Y} - \mathbb{E}\,(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}))\big)\\ &= \mathbb{E}\,\big((\mathbf{Y}^\top\mathbf{A}\mathbf{Y} - \operatorname{tr}(\mathbf{A}\mathbf{K}) - \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}^\top\mathbf{B}\mathbf{Y} - \operatorname{tr}(\mathbf{B}\mathbf{K}) - \boldsymbol{\mu}^\top\mathbf{B}\boldsymbol{\mu})\big)\,.\end{aligned}$$

- With the substitution $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$ or $\mathbf{Y} = \mathbf{Z} + \boldsymbol{\mu}$ it follows that

$$\begin{aligned}\operatorname{Cov}\big(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\, \mathbf{Y}^\top\mathbf{B}\mathbf{Y}\big) &= \mathbb{E}\,\big((\mathbf{Z}^\top\mathbf{A}\mathbf{Z} + 2\boldsymbol{\mu}^\top\mathbf{A}\mathbf{Z} - \operatorname{tr}(\mathbf{A}\mathbf{K}))(\mathbf{Z}^\top\mathbf{B}\mathbf{Z} + 2\boldsymbol{\mu}^\top\mathbf{B}\mathbf{Z} - \operatorname{tr}(\mathbf{B}\mathbf{K}))\big)\\ &= \mathbb{E}\,\big(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\big) + 2\boldsymbol{\mu}^\top\mathbf{A}\mathbb{E}\,\big(\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\big) + 2\boldsymbol{\mu}^\top\mathbf{B}\mathbb{E}\,\big(\mathbf{Z}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\big)\\ &\quad -\mathbb{E}\,\big(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\big)\operatorname{tr}(\mathbf{B}\mathbf{K}) - \mathbb{E}\,\big(\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\big)\operatorname{tr}(\mathbf{A}\mathbf{K})\\ &\quad +4\boldsymbol{\mu}^\top\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu} + \operatorname{tr}(\mathbf{A}\mathbf{K})\operatorname{tr}(\mathbf{B}\mathbf{K})\\ &= \mathbb{E}\,\big(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\big) + 2\boldsymbol{\mu}^\top\mathbf{A}\mathbb{E}\,\big(\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\big) + 2\boldsymbol{\mu}^\top\mathbf{B}\mathbb{E}\,\big(\mathbf{Z}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\big)\\ &\quad +4\boldsymbol{\mu}^\top\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu} - \operatorname{tr}(\mathbf{A}\mathbf{K})\operatorname{tr}(\mathbf{B}\mathbf{K})\,,\end{aligned}$$

  where the last equality is a result of Theorem 1.5 because $Z \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$, which implies $\mathbb{E}\,\big(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\big) = \operatorname{tr}(\mathbf{A}\mathbf{K})$.

- Since the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{K}$ are symmetric, it follows from Lemma 1.11 that

$$\begin{aligned}\mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right) &= \mathbb{E}\left(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\cdot\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\ell=1}^{n}\sum_{m=1}^{n} a_{ij}b_{\ell m}\mathbb{E}\left(Z_iZ_jZ_\ell Z_m\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\ell=1}^{n}\sum_{m=1}^{n}\left(a_{ij}k_{ji}b_{\ell m}k_{m\ell}+a_{ji}k_{i\ell}b_{\ell m}k_{mj}+a_{ij}k_{j\ell}b_{\ell m}k_{mi}\right)\\ &= \operatorname{tr}(\mathbf{A}\mathbf{K})\operatorname{tr}(\mathbf{B}\mathbf{K})+2\operatorname{tr}(\mathbf{A}\mathbf{K}\mathbf{B}\mathbf{K})\,.\end{aligned}$$

- Furthermore, Lemma 1.11 implies that

$$\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\right)=\left(\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\mathbb{E}\left(Z_iZ_jZ_\ell\right)\right)_\ell=\mathbf{o} \tag{29}$$

  and analogously $\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\mathbf{B}\mathbf{Z}\right)=\mathbf{o}$.

- This result and the above derived expression for $\operatorname{Cov}\left(\mathbf{Y}^\top\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right)$ imply the statement. □

Now, we derive the following formula for the covariance vector of linear or quadratic forms of normally distributed random vectors.

Theorem 1.7 *Let $\mathbf{Y}=(Y_1,\ldots,Y_n)^\top$ be an n–dimensional random vector with $\mathbf{Y}\sim\mathrm{N}(\boldsymbol{\mu},\mathbf{K})$ and let $\mathbf{A}=(a_{ij})$, $\mathbf{B}=(b_{ij})$ be arbitrary symmetric $n\times n$ matrices. Then it holds*

$$\operatorname{Cov}\left(\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right)=2\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}\,. \tag{30}$$

Proof

- As $\mathbb{E}\left(\mathbf{A}\mathbf{Y}\right)=\mathbf{A}\boldsymbol{\mu}$ and as it has been shown in Theorem 1.5 that

$$\mathbb{E}\left(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right)=\operatorname{tr}(\mathbf{B}\mathbf{K})+\boldsymbol{\mu}^\top\mathbf{B}\boldsymbol{\mu}\,,$$

  it follows that

$$\begin{aligned}\operatorname{Cov}\left(\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right) &= \mathbb{E}\left((\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}^\top\mathbf{B}\mathbf{Y}-\boldsymbol{\mu}^\top\mathbf{B}\boldsymbol{\mu}-\operatorname{tr}(\mathbf{B}\mathbf{K}))\right)\\ &= \mathbb{E}\left((\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu})((\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu})+2(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}\boldsymbol{\mu}-\operatorname{tr}(\mathbf{B}\mathbf{K}))\right).\end{aligned}$$

- Moreover, it holds that $\mathbb{E}\left(\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu}\right)=\mathbf{o}$ and from (29) it follows with $\mathbf{Z}=\mathbf{Y}-\boldsymbol{\mu}$ that

$$\mathbb{E}\left((\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu})\right)=\mathbf{A}\mathbb{E}\left((\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu})\right)=\mathbf{o}\,.$$

- Therefore, we get

$$\begin{aligned}\operatorname{Cov}\left(\mathbf{A}\mathbf{Y},\,\mathbf{Y}^\top\mathbf{B}\mathbf{Y}\right) &= 2\mathbb{E}\left((\mathbf{A}\mathbf{Y}-\mathbf{A}\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\mathbf{B}\boldsymbol{\mu}\right)\\ &= 2\mathbf{A}\mathbb{E}\left((\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})^\top\right)\mathbf{B}\boldsymbol{\mu}\\ &= 2\mathbf{A}\mathbf{K}\mathbf{B}\boldsymbol{\mu}\,.\end{aligned}$$

□

1.3.2 Noncentral χ^2–Distribution

To determine the distribution of quadratic forms of normally distributed random vectors we introduce the (parametric) family of the noncentral χ^2–distribution.

Definition Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $(X_1, \ldots, X_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$. Then the random variable

$$Z = (X_1, \ldots, X_n)(X_1, \ldots, X_n)^\top = \sum_{i=1}^{n} X_i^2$$

is distributed according to a *noncentral χ^2–distribution* with n degrees of freedom and the *noncentrality parameter* $\lambda = \boldsymbol{\mu}^\top \boldsymbol{\mu}$. (Notation: $Z \sim \chi^2_{n,\lambda}$)

Remark

- For $\boldsymbol{\mu} = \mathbf{o}$ we obtain the (central) χ^2–distribution χ^2_n with n degrees of freedom, which has already been introduced in Section I–1.3.1, as a special case.
- To derive a formula for the density of the noncentral χ^2–distribution we consider (in addition to the characteristic function) still another *integral transform* of probability densities.

Definition

- Let $f : \mathbb{R} \to [0, \infty)$ be the density of a real–valued random variable, such that the integral $\int_{-\infty}^{\infty} e^{tx} f(x)\, dx$ is well–defined for each $t \in (a, b)$ in a certain interval (a, b) with $a < b$.
- Then the mapping $\psi : (a, b) \to \mathbb{R}$ with

$$\psi(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx\,, \qquad \forall\, t \in (a, b) \tag{31}$$

is called the *moment generating function* of the density f.

The following *uniqueness theorem* for moment generating functions is true, which we state without proof.

Lemma 1.12

- *Let $f, f' : \mathbb{R} \to [0, \infty)$ be densities of real–valued random variables and let the corresponding moment generating functions $\psi : (a, b) \to \mathbb{R}$ and $\psi' : (a, b) \to \mathbb{R}$ be well–defined in a (common) interval (a, b) with $a < b$.*
- *It holds that $\psi(t) = \psi'(t)$ for each $t \in (a, b)$ if and only if $f(x) = f'(x)$ for almost all $x \in \mathbb{R}$.*

By using Lemma 1.12 we are now able to identify the density of the noncentral χ^2–distribution.

Theorem 1.8

- *Let the random variable $Z_{n,\lambda} : \Omega \to \mathbb{R}$ be distributed according to the $\chi^2_{n,\lambda}$–distribution with n degrees of freedom and noncentrality parameter λ.*
- *Then the density of $Z_{n,\lambda}$ is given by*

$$f_{Z_{n,\lambda}}(z) = \begin{cases} \exp\Big(-\dfrac{\lambda + z}{2}\Big) \displaystyle\sum_{j=0}^{\infty} \frac{\Big(\dfrac{\lambda}{2}\Big)^j z^{\frac{n}{2}+j-1}}{j!\, 2^{\frac{n}{2}+j}\, \Gamma\Big(\dfrac{n}{2}+j\Big)}\,, & \text{if } z > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

Proof

- Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $(X_1, \ldots, X_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$.
- The moment generating function $\psi_Z(t)$ of $Z = (X_1, \ldots, X_n)(X_1, \ldots, X_n)^\top = \sum_{j=1}^n X_j^2$ is well–defined for $t \in (-\infty, 1/2)$ and for each $t < 1/2$ it holds that

$$\begin{aligned}
\psi_Z(t) &= \mathbb{E}\exp\Big(t\sum_{j=1}^n X_j^2\Big) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\Big(t\sum_{j=1}^n x_j^2\Big)\prod_{j=1}^n \frac{1}{\sqrt{2\pi}}\exp\Big(\frac{1}{2}\,(x_j-\mu_j)^2\Big)\,dx_1\ldots dx_n \\
&= \left(\frac{1}{2\pi}\right)^{n/2}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\exp\left(t\sum_{j=1}^n x_j^2 - \frac{1}{2}\sum_{j=1}^n (x_j-\mu_j)^2\right)dx_1\ldots dx_n \\
&= \prod_{j=1}^n\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp\left(tx_j^2 - \frac{1}{2}(x_j-\mu_j)^2\right)dx_j\,.
\end{aligned}$$

- It is possible to rewrite the exponent of the last term as follows:

$$\begin{aligned}
tx_j^2 - \frac{1}{2}(x_j-\mu_j)^2 &= -\frac{1}{2}\,(-2tx_j^2 + x_j^2 - 2x_j\mu_j + \mu_j^2) \\
&= -\frac{1}{2}\left(x_j^2(1-2t) - 2x_j\mu_j + \mu_j^2(1-2t)^{-1} + \mu_j^2 - \mu_j^2(1-2t)^{-1}\right) \\
&= -\frac{1}{2}\left((x_j - \mu_j(1-2t)^{-1})^2(1-2t) + \mu_j^2(1-(1-2t)^{-1})\right).
\end{aligned}$$

- Hence, it holds that

$$\begin{aligned}
\psi_Z(t) &= \exp\left(-\frac{1}{2}\,(1-(1-2t)^{-1})\sum_{j=1}^n\mu_j^2\right)\prod_{j=1}^n\int_{-\infty}^{\infty}(2\pi)^{-1/2}\exp\Big(-\frac{(x_j-\mu_j(1-2t)^{-1})^2}{2(1-2t)^{-1}}\Big)\,dx_j \\
&= (1-2t)^{-n/2}\exp\Big(-\frac{\lambda}{2}\,(1-(1-2t)^{-1})\Big)
\end{aligned}$$

  as the integrand represents the density of the one–dimensional normal distribution (except for the constant factor $(1-2t)^{1/2}$); $\lambda = \boldsymbol{\mu}^\top\boldsymbol{\mu}$.
- On the other hand, the moment generating function $\psi(t)$ of the density $f_{Z_{n,\lambda}}(z)$ given in (32) can be written as

$$\psi(t) = \sum_{j=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^j}{j!}\int_0^{\infty}e^{tz}\,\frac{z^{n/2+j-1}e^{-z/2}}{2^{\frac{n}{2}+j}\,\Gamma\Big(\frac{n}{2}+j\Big)}\,dz\,,$$

  where the integral is the moment generating function of the (central) χ^2–distribution χ^2_{n+2j} with $n+2j$ degrees of freedom.
- Similar to the way the characteristic function (cf. Theorem I–1.5) is defined, the moment generating function of this distribution is given by

$$\psi_{\chi^2_{n+2j}}(t) = \frac{1}{(1-2t)^{n/2+j}}\,.$$

- Therefore, it holds that

$$\int_0^{\infty}e^{tz}\,\frac{z^{n/2+j-1}e^{-z/2}}{2^{\frac{n}{2}+j}\,\Gamma\Big(\frac{n}{2}+j\Big)}\,dz = \frac{1}{(1-2t)^{n/2+j}}\,,$$

and

$$\begin{aligned}\psi(t) &= e^{-\lambda/2}(1-2t)^{-n/2}\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{\lambda}{2}(1-2t)^{-1}\right)^j \\ &= (1-2t)^{-n/2}\exp\Big(-\frac{\lambda}{2}\left(1-(1-2t)^{-1}\right)\Big)\,.\end{aligned}$$

- Hence, $\psi(t) = \psi_Z(t)$ for each $t < 1/2$ and the statement follows from Lemma 1.12. □

1.3.3 Distributional Properties of Linear and Quadratic Forms

- *Recall*: The definition of the noncentral χ^2–distribution in Section 1.3.2 considers the sum of squares of the components of $\mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$–distributed random vectors.
- One can show that the (adequately modified) sum of squares is distributed according to the noncentral χ^2–distribution even if the considered normally distributed random vector has an *arbitrary* positive definite covariance matrix.
- Indeed, let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $\mathbf{K}$ be a symmetric and positive definite $n \times n$ matrix.
- If $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, Theorem 1.3 implies that

$$\mathbf{K}^{-1/2}\mathbf{Z} \sim \mathrm{N}(\mathbf{K}^{-1/2}\boldsymbol{\mu}, \mathbf{I})\,.$$

- Therefore, by the definition of the noncentral χ^2–distribution it follows that

$$\mathbf{Z}^\top\mathbf{K}^{-1}\mathbf{Z} = \left(\mathbf{K}^{-1/2}\mathbf{Z}\right)^\top\mathbf{K}^{-1/2}\mathbf{Z} \sim \chi^2_{n,\lambda}\,, \tag{33}$$

  where $\lambda = (\mathbf{K}^{-1/2}\boldsymbol{\mu})^\top\mathbf{K}^{-1/2}\boldsymbol{\mu} = \boldsymbol{\mu}^\top\mathbf{K}^{-1}\boldsymbol{\mu}$.

The distributional property (33) for quadratic forms of normally distributed random vectors has the following generalization. In this context Lemma 1.7 about the factorization of symmetric and positive semidefinite matrices is useful.

Theorem 1.9

- *Let* $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$*, where the covariance matrix* $\mathbf{K}$ *be positive definite. Moreover, let* $\mathbf{A}$ *be a symmetric* $n \times n$ *matrix with* $\mathrm{rk}(\mathbf{A}) = r \leq n$*.*
- *If the matrix* $\mathbf{AK}$ *is idempotent, i.e., if* $\mathbf{AK} = (\mathbf{AK})^2$*, it holds that* $\mathbf{Z}^\top\mathbf{AZ} \sim \chi^2_{r,\lambda}$*, where* $\lambda = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}$*.*

Proof

- Let the matrix $\mathbf{AK}$ be idempotent. Then it holds that

$$\mathbf{AK} = \mathbf{AKAK}\,.$$

- Since $\mathbf{K}$ is nondegenerate, it is allowed to multiply both sides of the above equation from the right by $\mathbf{K}^{-1}$. It follows

$$\mathbf{A} = \mathbf{AKA} \tag{34}$$

  or

$$\mathbf{x}^\top\mathbf{Ax} = \mathbf{x}^\top\mathbf{AKAx} = (\mathbf{Ax})^\top\mathbf{K}(\mathbf{Ax}) \geq 0$$

  for each $\mathbf{x} \in \mathbb{R}^n$, i.e., $\mathbf{A}$ is positive semidefinite.

- According to Lemma 1.7 there exists a decomposition

$$\mathbf{A} = \mathbf{H}\mathbf{H}^\top , \tag{35}$$

such that the $n \times r$ matrix $\mathbf{H}$ has full column rank r.

- Now, Lemma 1.2 implies that the inverse matrix $(\mathbf{H}^\top\mathbf{H})^{-1}$ exists.
- From Theorem 1.3 about the linear transformation of normally distributed random vectors it follows for the r–dimensional vector $\mathbf{Z}' = \mathbf{H}^\top\mathbf{Z}$ that

$$\mathbf{Z}' \sim \mathrm{N}(\mathbf{H}^\top\boldsymbol{\mu}, \mathbf{I}_r) \tag{36}$$

because

$$\begin{aligned} \mathbf{H}^\top\mathbf{K}\mathbf{H} &= (\mathbf{H}^\top\mathbf{H})^{-1}(\mathbf{H}^\top\mathbf{H})(\mathbf{H}^\top\mathbf{K}\mathbf{H})(\mathbf{H}^\top\mathbf{H})(\mathbf{H}^\top\mathbf{H})^{-1} \\ &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top(\mathbf{A}\mathbf{K}\mathbf{A})\mathbf{H}(\mathbf{H}^\top\mathbf{H})^{-1} \\ &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{A}\mathbf{H}(\mathbf{H}^\top\mathbf{H})^{-1} = \mathbf{I}_r , \end{aligned}$$

where the last three equalities follow from (34) and (35).

- As on the other hand

$$\mathbf{Z}^\top\mathbf{A}\mathbf{Z} \;=\; \mathbf{Z}^\top\mathbf{H}\mathbf{H}^\top\mathbf{Z} \;=\; \big(\mathbf{H}^\top\mathbf{Z}\big)^\top\mathbf{H}^\top\mathbf{Z} \;=\; (\mathbf{Z}')^\top\mathbf{Z}'$$

and since

$$\big(\mathbf{H}^\top\boldsymbol{\mu}\big)^\top\mathbf{H}^\top\boldsymbol{\mu} = \boldsymbol{\mu}^\top\mathbf{H}\mathbf{H}^\top\boldsymbol{\mu} = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu} ,$$

the statement follows from (36) and from the definition of the noncentral χ^2–distribution. □

Furthermore, the following criterion for the independence of linear and quadratic forms of normally distributed random vectors is useful. It can be considered as a (vectorial) generalization of Lemma I–5.3.

Theorem 1.10

- *Let* $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$*, where* $\mathbf{K}$ *is an arbitrary (symmetric and positive semidefinite) covariance matrix.*
- *Moreover, let* $\mathbf{A}$*,* $\mathbf{B}$ *be arbitrary* $r_1 \times n$ *and* $r_2 \times n$ *matrices with* $r_1, r_2 \leq n$ *and let* $\mathbf{C}$ *be a symmetric and positive semidefinite* $n \times n$ *matrix.*
- *If the additional condition*

$$\mathbf{A}\mathbf{K}\mathbf{B}^\top = \mathbf{0} \qquad \textit{or} \qquad \mathbf{A}\mathbf{K}\mathbf{C} = \mathbf{0} \tag{37}$$

is fulfilled, the random variables $\mathbf{A}\mathbf{Z}$ *and* $\mathbf{B}\mathbf{Z}$ *or* $\mathbf{A}\mathbf{Z}$ *and* $\mathbf{Z}^\top\mathbf{C}\mathbf{Z}$*, respectively, are independent.*

Proof

- First, we show that (37) implies the independence of the linear forms $\mathbf{A}\mathbf{Z}$ and $\mathbf{B}\mathbf{Z}$.
- Because of the uniqueness theorem for characteristic functions of random vectors (cf. Lemma 1.9), it suffices to show that $\mathbf{t}_2 \in \mathbb{R}^{r_2}$

$$\mathbb{E}\,\exp\big(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A}\mathbf{Z} + \mathbf{t}_2^\top\mathbf{B}\mathbf{Z})\big) = \mathbb{E}\,\exp\big(\mathrm{i}\,\mathbf{t}_1^\top\mathbf{A}\mathbf{Z}\big)\mathbb{E}\,\exp\big(\mathrm{i}\,\mathbf{t}_2^\top\mathbf{B}\mathbf{Z}\big)$$

for arbitrary $\mathbf{t}_1 \in \mathbb{R}^{r_1}$.

- From (37) it follows that

$$\mathbf{BKA}^\top = \Big((\mathbf{BKA}^\top)^\top\Big)^\top = \Big(\mathbf{AKB}^\top\Big)^\top = \mathbf{0}.$$

- Therefore, it holds for arbitrary $\mathbf{t}_1 \in \mathbb{R}^{r_1}$, $\mathbf{t}_2 \in \mathbb{R}^{r_2}$ that

$$(\mathbf{t}_1^\top\mathbf{A})\mathbf{K}(\mathbf{t}_2^\top\mathbf{B})^\top = \mathbf{t}_1^\top\mathbf{AKB}^\top\mathbf{t}_2 = \mathbf{0}\,, \qquad (\mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A})^\top = \mathbf{t}_2^\top\mathbf{BKA}^\top\mathbf{t}_1 = \mathbf{0}\,. \tag{38}$$

- Then the representation formula (25) for the characteristic function of normally distributed random vectors derived in Theorem 1.4 and (38) imply that

$$\begin{aligned}
&\mathbb{E}\,\exp\big(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{AZ} + \mathbf{t}_2^\top\mathbf{BZ})\big) = \mathbb{E}\,\exp\big(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\mathbf{Z}\big)\\
&= \exp\Big(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})^\top\Big)\\
&= \exp\Big(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A} + \mathbf{t}_2^\top\mathbf{B})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_1^\top\mathbf{A})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A})^\top - \frac{1}{2}\,(\mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_2^\top\mathbf{B})^\top\Big)\\
&= \exp\Big(\mathrm{i}\,(\mathbf{t}_1^\top\mathbf{A})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_1^\top\mathbf{A})\mathbf{K}(\mathbf{t}_1^\top\mathbf{A})^\top\Big)\exp\Big(\mathrm{i}\,(\mathbf{t}_2^\top\mathbf{B})\boldsymbol{\mu} - \frac{1}{2}\,(\mathbf{t}_2^\top\mathbf{B})\mathbf{K}(\mathbf{t}_2^\top\mathbf{B})^\top\Big)\\
&= \mathbb{E}\,\exp\big(\mathrm{i}\,\mathbf{t}_1^\top\mathbf{AZ}\big)\mathbb{E}\,\exp\big(\mathrm{i}\,\mathbf{t}_2^\top\mathbf{BZ}\big)\,.
\end{aligned}$$

- Now, it remains to show that the independence of $\mathbf{AZ}$ of $\mathbf{Z}^\top\mathbf{CZ}$ is a result of the second condition of (37).
- Let $\mathrm{rk}(\mathbf{C}) = r \le n$. According to Lemma 1.7 there is an $n \times r$ matrix $\mathbf{H}$ with $\mathrm{rk}(\mathbf{H}) = r$, such that $\mathbf{C} = \mathbf{HH}^\top$.
- Then it follows from (37) that $\mathbf{AKHH}^\top = \mathbf{0}$ and $\mathbf{AKHH}^\top\mathbf{H} = \mathbf{0}$.
- Because of Lemma 1.2, the $r \times r$ matrix $\mathbf{H}^\top\mathbf{H}$ has (full) rank $\mathrm{rk}(\mathbf{H}) = r$. Hence, $\mathbf{H}^\top\mathbf{H}$ is invertible.
- Finally, it follows that $\mathbf{AKH} = \mathbf{0}$.
- Therefore, the first part of the proof implies that the linear forms $\mathbf{AZ}$ and $\mathbf{H}^\top\mathbf{Z}$ are independent.
- Because of

$$\mathbf{Z}^\top\mathbf{CZ} = \mathbf{Z}^\top\mathbf{HH}^\top\mathbf{Z} = (\mathbf{H}^\top\mathbf{Z})^\top\mathbf{H}^\top\mathbf{Z}\,,$$

the transformation theorem for independent random vectors (cf. Theorem I–1.8) implies that also $\mathbf{AZ}$ and $\mathbf{Z}^\top\mathbf{CZ}$ are independent. □

2 Linear Models; Design Matrix with Full Rank

Recall (cf. Chapter 5 of the lecture "Stochastik I"):

- In *simple linear regression* one considers two datasets $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and $(y_1, \ldots, y_n) \in \mathbb{R}^n$, which shall be modeled stochastically.
- In doing so, we perceive the vectors $(x_1, y_1), \ldots, (x_n, y_n)$ as realizations of n random vectors, say $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are typically *not* identically distributed.
- We interpret the random variables $Y_1, \ldots, Y_n$ as *response variables* and assume that they depend on the *predictor variables* $X_1, \ldots, X_n$ in the following way:

$$Y_i = \varphi(X_i) + \varepsilon_i\,, \qquad \forall\, i = 1, \ldots, n\,, \tag{1}$$

  where
  - $\varphi : \mathbb{R} \to \mathbb{R}$ is an arbitrary Borel measurable function, the so-called *regression function*, and
  - $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \mathbb{R}$ are random variables, so-called *error terms*, which can be used to model random errors, e.g., errors in measurement.
- We are dealing with an important special case if the regression function $\varphi : \mathbb{R} \to \mathbb{R}$ is a linear function, the so-called *regression line*, i.e., if there are real numbers $\beta_1, \beta_2 \in \mathbb{R}$, such that

$$\varphi(x) = \beta_1 + \beta_2 x\,, \qquad \forall\, x \in \mathbb{R}\,, \tag{2}$$

  where β_1 is called the *intercept* and β_2 is called the *regression coefficient*.
- The quantities $\beta_1, \beta_2 \in \mathbb{R}$ are unknown model parameters, which are to be estimated from the observed data $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and $(y_1, \ldots, y_n) \in \mathbb{R}^n$.

We now consider the following *multivariate* generalization of the simple linear regression model and let $m, n \geq 2$ be arbitrary natural numbers, such that $m \leq n$.

- We assume that the response variables $Y_1, \ldots, Y_n$ depend on *vectorial* m–dimensional predictor variables $(X_{11}, \ldots, X_{1m})^\top, \ldots, (X_{n1}, \ldots, X_{nm})^\top$, i.e.,

$$Y_i = \varphi(X_{i1}, \ldots, X_{im}) + \varepsilon_i\,, \qquad \forall\, i = 1, \ldots, n\,, \tag{3}$$

  where
  - the regression function $\varphi : \mathbb{R}^m \to \mathbb{R}$ is given by

$$\varphi(x_1, \ldots, x_m) = \beta_1 x_1 + \ldots + \beta_m x_m\,, \qquad \forall\, (x_1, \ldots, x_m)^\top \in \mathbb{R}^m \tag{4}$$

    with (unknown) regression coefficients $\beta_1, \ldots, \beta_m \in \mathbb{R}$ and
  - the random error terms $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \mathbb{R}$ satisfy the following requirements:

$$\mathbb{E}\,\varepsilon_i = 0\,, \qquad \operatorname{Var}\varepsilon_i = \sigma^2\,, \qquad \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0\,, \qquad \forall\, i, j = 1, \ldots, n \text{ with } i \neq j \tag{5}$$

    for a certain (unknown) $\sigma^2 > 0$.
- Here we just consider the case of deterministic predictor variables $(X_{11}, \ldots, X_{1m})^\top, \ldots, (X_{n1}, \ldots, X_{nm})^\top$, i.e., we put

$$(X_{11}, \ldots, X_{1m})^\top = (x_{11}, \ldots, x_{1m})^\top, \ldots, (X_{n1}, \ldots, X_{nm})^\top = (x_{n1}, \ldots, x_{nm})^\top$$

  for certain vectors $(x_{11}, \ldots, x_{1m})^\top, \ldots, (x_{n1}, \ldots, x_{nm})^\top \in \mathbb{R}^m$.

Remark

- In matrix notation the model given in (3) and (4) can be expressed as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{6}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{7}$$

- Here $\mathbf{X}$ is called the *design matrix* of the regression model.

2.1 Method of Least Squares

The goal of this section consists in estimating the unknown model parameters $\beta_1, \dots, \beta_m$ and σ^2 from the observed data $(x_{11}, \dots, x_{1m})^\top, \dots, (x_{n1}, \dots, x_{nm})^\top \in \mathbb{R}^m$ and $(y_1 \dots, y_n)^\top \in \mathbb{R}^n$.

- Similar to the way this is done in Section I–5.1 we consider the *method of least squares* in order to determine estimators $\widehat{\beta}_1, \dots, \widehat{\beta}_m$ for the unknown regression coefficients $\beta_1, \dots, \beta_m$.
- In detail, this means that a random vector $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \dots, \widehat{\beta}_m)^\top$ is to be determined, such that the *mean squared error*

$$e(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - (\beta_1 x_{i1} + \dots + \beta_m x_{im})\big)^2 \tag{8}$$

is minimal for $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

Remark Except the model assumptions made in (5) no further preconditions concerning the distribution of the random error terms $\varepsilon_1, \dots, \varepsilon_n : \Omega \to \mathbb{R}$ are required up to now.

2.1.1 Normal Equation

It can easily be shown that the function $e(\boldsymbol{\beta})$ considered in (8) has a uniquely determined minimum if the design matrix $\mathbf{X}$ has full (column) rank, i.e., $\operatorname{rk}(\mathbf{X}) = m$.

Theorem 2.1 *Let* $\operatorname{rk}(\mathbf{X}) = m$.

- *The mean squared error* $e(\boldsymbol{\beta})$ *in* (8) *is minimal if and only if* $\boldsymbol{\beta}$ *is a solution of the so–called normal equation:*

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}\,. \tag{9}$$

- *The normal equation* (9) *has the uniquely determined solution*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\,. \tag{10}$$

Proof

- The function $e(\boldsymbol{\beta})$ given in (8) is differentiable, where

$$e'(\boldsymbol{\beta}) = \Big(\frac{\partial e(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial e(\boldsymbol{\beta})}{\partial \beta_m}\Big)^\top = \frac{2}{n}\left(\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top\mathbf{Y}\right)$$

  and

$$e''(\boldsymbol{\beta}) = \Big(\frac{\partial^2 e(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\Big) = \frac{2}{n}\,\mathbf{X}^\top\mathbf{X}\,.$$

- Setting $e'(\boldsymbol{\beta}) = \mathbf{o}$ results in the normal equation (9).
- Moreover, it follows from Lemma 1.2 that $\operatorname{rk}(\mathbf{X}^\top\mathbf{X}) = m$.
  - Hence, the $m \times m$ matrix $\mathbf{X}^\top\mathbf{X}$ (and consequently also $e''(\boldsymbol{\beta})$) is invertible and positive definite.
  - Therefore, $e(\boldsymbol{\beta})$ is minimal if and only if $\boldsymbol{\beta}$ is a solution of (9).
- As the $m \times m$ matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, we know that (9) has a uniquely determined solution $\widehat{\boldsymbol{\beta}}$, which is given by (10). □

Remark The estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ for $\boldsymbol{\beta}$ is a linear transformation of the random sample $\mathbf{Y}$, i.e., $\widehat{\boldsymbol{\beta}}$ is a *linear estimator* for $\boldsymbol{\beta}$.

Examples (simple and multiple linear regression model)

- For $m = 2$ and

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{11}$$

  we obtain the *simple linear regression model* already considered in Section I–5.1 as a special case.
- The design matrix $\mathbf{X}$ in (11) has full rank $\operatorname{rk}(\mathbf{X}) = 2$ if and only if not all $x_1, \ldots, x_n$ are equal.
- The estimators $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2)$ for the intercept β_1 and the regression coefficient β_2, considered in (10) (see also Theorem I–5.1), are then given by

$$\widehat{\beta}_2 = \frac{s_{xy}^2}{s_{xx}^2} \qquad \text{and} \qquad \widehat{\beta}_1 = \overline{y}_n - \widehat{\beta}_2\overline{x}_n\,, \tag{12}$$

  respectively, where $\overline{x}_n$, $\overline{y}_n$ denote the sample means, i.e.,

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i\,, \qquad \overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i\,,$$

  and the sample variances s_{xx}^2, s_{yy}^2 and sample covariance s_{xy}^2 are given by

$$s_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^n \big(x_i - \overline{x}_n\big)^2\,, \quad s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad \text{and} \quad s_{yy}^2 = \frac{1}{n-1}\sum_{i=1}^n \big(y_i - \overline{y}_n\big)^2\,.$$

- For $m > 2$ and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \ldots & x_{nm} \end{pmatrix} \tag{13}$$

  we obtain the so-called *multiple linear regression model*.

2.1.2 Properties of the LS–Estimator $\widehat{\boldsymbol{\beta}}$

From now on in Section 2.1, we always assume that the design matrix $\mathbf{X}$ has full (column) rank and derive three different *properties* of the LS–estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)^\top$ given in (10).

Theorem 2.2 *The estimator $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, i.e., $\mathbb{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^m$.*

Proof

Due to $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{o}$ it follows from (6) and (10) that

$$\begin{aligned} \mathbb{E}\,\widehat{\boldsymbol{\beta}} &\overset{(10)}{=} \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right) \overset{(6)}{=} \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right) \;=\; \boldsymbol{\beta} + \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\,\boldsymbol{\varepsilon} \;=\; \boldsymbol{\beta}\,. \end{aligned}$$

□

The LS–estimator $\widehat{\boldsymbol{\beta}}$ additionally has the following *minimum variance* property. We denote by $\mathcal{L}$ the family of all unbiased linear estimators $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y} + \mathbf{a}$ for $\boldsymbol{\beta}$, where $\mathbf{A}$ is an $(m \times n)$–dimensional matrix and $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$.

Theorem 2.3 *For all $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_m) \in \mathcal{L}$ it holds that*

$$\operatorname{Var}\widehat{\beta}_i \le \operatorname{Var}\widetilde{\beta}_i\,, \qquad \forall\, i = 1, \ldots, m\,, \tag{14}$$

where the equality in (14) is true for all $i = 1, \ldots, m$ if and only if $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$.

Proof

- As it is assumed that the linear estimator $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y} + \mathbf{a}$ is unbiased for $\boldsymbol{\beta}$, one has

  $$\boldsymbol{\beta} \;=\; \mathbb{E}\,\widetilde{\boldsymbol{\beta}} \;=\; \mathbb{E}\,(\mathbf{A}\mathbf{Y}+\mathbf{a}) \overset{(6)}{=} \mathbb{E}\left(\mathbf{A}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right) + \mathbf{a} \;=\; \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbb{E}\,\boldsymbol{\varepsilon} + \mathbf{a} \;=\; \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{a}$$

  for all $\boldsymbol{\beta} \in \mathbb{R}^m$, where the last equality arises from $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{o}$.

- Herefrom, it follows that

  $$\mathbf{A}\mathbf{X} = \mathbf{I} \qquad \text{and} \qquad \mathbf{a} = \mathbf{o}\,. \tag{15}$$

- Hence, one has

  $$\widetilde{\boldsymbol{\beta}} \;=\; \mathbf{A}\mathbf{Y} \;=\; \mathbf{A}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \;=\; \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} \;=\; \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}\,,$$

  i.e., each linear unbiased estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is of the form

  $$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}\,. \tag{16}$$

- For the covariance matrix $\operatorname{Cov}(\widetilde{\boldsymbol{\beta}})$ of the random vector $\widetilde{\boldsymbol{\beta}}$ it thus holds that

  $$\operatorname{Cov}(\widetilde{\boldsymbol{\beta}}) \;=\; \mathbb{E}\left((\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right) \;=\; \mathbb{E}\left((\mathbf{A}\boldsymbol{\varepsilon})(\mathbf{A}\boldsymbol{\varepsilon})^\top\right) \;=\; \mathbf{A}\mathbb{E}\,(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\mathbf{A}^\top \;=\; \sigma^2\mathbf{A}\mathbf{A}^\top\,,$$

  i.e.,

  $$\operatorname{Cov}(\widetilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}\mathbf{A}^\top\,. \tag{17}$$

- Furthermore, it results from (17) with $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ that the covariance matrix $\mathrm{Cov}\,(\widehat{\boldsymbol{\beta}})$ of the LS–estimator $\widehat{\boldsymbol{\beta}}$ is given by
$$\mathrm{Cov}\,(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \tag{18}$$
because
$$\begin{aligned}\mathrm{Cov}\,(\widehat{\boldsymbol{\beta}}) &= \sigma^2\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)^\top \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\,.\end{aligned}$$
- So in order to prove the validity of (14), it has to be shown that
$$\big((\mathbf{X}^\top\mathbf{X})^{-1}\big)_{ii} \le \big(\mathbf{A}\mathbf{A}^\top\big)_{ii}\,, \qquad \forall\, i = 1,\ldots,m\,. \tag{19}$$
- For $\mathbf{D} = \mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ one has
$$\begin{aligned}\mathbf{A}\mathbf{A}^\top &= \big(\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\big(\mathbf{D} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)^\top \\ &= \mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}^\top + \mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}\end{aligned}$$
because due to (15) it holds that
$$\mathbf{D}\mathbf{X} = \big(\mathbf{A} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\mathbf{X} = \mathbf{A}\mathbf{X} - \mathbf{I} = \mathbf{I} - \mathbf{I} = \mathbf{0}\,,$$
where $\mathbf{0}$ denotes the zero matrix.
- As for $\mathbf{D} = (d_{ij})$ the inequality $\big(\mathbf{D}\mathbf{D}^\top\big)_{ii} = \sum_{j=1}^m d_{ij}^2 \ge 0$ is fulfilled, this gives (19).
- Moreover, it becomes clear that the equality in (19) for each $i = 1,\ldots,m$ holds if and only if $\mathbf{D} = \mathbf{0}$, i.e., $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. □

Remark

- It follows from Theorems 2.1 and 2.2 that $\widehat{\boldsymbol{\beta}} \in \mathcal{L}$. Moreover, it arises from Theorem 2.3 that $\widehat{\boldsymbol{\beta}}$ is the *best unbiased linear estimator* for $\boldsymbol{\beta}$ in terms of (14).
- We now derive a sufficient condition for $\widehat{\boldsymbol{\beta}}$ to be a weakly consistent estimator for $\boldsymbol{\beta}$, where the sample size n, i.e., the number of rows of the design matrix $\mathbf{X} = \mathbf{X}_n$ tends to ∞.
- *Recall*: An estimator $\widetilde{\boldsymbol{\beta}}_n = \widetilde{\boldsymbol{\beta}}(Y_1,\ldots,Y_n)$ for $\boldsymbol{\beta}$ is called *weakly consistent* if
$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}}(|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \varepsilon) = 0\,, \qquad \forall\, \varepsilon > 0,\, \boldsymbol{\beta} \in \mathbb{R}^m\,.$$
- Under similar conditions, it can also be shown that $\widehat{\boldsymbol{\beta}}_n$ is asymptotically normally distributed if $n \to \infty$ (cf. Section III.3.2 in Pruscha (2000)).

Theorem 2.4 *Let $f : \mathbb{N} \to \mathbb{R}\setminus\{0\}$ be a function satisfying $\lim_{n\to\infty} f(n) = 0$, such that the limit*
$$\mathbf{Q} = \lim_{n\to\infty}\big(f(n)\mathbf{X}_n^\top\mathbf{X}_n\big) \tag{20}$$
exists and the $m \times m$ matrix $\mathbf{Q}$ is invertible. Then $\widehat{\boldsymbol{\beta}}_n$ is a weakly consistent estimator for $\boldsymbol{\beta}$.

Proof

- As $\widehat{\boldsymbol{\beta}}_n$ is unbiased (cf. Theorem 2.2), it holds for each $n \geq m$ that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\beta}}(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \varepsilon) &= \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}|^2 > \varepsilon^2) = \mathbb{P}_{\boldsymbol{\beta}}\Big(\sum_{i=1}^{m}\big(\widehat{\beta}_{in} - \beta_i\big)^2 > \varepsilon^2\Big) \\
&\leq \mathbb{P}_{\boldsymbol{\beta}}\Big(\bigcup_{i=1}^{m}\big\{(\widehat{\beta}_{in} - \beta_i)^2 > \frac{\varepsilon^2}{m}\big\}\Big) \leq \sum_{i=1}^{m}\mathbb{P}_{\boldsymbol{\beta}}\Big(\big(\widehat{\beta}_{in} - \beta_i\big)^2 > \frac{\varepsilon^2}{m}\Big) \\
&\leq \frac{m}{\varepsilon^2}\sum_{i=1}^{m}\operatorname{Var}\widehat{\beta}_{in}\,,
\end{aligned}$$

  where the last inequality results from the Chebyshev inequality (cf. Theorem WR-4.18).

- Hence, it suffices to show that

$$\lim_{n\to\infty}\operatorname{Var}\widehat{\beta}_{in} = 0\,, \qquad \forall\, i = 1, \ldots, m\,. \tag{21}$$

- The matrix $\mathbf{Q}^{-1}$ is well–defined because we assume that the (limiting) matrix $\mathbf{Q}$ is invertible. Moreover, it follows from (20) that

$$\mathbf{Q}^{-1} = \lim_{n\to\infty}\big(f(n)\mathbf{X}_n^\top\mathbf{X}_n\big)^{-1}\,.$$

- From the formula for the covariance matrix of the random vector $\widehat{\beta}_n$ derived in (18), it now results that

$$\lim_{n\to\infty}\operatorname{Cov}(\widehat{\beta}_n) = \sigma^2\lim_{n\to\infty}\big(\mathbf{X}_n^\top\mathbf{X}_n\big)^{-1} = \sigma^2\lim_{n\to\infty}f(n)\lim_{n\to\infty}\big(f(n)\mathbf{X}_n^\top\mathbf{X}_n\big)^{-1} = \Big(\sigma^2\lim_{n\to\infty}f(n)\Big)\mathbf{Q}^{-1} = \mathbf{0}\,.$$

- This particularly implies (21). □

2.1.3 Unbiased Estimation of the Variance σ^2 of the Error Terms

- Besides the conditions on the error terms $\varepsilon_1, \ldots, \varepsilon_n$ formulated in (5), we now assume that $n > m$. Furthermore, we again assume that the design matrix has full rank, i.e., $\operatorname{rk}(\mathbf{X}) = m$.

- By generalizing the approach we considered in Section I–5.1.3 for the estimation of σ^2 in the simple linear regression model, we now consider

$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,. \tag{22}$$

- For normally distributed error terms, S^2 can be regarded as a modified version of a maximum–likelihood estimator for σ^2; cf. Section 2.2.

We show that the random variable S^2 defined in (22) gives an unbiased estimator for σ^2. Here, the following lemmata are useful.

Lemma 2.1 *The $n \times n$ matrix*

$$\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \tag{23}$$

is idempotent and symmetric, i.e.,

$$\mathbf{G} = \mathbf{G}^2 \qquad \textit{and} \qquad \mathbf{G} = \mathbf{G}^\top\,. \tag{24}$$

Proof

- The second part of the statement in (24) follows directly from the definition of $\mathbf{G}$ and the computation rules for transposed matrices because

$$\mathbf{G}^\top \;=\; \Big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)^\top \;=\; \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \;=\; \mathbf{G}\,.$$

- Furthermore, it holds that

$$\begin{aligned}\mathbf{G}^2 &= \Big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\Big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\\ &= \mathbf{I} - 2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\\ &= \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{G}\,.\end{aligned}$$

□

Lemma 2.2 *For the $n \times n$ matrix $\mathbf{G}$ given in (23) it holds that $\operatorname{tr}(\mathbf{G}) = n - m$.*

Proof

- One can easily see (cf. Lemmas 1.1 and 1.3) that
  - $\operatorname{tr}(\mathbf{A} - \mathbf{B}) = \operatorname{tr}(\mathbf{A}) - \operatorname{tr}(\mathbf{B})$ for arbitrary $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$,
  - $\operatorname{tr}(\mathbf{CD}) = \operatorname{tr}(\mathbf{DC})$ for arbitrary $n \times m$ matrices $\mathbf{C}$ and arbitrary $m \times n$ matrices $\mathbf{D}$.
- Herefrom and from the definition of $\mathbf{G}$ in (23) it follows that

$$\begin{aligned}\operatorname{tr}(\mathbf{G}) &= \operatorname{tr}\Big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big) \;=\; \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}\Big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\\ &= \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}\Big(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big) \;=\; \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}(\mathbf{I}_m) \;=\; n - m\,,\end{aligned}$$

  where $\mathbf{I}_\ell$ denotes the $(\ell \times \ell)$–dimensional identity matrix.

□

Theorem 2.5 *It holds that $\mathbb{E}\,S^2 = \sigma^2$ for any $\sigma^2 > 0$, i.e., S^2 is an unbiased estimator for σ^2.*

Proof

- It obviously holds that

$$\mathbf{GX} = \big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big)\mathbf{X} = \mathbf{0}\,. \tag{25}$$

- Herefrom, it follows with the aid of (10) and (23) that

$$\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \overset{(10)}{=} \mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \overset{(23)}{=} \mathbf{GY} \;=\; \mathbf{GX}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\varepsilon} \overset{(25)}{=} \mathbf{G}\boldsymbol{\varepsilon}\,.$$

- Hence, for the estimator S^2 introduced in (22) it is true that

$$\begin{aligned}S^2 &= \frac{1}{n-m}\,(\mathbf{G}\boldsymbol{\varepsilon})^\top(\mathbf{G}\boldsymbol{\varepsilon}) = \frac{1}{n-m}\,\boldsymbol{\varepsilon}^\top\mathbf{G}^\top\mathbf{G}\boldsymbol{\varepsilon} = \frac{1}{n-m}\,\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\\ &= \frac{1}{n-m}\,\operatorname{tr}\big(\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\big) = \frac{1}{n-m}\,\operatorname{tr}\big(\mathbf{G}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\,,\end{aligned}$$

  on account of $\mathbf{G}^\top\mathbf{G} = \mathbf{G}^2 = \mathbf{G}$ (cf. Lemma 2.1).
- Due to $\mathbb{E}\,(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2\mathbf{I}_n$ this leads to

$$\mathbb{E}\,S^2 = \frac{1}{n-m}\,\operatorname{tr}\big(\mathbf{G}\mathbb{E}\,(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\big) = \frac{1}{n-m}\,\operatorname{tr}\big(\mathbf{G}\sigma^2\mathbf{I}_n\big) = \frac{\sigma^2}{n-m}\,\operatorname{tr}\big(\mathbf{G}\big) = \sigma^2\,,$$

  where the last equality results from Lemma 2.2.

□

2.2 Normally Distributed Error Terms

- In addition to the model assumptions that were made at the beginning of Chapter 2, we now assume that the random error terms $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \mathbb{R}$ are independent and normally distributed, i.e., $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ for each $i = 1, \ldots, n$.
- Moreover, let $\mathrm{rk}(\mathbf{X}) = m$ and $n > m$.
- According to Theorem 1.3, the distributions of the vector $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ of response variables and of the LS–estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ are given by

$$\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}) \tag{26}$$

and

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\big(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big)\,, \tag{27}$$

respectively.

2.2.1 Maximum–Likelihood Estimation

- A *parametric model* for the distribution of the vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ of the sampling variables $Y_1, \ldots, Y_n$ is given by (26).
- This means that aside from the method of least squares, which was discussed in Section 2.1, we can now also use maximum–likelihood estimation in order to construct estimators for the unknown model parameters $\boldsymbol{\beta}$ and σ^2.
- It follows from (1.13) and (26) that

$$f_{\mathbf{Y}}(\mathbf{y}) = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big) \tag{28}$$

for each $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$.

- Hence, we consider the *likelihood function*

$$L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big) \tag{29}$$

or the *loglikelihood function*

$$\log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\,\frac{n}{2}\,\log(2\pi) - \frac{n}{2}\,\log(\sigma^2) - \frac{1}{2\sigma^2}\,|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2\,. \tag{30}$$

- We want to find estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ for $\boldsymbol{\beta}$, σ^2, such that

$$L(\mathbf{Y}; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \sup_{\boldsymbol{\beta}\in\mathbb{R}^m,\, \sigma^2>0} L(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2) \tag{31}$$

with probability 1 or equivalently such that

$$\log L(\mathbf{Y}; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \sup_{\boldsymbol{\beta}\in\mathbb{R}^m,\, \sigma^2>0} \log L(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2) \tag{32}$$

with probability 1.

Remark The maximization in (31) or (32) can be carried out in two steps: first with respect to $\boldsymbol{\beta}$ and then with respect to σ^2. Due to (30), the first step is identical with the minimization method considered in Section 2.1.1.

Theorem 2.6 *The solutions of the maximization problems* (31) *and* (32) *are uniquely determined and given by*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{33}$$

and

$$\widehat{\sigma}^2 = \frac{1}{n} \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^\top \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right), \tag{34}$$

respectively.

Proof

- For arbitrary but fixed $\mathbf{y} \in \mathbb{R}^n$ and $\sigma^2 > 0$, we first consider the mapping

$$\mathbb{R}^m \ni \boldsymbol{\beta} \mapsto \log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2). \tag{35}$$

- In Theorem 2.1 we have shown that the mapping given in (35) has the uniquely determined global maximum $\widehat{\boldsymbol{\beta}}(\mathbf{y}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, which does not depend on σ^2.
- For each (fixed) $\mathbf{y} \in \mathbb{R}^n$, we now consider the mapping

$$(0, \infty) \ni \sigma^2 \mapsto \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2). \tag{36}$$

- This mapping is continuous and it obviously holds that

$$\lim_{\sigma^2 \to \infty} \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2) = -\infty.$$

- As $n > m$ is assumed, the n–dimensional absolutely continuous random vector $\mathbf{Y}$ only takes values in the m–dimensional subset $\{\mathbf{Xz} : \mathbf{z} \in \mathbb{R}^m\}$ of $\mathbb{R}^n$ with probability 0.
- Therefore, we have that $|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}|^2 > 0$ with probability 1 and

$$\lim_{\sigma^2 \to 0} \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2) = -\infty$$

  for almost every $\mathbf{y} \in \mathbb{R}^n$.
- Thus, for almost every $\mathbf{y} \in \mathbb{R}^n$, the mapping given in (36) has at least one global maximum in $(0, \infty)$.
- For each of these maxima, it holds that

$$\frac{\partial \log L(\mathbf{y}; \widehat{\boldsymbol{\beta}}(\mathbf{y}), \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y})) = 0.$$

- The (uniquely determined) solution of this equation is

$$\widehat{\sigma}^2(\mathbf{y}) = \frac{1}{n} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{y})).$$

□

Remark

- The ML–estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ derived in Theorem 2.6 coincides with the LS–estimator derived in Theorem 2.1.
- In contrast, the ML–estimator $\widehat{\sigma}^2$ for σ^2 differs from the (unbiased) estimator S^2 for σ^2 considered in Section 2.1.3 in a constant proportionality factor because

$$\widehat{\sigma}^2 = \frac{n-m}{n} S^2.$$

2.2.2 Distributional Properties of $\widehat{\boldsymbol{\beta}}$ and S^2

- Apart from the fact that $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\big(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big)$ is normally distributed, which was already mentioned in (27), it is also possible to determine the distribution of the estimator

$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,. \tag{37}$$

  for the variance σ^2 of the error terms.

- For this purpose we use the representation formula

$$\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{G}\boldsymbol{\varepsilon}\,, \tag{38}$$

  which we have shown in the proof of Theorem 2.5, where $\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

From the condition derived in Theorem 1.9, under which a quadratic form of normally distributed random vectors follows a χ^2–distribution, we obtain the following result.

Theorem 2.7 *It holds that*

$$\frac{(n-m)S^2}{\sigma^2} \sim \chi^2_{n-m}\,, \tag{39}$$

i.e., the random variable $(n-m)S^2/\sigma^2$ is distributed according to the (central) χ^2–distribution with $n-m$ degrees of freedom.

Proof

- In Lemma 2.1 we have shown that the matrix $\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is idempotent and symmetric.
- Herefrom and from (38) it follows that

$$\begin{aligned}\frac{(n-m)S^2}{\sigma^2} &= \frac{1}{\sigma^2}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \;=\; \frac{1}{\sigma^2}\,(\mathbf{G}\boldsymbol{\varepsilon})^\top\mathbf{G}\boldsymbol{\varepsilon} \;=\; \frac{1}{\sigma^2}\,\boldsymbol{\varepsilon}^\top\mathbf{G}^\top\mathbf{G}\boldsymbol{\varepsilon} \\ &= \big(\sigma^{-1}\boldsymbol{\varepsilon}\big)^\top\mathbf{G}\big(\sigma^{-1}\boldsymbol{\varepsilon}\big)\,.\end{aligned}$$

- As $\sigma^{-1}\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \mathbf{I})$ and as the matrix $\mathbf{GI} = \mathbf{G}$ is idempotent, it suffices to show that $\mathrm{rk}(\mathbf{G}) = n - m$ due to Theorem 1.9.
- This is a result of Lemma 1.3 and 2.2 because

$$\mathrm{rk}(\mathbf{G}) \overset{\text{Lemma 1.3}}{=} \mathrm{sp}(\mathbf{G}) \overset{\text{Lemma 2.2}}{=} n - m\,.$$

□

Moreover, we use the criterion for the independence of linear and quadratic forms of normally distributed random vectors, which has been derived in Theorem 1.10 in order to show the following result.

Theorem 2.8 *The estimators $\widehat{\boldsymbol{\beta}}$ and S^2 for $\boldsymbol{\beta}$ and σ^2, respectively, are independent.*

Proof

- From $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ it follows that

$$\widehat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \;=\; (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} + \boldsymbol{\beta}\,.$$

- Furthermore, we have shown in the proof of Theorem 2.7 that the estimator

$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

  can be written as a quadratic form of $\boldsymbol{\varepsilon}$:

$$S^2 = \frac{1}{n-m}\,\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\,, \qquad \text{where } \mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

- Due to $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I})$ and

$$\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big)\Big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\Big) = \mathbf{0}\,,$$

  it follows from Theorem 1.10 that the linear form $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$ and the quadratic form $\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}$ are independent.
- This implies that also the random variables $\widehat{\boldsymbol{\beta}}$ and S^2 are independent. □

2.2.3 Statistical Tests for the Regression Coefficients

- By use of the distributional properties of linear and quadratic forms of normally distributed random vectors which have been derived in Sections 1.3.3 and 2.2.2, we are able to construct statistical t–tests and F–tests for the verification of hypotheses about the regression coefficients $\beta_1, \ldots, \beta_m$.
- In doing so, we still consider the (independent) estimators $\widehat{\boldsymbol{\beta}}$ and S^2 for $\boldsymbol{\beta}$ and σ^2, where

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \sim \mathrm{N}\big(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big) \tag{40}$$

  and

$$\frac{(n-m)S^2}{\sigma^2} = \frac{1}{\sigma^2}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \sim \chi^2_{n-m}\,. \tag{41}$$

We first discuss the following F–test, which is also called a *test of model significance*.

- Here, the null hypothesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ is tested (against the alternative $H_1 : \beta_j \neq 0$ for at least one $j \in \{1, \ldots, m\}$).
- The choice of the test statistic is motivated by the following decomposition.

Theorem 2.9 *With the notation* $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$*, it holds that*

$$\mathbf{Y}^\top\mathbf{Y} = \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}} + \big(\mathbf{Y} - \widehat{\mathbf{Y}}\big)^\top\big(\mathbf{Y} - \widehat{\mathbf{Y}}\big)\,. \tag{42}$$

Proof

- We have that

$$\begin{aligned}\mathbf{Y}^\top\mathbf{Y} &= \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \Big((Y_i-\widehat{Y}_i)+\widehat{Y}_i\Big)^2 = \sum_{i=1}^n \widehat{Y}_i^2 + \sum_{i=1}^n (Y_i-\widehat{Y}_i)^2 + 2\underbrace{\sum_{i=1}^n (Y_i-\widehat{Y}_i)\widehat{Y}_i}_{=0} \\ &= \sum_{i=1}^n \widehat{Y}_i^2 + \sum_{i=1}^n (Y_i-\widehat{Y}_i)^2 = \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}} + \big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)^\top\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)\,.\end{aligned}$$

- Here, the last but one equality holds due to the following consideration:

$$\begin{aligned}\sum_{i=1}^n (Y_i-\widehat{Y}_i)\widehat{Y}_i &= (\mathbf{Y}-\widehat{\mathbf{Y}})^\top\widehat{\mathbf{Y}} = (\mathbf{Y}^\top-\widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top)\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Y}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\top\underbrace{\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}}_{=\mathbf{X}^\top\mathbf{Y}} \\ &= \mathbf{Y}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{Y} = \big(\mathbf{Y}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}\big)^\top - \widehat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{Y} = 0\,.\end{aligned}$$

□

Remark

- The first summand $\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ on the right side of (42) is the squared length of the vector $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ of the estimated target values $\widehat{Y}_1,\ldots,\widehat{Y}_n$.
- The second component of the decomposition (42), i.e., the sum of the deviation squares $\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)^\top\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)$, is called *residual variance*.
- Sometimes the so–called *coefficient of determination* R^2 is considered as well, which is given by

$$R^2 = 1 - \frac{\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)^\top\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)}{\sum_{i=1}^n (Y_i-\overline{Y})^2}\,, \qquad \text{where} \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i\,.$$

Our model assumption that the design matrix $\mathbf{X}$ has full rank, i.e., $\mathrm{rk}(\mathbf{X}) = m$, implies that the inequality $(\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta} > 0$ holds if the hypothesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ is not true.

- Therefore, it is natural to reject the hypothesis H_0 if the squared length $\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ of the random vector $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is sufficiently large.
- In order to decide what "sufficiently large" means in this context, we also consider the variability σ^2 of the data.
  - Assuming that $H_0 : \boldsymbol{\beta} = \mathbf{o}$ is true, it holds that

$$\mathbb{E}\,\big(\mathbf{Y}^\top\mathbf{Y}\big) = \mathbb{E}\,\big(\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\big) = \mathbb{E}\,\Big(\sum_{i=1}^n \varepsilon_i^2\Big) = \sum_{i=1}^n \mathbb{E}\,\varepsilon_i^2 = n\sigma^2\,.$$

  - In this case, due to Theorem 2.9, one has

$$n\sigma^2 = \mathbb{E}\,\big(\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}\big) + \mathbb{E}\,\big((\mathbf{Y}-\widehat{\mathbf{Y}})^\top\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)\big)\,,$$

  - which is the reason why the quotient of $\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ and the sum of deviation squares $\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)^\top\big(\mathbf{Y}-\widehat{\mathbf{Y}}\big)$ is considered for testing the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{o}$.

- More precisely, we consider the following test statistic

$$T_{\text{mod}} = \frac{\widehat{\boldsymbol{\beta}}^{\top}(\mathbf{X}^{\top}\mathbf{X})\widehat{\boldsymbol{\beta}}}{mS^2} \,. \tag{43}$$

For being able to construct a test of the hypothesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ based on T_{mod}, the distribution of the test statistic T_{mod} has to be determined.

Theorem 2.10 *Assuming that $H_0 : \beta_1 = \ldots = \beta_m = 0$ is true, it holds that*

$$T_{\text{mod}} \sim \mathrm{F}_{m,n-m} \,, \tag{44}$$

i.e., the test statistic T_{mod} given in (43) has an F–distribution with $(m, n-m)$ degrees of freedom.

Proof

- Assuming that $H_0 : \beta_1 = \ldots = \beta_m = 0$ is true, it holds that $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$ with $\mathbf{K} = \sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1}$.
- This implies that $(\sigma^{-1}\mathbf{X})^{\top}(\sigma^{-1}\mathbf{X})\mathbf{K} = (\sigma^{-1}\mathbf{X})^{\top}(\sigma^{-1}\mathbf{X})\sigma^2(\mathbf{X}^{\top}\mathbf{X})^{-1} = \mathbf{I}$, i.e., in particular, that the matrix $(\sigma^{-1}\mathbf{X})^{\top}(\sigma^{-1}\mathbf{X})\mathbf{K}$ is idempotent.
- Now it follows from Theorem 1.9 that the quadratic form $\sigma^{-2}\widehat{\boldsymbol{\beta}}^{\top}(\mathbf{X}^{\top}\mathbf{X})\widehat{\boldsymbol{\beta}}$ has a (noncentral) χ^2–distribution with m degrees of freedom.
- Moreover, we have shown in Theorem 2.7 that the radom variable $(n-m)S^2/\sigma^2$ has a (central) χ^2–distribution with $n-m$ degrees of freedom.
- In Theorem 2.8 we have shown that $\widehat{\boldsymbol{\beta}}$ and S^2 are independent.
- Thus, it follows from the transformation theorem for independent random vectors (cf. Theorem I–1.8) that the random variables $\sigma^{-2}\widehat{\boldsymbol{\beta}}^{\top}(\mathbf{X}^{\top}\mathbf{X})\widehat{\boldsymbol{\beta}}$ and $(n-m)S^2/\sigma^2$ are independent as well.
- Now the statement follows from the definition of the F–distribution, cf. Section I–3.1.3. □

Remark

- When testing the hypothesis $H_0 : \beta_1 = \ldots = \beta_m = 0$ with a significance level of $\alpha \in (0,1)$ (against the alternative $H_1 : \beta_j \neq 0$ for at least one $j \in \{1, \ldots, m\}$), the null hypothesis H_0 is rejected if

$$T_{\text{mod}} > \mathrm{F}_{m,n-m,1-\alpha} \,, \tag{45}$$

  where $\mathrm{F}_{m,n-m,1-\alpha}$ denotes the $1-\alpha$ quantile of the F–distribution with $(m, n-m)$ degrees of freedom.
- In a similar way, an F–test for the verification of the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ with significance level $\alpha \in (0,1)$ (against the alternative $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$) for an arbitrary hypothetical parameter vector $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})$ can be constructed.
- Proceeding as in the proof of Theorem 2.10, one can show that if $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ is true, the test statistic

$$T_{\boldsymbol{\beta}_0} = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\top}(\mathbf{X}^{\top}\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{mS^2} \tag{46}$$

  has an F–distribution with $(m, n-m)$ degrees of freedom.
- Thus, the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ is rejected if

$$T_{\boldsymbol{\beta}_0} > \mathrm{F}_{m,n-m,1-\alpha} \,. \tag{47}$$

For the verification of hypotheses about *single* components of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^{\top}$, t–tests are used instead.

- Let $j \in \{1, \ldots, m\}$. In order to test a hypothetical value $\beta_{0,j}$ of the j–th component β_j of the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$, we consider the test statistic

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{S\sqrt{x^{jj}}}\,, \tag{48}$$

  where x^{ij} denotes the entry of the (inverse) matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ at position (i, j).

- From (40) – (41) and from the independence of $\widehat{\boldsymbol{\beta}}$ and S^2 it follows that $T_j \sim \mathrm{t}_{n-m}$.

- When testing the hypothesis $H_0 : \beta_j = \beta_{0,j}$ with a significance level of $\alpha \in (0, 1)$ (against the alternative $H_1 : \beta_j \neq \beta_{0,j}$), the null hypothesis H_0 is rejected if

$$\frac{|\widehat{\beta}_j - \beta_{0,j}|}{S\sqrt{x^{jj}}} > \mathrm{t}_{n-m,1-\alpha/2}\,, \tag{49}$$

  where $\mathrm{t}_{n-m,1-\alpha/2}$ denotes the $(1-\alpha/2)$–quantile of the t–distribution with $n-m$ degrees of freedom.

Remark

- The test of the hypothesis $H_0 : \beta_j = 0$ (against the alternative $H_1 : \beta_j \neq 0$) is particularly interesting because by using it, one can verify how far the response variables $Y_1, \ldots, Y_n$ depend on the j–th predictor at all.
- In this test of significance of the j–th predictor, the null hypothesis $H_0 : \beta_j = 0$ is rejected if

$$\frac{|\widehat{\beta}_j|}{S\sqrt{x^{jj}}} > \mathrm{t}_{n-m,1-\alpha/2}\,. \tag{50}$$

The tests we considered up to now in this section are special cases of the following *ubiquitious test*. Here, an *arbitrary* part of the components of the parameter vector $\boldsymbol{\beta}$ is tested.

- For $\ell \in \{1, \ldots, m\}$ and $\beta_{0\ell}, \ldots, \beta_{0m} \in \mathbb{R}$ the hypothesis

$$H_0 : \beta_\ell = \beta_{0\ell}, \ldots, \beta_m = \beta_{0m} \qquad \text{versus} \qquad H_1 : \beta_j \neq \beta_{0j} \text{ for at least one } j \in \{\ell, \ldots, m\} \tag{51}$$

  shall be tested.

- For this purpose, we consider the following $(m-\ell+1) \times (m-\ell+1)$–dimensional submatrix $\mathbf{K}_{\mathrm{uni}}$ of the matrix $(\mathbf{X}^\top\mathbf{X})^{-1} = (x^{ij})$ with

$$\mathbf{K}_{\mathrm{uni}} = \begin{pmatrix} x^{\ell\ell} & \ldots & x^{\ell m} \\ \vdots & & \vdots \\ x^{m\ell} & \ldots & x^{mm} \end{pmatrix}.$$

  – One can show that the inverse matrix $\mathbf{K}_{\mathrm{uni}}^{-1}$ is well–defined because $\mathbf{K}_{\mathrm{uni}} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top$, where $\mathbf{H} = (\mathbf{0}, \mathbf{I})$, the null matrix $\mathbf{0}$ has the dimension $(m-\ell+1) \times (\ell-1)$ and the identity matrix $\mathbf{I}$ has the dimension $(m-\ell+1) \times (m-\ell+1)$.

  – Herefrom and from Lemma 1.8 it follows that the matrix $\mathbf{K}_{\mathrm{uni}}$ is positive definite and thus invertible.

- A possible approach to solve the testing problem given in (51) is then given by the test statistic

$$T_{\mathrm{uni}} = \frac{\big(\widehat{\boldsymbol{\beta}}_{\mathrm{uni}} - \boldsymbol{\beta}_{\mathrm{uni}}\big)^\top \mathbf{K}_{\mathrm{uni}}^{-1} \big(\widehat{\boldsymbol{\beta}}_{\mathrm{uni}} - \boldsymbol{\beta}_{\mathrm{uni}}\big)}{(m-\ell+1)S^2}, \tag{52}$$

  where $\widehat{\boldsymbol{\beta}}_{\mathrm{uni}} = (\widehat{\beta}_\ell, \ldots, \widehat{\beta}_m)$ and $\boldsymbol{\beta}_{\mathrm{uni}} = (\beta_{0\ell}, \ldots, \beta_{0m})$.

- The following Theorem 2.11 implies that, assuming that the null hypothesis H_0 formulated in (51) is true,

$$T_{\text{uni}} \sim \mathrm{F}_{m-\ell+1,n-m}\,. \tag{53}$$

- Hence, the hypothesis $H_0 : \beta_\ell = \beta_{0,\ell}, \ldots, \beta_m = \beta_{0m}$ is rejected if

$$T_{\text{uni}} > \mathrm{F}_{m-\ell+1,n-m,1-\alpha}\,. \tag{54}$$

We now discuss one further test, a general *test for linear forms* of the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$.

- Let $r \in \{1, \ldots, m\}$, let $\mathbf{H}$ be an $r \times m$ matrix with full rank $\mathrm{rk}(\mathbf{H}) = r$, and let $\mathbf{c} \in \mathbb{R}^r$.
- The hypothesis to be tested is

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c} \qquad \text{versus} \qquad H_1 : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{c}\,, \tag{55}$$

  where the following test statistic $T_{\mathbf{H}}$ is considered:

$$T_{\mathbf{H}} = \frac{\big(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{c}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{c}\big)}{rS^2}\,. \tag{56}$$

Theorem 2.11 *Assuming that $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ is true, it holds that*

$$T_{\mathbf{H}} \sim \mathrm{F}_{r,n-m}\,, \tag{57}$$

i.e., the test statistic $T_{\mathbf{H}}$ given in (56) *has an* F*–distribution with $(r, n-m)$ degrees of freedom.*

Proof

- As the design matrix $\mathbf{X}$ has full rank, the symmetric matrix $\mathbf{X}^\top\mathbf{X}$ is positive definite.
  - Due to Lemma 1.8 the matrices $(\mathbf{X}^\top\mathbf{X})^{-1}$ and $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top$ are then positive definite as well,
  - i.e., in particular, the matrix $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top$ has full rank and thus is invertible.
- Therefore, the quantity $\mathbf{Z}^\top(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top)^{-1}\mathbf{Z}$ considered in (56) is well–defined, where

$$\mathbf{Z} = \mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{c} \qquad \text{with } \widehat{\boldsymbol{\beta}} \sim \mathrm{N}\big(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big).$$

- Theorem 1.3 implies that, assuming that $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ is true, it holds that

$$\mathbf{Z} \sim \mathrm{N}\big(\mathbf{o}, \sigma^2\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big)\,.$$

- Furthermore, the $r \times r$ matrix $\mathbf{A} = \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big)^{-1}$ is symmetric because

$$\begin{aligned}\mathbf{A}^\top &= \big(\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big)^{-1}\big)^\top = \big(\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big)^\top\big)^{-1} = \big(\mathbf{H}\big((\mathbf{X}^\top\mathbf{X})^{-1}\big)^\top\mathbf{H}^\top\big)^{-1} \\ &= \big(\mathbf{H}\big((\mathbf{X}^\top\mathbf{X})^\top\big)^{-1}\mathbf{H}^\top\big)^{-1} = \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big)^{-1} = \mathbf{A}\,.\end{aligned}$$

- As the matrix $\big(\sigma^{-2}\mathbf{A}\big)\big(\sigma^2\mathbf{H}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{H}^\top\big) = \mathbf{I}$ obviously is idempotent, Theorem 1.9 implies that $\sigma^{-2}\mathbf{Z}^\top\mathbf{A}\mathbf{Z}$ is a random variable having a χ^2_r–distribution.
- The rest of the proof continues in the same way as the proof of Theorem 2.10. □

Remark The null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$ is rejected if $T_{\mathbf{H}} > \mathrm{F}_{r,n-m,1-\alpha}$, where $T_{\mathbf{H}}$ is the test statistic given in (56).

2.2.4 Confidence Regions; Prediction of Response Variables

- *Recall:* In Section 2.2.3 we have considered the test statistic $T_j = (\widehat{\beta}_j - \beta_j)/(S\sqrt{x^{jj}})$, where x^{ij} denotes the entry of the (inverse) matrix $(\mathbf{X}^\top \mathbf{X})^{-1}$ at position (i, j).
- In doing so, we have shown that $T_j \sim \mathrm{t}_{n-m}$ for each $j \in \{1, \ldots, m\}$.
- This leads to the following confidence intervals with confidence level $1 - \alpha \in (0, 1)$ for each single regression coefficient β_j.
- It holds for each $j \in \{1, \ldots, m\}$ with probability $1 - \alpha$ that

$$\widehat{\beta}_j - \mathrm{t}_{n-m,1-\alpha/2} S\sqrt{x^{jj}} < \beta_j < \widehat{\beta}_j + \mathrm{t}_{n-m,1-\alpha/2} S\sqrt{x^{jj}}\,. \tag{58}$$

Remark

- In the same way as in the proof of Theorem I–5.8 a *common confidence region* with confidence level $1 - \alpha \in (0, 1)$ for all m regression coefficients $\beta_1, \ldots, \beta_m$ can be derived by use of the Bonferroni inequality (cf. Lemma I–5.4).
- Indeed, the probability that

$$\widehat{\beta}_j - \mathrm{t}_{n-m,1-\alpha/2m} S\sqrt{x^{jj}} < \beta_j < \widehat{\beta}_j + \mathrm{t}_{n-m,1-\alpha/2m} S\sqrt{x^{jj}} \tag{59}$$

  for all $j = 1, \ldots, m$ at the same time is at least equal to $1 - \alpha$.
- Moreover, Theorem 2.10 leads to an *exact* common confidence region with confidence level $1 - \alpha$ for all m regression coefficients $\beta_1, \ldots, \beta_m$.
  - It holds (cf. (46) – (47)) that

$$\mathbb{P}_{\boldsymbol{\beta}}\Big(\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{mS^2} < \mathrm{F}_{m,n-m,1-\alpha} \Big) = 1 - \alpha\,.$$

  - Here, the confidence region E with

$$E = \Big\{ \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m) : \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{mS^2} < \mathrm{F}_{m,n-m,1-\alpha} \Big\}$$

    forms a (random) *ellipsoid* with center $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)$.
- One can show that the ellipsoid E can be embedded into an m–dimensional paraxial cuboid $E' \supset E$, where

$$E' = \prod_{j=1}^{m} \Big(\widehat{\beta}_j - S\sqrt{m x^{jj} \mathrm{F}_{m,n-m,1-\alpha}},\ \widehat{\beta}_j + S\sqrt{m x^{jj} \mathrm{F}_{m,n-m,1-\alpha}} \Big)\,.$$

- The confidence region E' has a simpler form than E. However, due to $E' \supset E$ it is clear that E' is an estimation, which is less accurate than E.

In a similar way, a *confidence interval for the expected target value*

$$\varphi(x_{01}, \ldots, x_{0m}) = \beta_1 x_{01} + \ldots + \beta_m x_{0m}$$

corresponding to a given vector $\mathbf{x}_0 = (x_{01}, \ldots, x_{0m})^\top \in \mathbb{R}^m$ of values $x_{01}, \ldots, x_{0m}$ of the m predictor variables can be derived.

- For this purpose, we consider the $1 \times m$ matrix $\mathbf{H} = (x_{01}, \ldots, x_{0m})$ $(= \mathbf{x}_0^\top)$.

- Then, Theorem 2.11 implies that

$$\sqrt{T_{\mathbf{H}}} = \frac{|\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - \varphi(\mathbf{x}_0)|}{S\sqrt{\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}} \stackrel{\mathrm{d}}{=} |T|\,,$$

  where T is a random variable having a t–distribution with $n - m$ degrees of freedom.

- Hence, it holds with probability $1 - \alpha$ that

$$\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - Z_0 < \varphi(\mathbf{x}_0) < \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 + Z_0\,, \tag{60}$$

  where

$$Z_0 = \mathrm{t}_{n-m,1-\alpha/2} S \sqrt{\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}\,.$$

Remark

- Analogously, one can derive a *prediction interval* for the response variable $Y_0 = \beta_1 x_{01} + \ldots + \beta_m x_{0m} + \varepsilon_0$, where the error term ε_0 is normally distributed and independent of the error terms $\varepsilon_1, \ldots, \varepsilon_n$; $\varepsilon_0 \sim \mathrm{N}(0, \sigma^2)$.
- Indeed, it is $\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - Y_0 \sim \mathrm{N}\big(0, \sigma^2(1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0)\big)$ and thus

$$\widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 - Z_0' < Y_0 < \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_0 + Z_0'\,, \tag{61}$$

  with probability $1 - \alpha$, where $Z_0' = \mathrm{t}_{n-m,1-\alpha/2} S \sqrt{1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}$.

2.2.5 Confidence Band

In this section we assume that the design matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \ldots & x_{nm} \end{pmatrix}, \tag{62}$$

i.e., we consider the (multiple) linear regression model.

- In the definition of the regression function $\varphi(x_1, \ldots, x_m) = \beta_1 x_1 + \ldots + \beta_m x_m$ in (4), we now set $x_1 = 1$ and determine a *confidence band for the regression hyperplane*

$$y = \varphi(1, x_2, \ldots, x_m) = \beta_1 + \beta_2 x_2 + \ldots + \beta_m x_m\,, \qquad \forall\, x_2, \ldots, x_m \in \mathbb{R}\,.$$

- This means that we need to find a number $a_\gamma > 0$, such that with the given (coverage) probability $\gamma = 1 - \alpha \in (0, 1)$ it holds that

$$\widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_m x_m - a_\gamma Z_{\mathbf{x}} < \varphi(1, x_2, \ldots, x_m) < \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_m x_m + a_\gamma Z_{\mathbf{x}} \tag{63}$$

  for each $\mathbf{x} = (1, x_2, \ldots, x_m) \in \mathbb{R}^m$ simultaneously, where

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \qquad \text{and} \qquad Z_{\mathbf{x}} = S\sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}}\,.$$

For solving this problem, the following lemma is useful.

Lemma 2.3 *It holds with probability* 1 *that*

$$\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\left((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\right)^2}{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} = (\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})\,, \tag{64}$$

where $\mathbb{R}_1^{m-1}$ *denotes the set of all vectors* $\mathbf{x}\in\mathbb{R}^m$ *with* $\mathbf{x}=(1,x_2,\ldots,x_m)^\top$.

Proof

- From Lemmas 1.6 and 1.8 it follows that $(\mathbf{X}^\top\mathbf{X})^{-1}=\mathbf{H}\mathbf{H}^\top$ for an invertible $m\times m$ matrix $\mathbf{H}$.
- Therefore, the expression
$$(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} = ((\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon})^\top\mathbf{H}^\top\mathbf{x}$$
can be perceived as the scalar product of the m–dimensional vectors $(\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon}$ and $\mathbf{H}^\top\mathbf{x}$.
- Analogously, it holds that
$$(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} = ((\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon} \qquad \text{and} \qquad \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x} = (\mathbf{H}^\top\mathbf{x})^\top\mathbf{H}^\top\mathbf{x}\,.$$
- From this result and from the inequality
$$|\mathbf{y}^\top\mathbf{z}| \le \sqrt{\mathbf{y}^\top\mathbf{y}}\,\sqrt{\mathbf{z}^\top\mathbf{z}} \qquad \forall\,\mathbf{y},\mathbf{z}\in\mathbb{R}^m \tag{65}$$
together with $\mathbf{y}=(\mathbf{X}\mathbf{H})^\top\boldsymbol{\varepsilon}$ and $\mathbf{z}=\mathbf{H}^\top\mathbf{x}$ it follows that
$$\left|(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\right| \le \sqrt{(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})}\,\sqrt{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}$$
and thus
$$\frac{\left((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\right)^2}{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} \le (\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})\,. \tag{66}$$
- As the random vector $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^\top$ has independent and absolutely continuous components, we get that $\sum_{i=1}^n\varepsilon_i\neq 0$ with probability 1.
- Now let $\sum_{i=1}^n\varepsilon_i\neq 0$. Then it follows from the form of the design matrix $\mathbf{X}$ considered in (62) that the vector $\mathbf{x}=\mathbf{X}^\top\boldsymbol{\varepsilon}/\sum_{i=1}^n\varepsilon_i$ belongs to $\mathbb{R}_1^{m-1}$ and that in this case the equality in (66) holds . □

The following result, which is a vectorial generalization of Theorem I–5.9, leads to the desired confidence band.

Theorem 2.12 *Let* $a_\gamma=\sqrt{m\,\mathrm{F}_{m,n-m,\gamma}}$. *Then it holds that*

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\left(\widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\varphi(\mathbf{x})\right)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} \le a_\gamma^2\right) = \gamma\,. \tag{67}$$

Proof

- For each $\mathbf{x}\in\mathbb{R}_1^{m-1}$ it holds that
$$\begin{aligned}\widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\varphi(\mathbf{x}) &= \widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\boldsymbol{\beta}^\top\mathbf{x} = \left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right)^\top\mathbf{x}-\boldsymbol{\beta}^\top\mathbf{x} = \left(\boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\right)^\top\mathbf{x}-\boldsymbol{\beta}^\top\mathbf{x}\\ &= (\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\end{aligned}$$
and thus
$$\begin{aligned}\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\left(\widehat{\boldsymbol{\beta}}^\top\mathbf{x}-\varphi(\mathbf{x})\right)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} &= \max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\left((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\right)^2}{S^2\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} = \frac{1}{S^2}\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\left((\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\right)^2}{\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}}\\ &= \frac{(\mathbf{X}^\top\boldsymbol{\varepsilon})^\top(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\boldsymbol{\varepsilon})}{S^2}\,,\end{aligned}$$
where the last equality follows from Lemma 2.3.

- Therefore, one has
$$\max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\big(\widehat{\boldsymbol{\beta}}^\top \mathbf{x} - \varphi(\mathbf{x})\big)^2}{S^2 \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}} = \frac{(\mathbf{X}^\top \boldsymbol{\varepsilon})^\top (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \boldsymbol{\varepsilon})}{S^2} \,. \tag{68}$$
- Due to $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2 \mathbf{I})$ and
$$\mathbf{X}^\top \Big(\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \Big) = \mathbf{0} \,,$$
Theorem 1.10 implies that $\mathbf{X}^\top \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^\top \big(\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \big) \boldsymbol{\varepsilon}$ are independent.
- Hence, it follows from the representation formula
$$S^2 = \frac{1}{n-m} \, \boldsymbol{\varepsilon}^\top \big(\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \big) \boldsymbol{\varepsilon},$$
which has already been derived in Theorem 2.7, that also the random variables $(\mathbf{X}^\top \boldsymbol{\varepsilon})^\top (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \boldsymbol{\varepsilon})$ and S^2 are independent.
- In Theorem 2.7 we have shown that
$$(n-m) S^2 / \sigma^2 \sim \chi^2_{n-m} \,.$$
- Moreover, Theorem 1.9 implies that
$$(\mathbf{X}^\top \boldsymbol{\varepsilon})^\top (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \boldsymbol{\varepsilon}) / \sigma^2 \sim \chi^2_m$$
since the $m \times m$ (covariance) matrix $\mathbf{X}^\top \mathbf{X}$ of the normally distributed random vector $\mathbf{X}^\top \boldsymbol{\varepsilon}$ has full rank and since the matrix $(\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) = \mathbf{I}$ is idempotent.
- Due to (68) we have altogether shown that
$$\frac{1}{m} \max_{\mathbf{x}\in\mathbb{R}_1^{m-1}} \frac{\big(\widehat{\boldsymbol{\beta}}^\top \mathbf{x} - \varphi(\mathbf{x})\big)^2}{S^2 \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}} \sim \mathrm{F}_{m,n-m} \,.$$
- For the threshold value considered in (63) and (67), we hence obtain $a_\gamma = \sqrt{m \, \mathrm{F}_{m,n-m,\gamma}}$. □

3 Arbitrary Design Matrix; Generalized Inverse

- We now consider the following generalization of the linear model discussed in Chapter 2,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{1}$$

for which we have assumed so far that the design matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \tag{2}$$

is an $(n \times m)$–dimensional matrix with full (column) rank $\operatorname{rk}(\mathbf{X}) = m$, where $n \geq m$.

- In contrast to this, in this chapter we will consider the case that $\operatorname{rk}(\mathbf{X}) \leq m$, i.e., we allow that the design matrix $\mathbf{X}$ does *not* have full rank.
- As we did in Section 2.1, we first assume for the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ that

$$\mathbb{E}\,\varepsilon_i = 0\,, \qquad \operatorname{Var}\varepsilon_i = \sigma^2\,, \qquad \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0\,, \qquad \forall\, i, j = 1, \dots, n \text{ with } i \neq j \tag{3}$$

for a certain (unknown) $\sigma^2 > 0$.

3.1 Analysis of Variance as a Linear Model

To begin with, we discuss two examples of problems leading to linear models whose design matrix does not have full rank, cf. Section 3.4.

The term „analysis of variance" does not mean that variances of random variables are analyzed in this context, but refers to the analysis of the *variability of expectations*. In literature, ANOVA is typically used as an abbreviation.

3.1.1 One–Factor Analysis of Variance; ANOVA Null Hypothesis

- In a one–factor analysis of variance, we assume that the random sample $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ can be partitioned into k classes of subsamples $(Y_{ij},\ j = 1, \dots, n_i)$, where
  - $n_i > 1$ for each $i = 1, \dots, k$ and $\sum_{i=1}^{k} n_i = n$
  - and the sampling variables belonging to the same class have the same expectation θ_i.
- In other words: We assume that

$$Y_{ij} = \theta_i + \varepsilon_{ij}\,, \qquad \forall\, i = 1, \dots, k,\ j = 1, \dots, n_i\,, \tag{4}$$

where $\theta_1, \dots, \theta_k \in \mathbb{R}$ are (unknown) parameters and the error terms $\varepsilon_{ij} : \Omega \to \mathbb{R}$ are uncorrelated with

$$\mathbb{E}\,\varepsilon_{ij} = 0\,, \qquad \operatorname{Var}\varepsilon_{ij} = \sigma^2\,, \qquad \forall\, i = 1, \dots, k,\ j = 1, \dots, n_i\,. \tag{5}$$

Remark

- The numbers $i = 1, \dots, k$ of the classes $(Y_{ij},\ j = 1, \dots, n_i)$ are interpreted as *levels of a predictor variable*.

- The model assumptions made above imply in particular that the observed values $y_1, \ldots, y_n$ of the response variables $Y_1, \ldots, Y_n$ can be structured in table form as follows:

level	1	2	3	...	k
	y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
	y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	y_{k3}
			y_{3n_3}		$\vdots$
	y_{1n_1}				
		y_{2n_2}			y_{kn_k}

We show that the classical ANOVA null hypothesis $H_0 : \theta_1 = \ldots = \theta_k$ can be expressed by use of so–called contrasts.

- For this purpose, we consider the following set $\mathcal{A} \subset \mathbb{R}^k$ with

$$\mathcal{A} = \left\{ \mathbf{a} = (a_1, \ldots, a_k)^\top : \mathbf{a} \neq \mathbf{o}, \sum_{i=1}^{k} a_i = 0 \right\}.$$

- Let $\mathbf{t} = (t_1, \ldots, t_k)^\top \in \mathbb{R}^k$ be an arbitrary vector of variables and let $\mathbf{a} = (a_1, \ldots, a_k)^\top \in \mathcal{A}$ be a vector of (known) constants. The mapping $\mathbf{t} \to \sum_{i=1}^{k} a_i t_i$ is then called a *contrast*.

Lemma 3.1 *Let $\theta_1, \ldots, \theta_k \in \mathbb{R}$ be arbitrary real numbers. For the validity of $\theta_1 = \ldots = \theta_k$ it is then necessary and sufficient that*

$$\sum_{i=1}^{k} a_i \theta_i = 0 \qquad \forall\, \mathbf{a} \in \mathcal{A}\,. \tag{6}$$

Proof

- If $\theta_1 = \ldots = \theta_k = \theta$ is true, we get for each $\mathbf{a} \in \mathcal{A}$ that

$$\sum_{i=1}^{k} a_i \theta_i = \theta \sum_{i=1}^{k} a_i = 0\,.$$

- In order to show the sufficiency of the condition, we consider the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_{k-1} \in \mathcal{A}$ with

$$\mathbf{a}_1 = (1, -1, 0, \ldots, 0)^\top, \quad \mathbf{a}_2 = (0, 1, -1, 0, \ldots, 0)^\top, \quad \ldots, \quad \mathbf{a}_{k-1} = (0, \ldots, 0, 1, -1)^\top.$$

- For each $i \in \{1, \ldots, k-1\}$ the validity of condition (6) for $\mathbf{a}_i$ implies that $-\theta_i + \theta_{i+1} = 0$, i.e., $\theta_i = \theta_{i+1}$. Therefore, it follows that $\theta_1 = \ldots = \theta_k$. □

Remark

- Due to Lemma 3.1, the classical ANOVA null hypothesis $H_0 : \theta_1 = \ldots = \theta_k$ is equivalent to the hypothesis $H_0 : \sum_{i=1}^{k} a_i \theta_i = 0$ for each $\mathbf{a} = (a_1, \ldots, a_k)^\top \in \mathcal{A}$.

- Moreover, assuming that H_0 is true, it is obvious that
  - $\sum_{i=1}^k a_i \overline{Y}_{i\cdot}$ with $\overline{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$ is an unbiased estimator for $\sum_{i=1}^k a_i\theta_i = 0$ for each $\mathbf{a} \in \mathcal{A}$,
  - the variance of $\sum_{i=1}^k a_i \overline{Y}_{i\cdot}$ is given by $\mathrm{Var} \sum_{i=1}^k a_i \overline{Y}_{i\cdot} = \sigma^2 \sum_{i=1}^k a_i^2/n_i$
  - and
$$S_p^2 = \frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} \big(Y_{ij} - \overline{Y}_{i\cdot}\big)^2 \tag{7}$$
  is an unbiased estimator for σ^2, the so–called *pooled sample variance*.
- Hence, it is reasonable to reject $H_0 : \theta_1 = \ldots = \theta_k$ if the supremum over $\mathbf{a} \in \mathcal{A}$ of the (suitably normed) absolute values of $\sum_{i=1}^k a_i \overline{Y}_{i\cdot}$ exceeds a certain threshold value, where the test statistic $\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2$ is considered with $T_{\mathbf{a}} = \Big(\sum_{i=1}^k a_i \overline{Y}_{i\cdot}\Big) \Big/ \sqrt{S_p^2 \sum_{i=1}^k a_i^2/n_i}$.
- In a similar way as in the proof of Lemma 2.3 one can show that, assuming that $H_0 : \theta_1 = \ldots = \theta_k$ is true, it holds that
$$\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2 = \frac{\sum_{i=1}^k n_i \big(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\big)^2}{S_p^2}\,, \tag{8}$$
  where $\overline{Y}_{\cdot\cdot} = \sum_{i=1}^k n_i \overline{Y}_{i\cdot} \Big/ \sum_{i=1}^k n_i$.

The following *decomposition* implies an intuitive interpretation of numerator and denominator of the test statistic $\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2$ considered in (8), cf. also Theorem 2.9.

Theorem 3.1 *It holds that*
$$\sum_{i=1}^k \sum_{j=1}^{n_i} \big(Y_{ij} - \overline{Y}_{\cdot\cdot}\big)^2 = \sum_{i=1}^k n_i \big(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\big)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} \big(Y_{ij} - \overline{Y}_{i\cdot}\big)^2 . \tag{9}$$

Proof By expanding the left–hand side of (9), one obtains that
$$\begin{aligned}
\sum_{i=1}^k \sum_{j=1}^{n_i} \big(Y_{ij} - \overline{Y}_{\cdot\cdot}\big)^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} \big((Y_{ij} - \overline{Y}_{i\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\big)^2 \\
&= \sum_{i=1}^k \sum_{j=1}^{n_i} \big((Y_{ij} - \overline{Y}_{i\cdot})^2 + 2(Y_{ij} - \overline{Y}_{i\cdot})(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2\big) \\
&= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2 + 2 \sum_{i=1}^k (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) \underbrace{\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})}_{=0} + \sum_{i=1}^k n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 .
\end{aligned}$$
□

Remark

- The double sum on the left–hand side of (9) can be interpreted as a measure for the (total) *variability of the sampling variables* $\{Y_{ij},\, i = 1, \ldots, k,\, j = 1, \ldots, n_i\}$.
- The first sum on the right–hand side of (9) is a measure for the variability *between* the levels of the predictor variable, while the double sum on the right–hand side of (9) is a measure for the variability *within* the levels of the predictor variable.
- So due to the definition of S_p^2 given in (7), the test statistic considered in (8) is proportional to the ratio of the variability between the levels and the variability within the levels of the predictor variable.
- Therefore, the ANOVA null hypothesis $H_0 : \theta_1 = \ldots = \theta_k$ is rejected if the variability between the levels is significantly higher than the variability within the levels of the predictor variable.

3.1.2 Reparametrization of the Expectations

The model of one–factor analysis of variance considered in Section 3.1.1 can be represented as a linear model in two different ways.

- In both cases, the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is "structured", i.e., we use the notation $\mathbf{Y} = \big(Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{k1}, \ldots, Y_{kn_k}\big)^\top$, where $n_1 + \ldots + n_k = n$.
- The random vector $\mathbf{Y}$ is represented in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the design matrix $\mathbf{X}$ and the parameter vector $\boldsymbol{\beta}$ are chosen differently in each case.
  - While $\mathbf{X}$ has full rank in the first case, it does not have full rank in the second case.
  - The second (reparametrized) representation is especially useful for the application of general estimation and test methods, which are discussed in Sections 3.2 and 3.3.
  - If the error terms are normally distributed, we can thus determine the distribution of the test statistic $\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2$ considered in (8), assuming that $H_0 : \theta_1 = \ldots = \theta_k$ is true, cf. formula (89) in Section 3.4.1.

Case 1

- In this case the design matrix $\mathbf{X}$ is given by the $n \times k$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \tag{10}$$

  and the parameter vector $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = (\theta_1, \ldots, \theta_k)^\top$.

Case 2

- We consider the following *reparametrization* of the expectations $\theta_1, \ldots, \theta_k$, which correspond to the levels of the predictor variable.
- Let $\mu \in \mathbb{R}$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ be real numbers, such that

$$\theta_i = \mu + \alpha_i\,, \qquad \forall\, i = 1, \ldots, k \tag{11}$$

  and

$$\sum_{i=1}^{k} n_i \alpha_i = 0\,. \tag{12}$$

- Then the random sample $\mathbf{Y}$ of the model of one–factor analysis of variance can also be written in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the design matrix $\mathbf{X}$, however, is now given by the $n \times (k+1)$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{13}$$

and the parameter vector $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_k)^\top$.

Remark

- The linear additional condition (12) for the components $\alpha_1, \dots, \alpha_k$ of the parameter vector $\boldsymbol{\beta}$ ensures that the representation (11) – (12) of the expectations $\theta_1, \dots, \theta_k$ is unique.
- Furthermore, (11) and (12) imply that

$$\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbb{E}\, Y_{ij} = \mu\,,$$

  where
  - the parameter μ can be interpreted as *general mean* of the expectations $\mathbb{E}\, Y_{ij}$ of the sampling variables Y_{ij} and
  - the (deviation) parameter α_i is called the *effect* of the i-th level of the predictor variable.
- For the design matrix $\mathbf{X}$ given in (13) it holds that $\operatorname{rk}(\mathbf{X}) = k$, i.e., the $n \times (k+1)$–dimensional matrix $\mathbf{X}$ does not have full column rank.

Theorem 3.2 *It holds that*

$$\mathbb{E}\, \overline{Y}_{..} = \mu \qquad \text{and} \qquad \mathbb{E}\left(\overline{Y}_{i\,.} - \overline{Y}_{..}\right) = \alpha_i \tag{14}$$

for each $i = 1, \dots, k$, i.e., $\overline{Y}_{..}$ and $\overline{Y}_{i\,.} - \overline{Y}_{..}$ define unbiased estimators for the model parameters μ and α_i, respectively.

Proof It follows from the definition of $\overline{Y}_{..}$ that

$$\mathbb{E}\, \overline{Y}_{..} \;=\; \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbb{E}\, Y_{ij} \;=\; \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \theta_i \;=\; \mu + \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \alpha_i \;=\; \mu\,,$$

where the last equality follows from the reparametrization condition (12). The second part of the statement in (14) can be proved analogously. □

3.1.3 Two–Factor Analysis of Variance

- We now modify the model of one–factor analysis of variance introduced in Section 3.1.1 and assume that the response variables $Y_1, \dots, Y_n$ depend on *two* predictor variables.
- Thus, we partition the random sample $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ into $k_1 \cdot k_2$ subsamples $(Y_{i_1 i_2 j},\, j = 1, \dots, n_{i_1 i_2})$, where $n_{i_1 i_2} > 1$ for all $i_1 = 1, \dots, k_1$ and $i_2 = 1, \dots, k_2$ and

$$\sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} n_{i_1 i_2} = n\,.$$

- We assume that the sampling variables belonging to the same class have the same expectation $\theta_{i_1 i_2}$ in each case.
- In other words: We assume that

$$Y_{i_1 i_2 j} = \theta_{i_1 i_2} + \varepsilon_{i_1 i_2 j}\,, \qquad \forall\, i_1 = 1, \dots, k_1,\ i_2 = 1, \dots, k_2,\ j = 1, \dots, n_{i_1 i_2}\,, \tag{15}$$

  where $\theta_{i_1 i_2} \in \mathbb{R}$ are (unknown) parameters and the error terms $\varepsilon_{i_1 i_2 j} : \Omega \to \mathbb{R}$ are uncorrelated with

$$\mathbb{E}\,\varepsilon_{i_1 i_2 j} = 0\,, \qquad \operatorname{Var} \varepsilon_{i_1 i_2 j} = \sigma^2\,, \qquad \forall\, i_1 = 1, \dots, k_1,\ i_2 = 1, \dots, k_2,\ j = 1, \dots, n_{i_1 i_2}\,. \tag{16}$$

Remark

- The representation (15) of the sampling variables $Y_{i_1 i_2 j}$ leads to the same form of linear model as it was considered in Case 1 of Section 3.1.2.
- The numbers $i_1 = 1, \dots, k_1$ and $i_2 = 1, \dots, k_2$ of the classes $(Y_{i_1 i_2 j},\, j = 1, \dots, n_{i_1 i_2})$ are again interpreted as *levels* of the corresponding predictor variable.
- Here, the design matrix $\mathbf{X}$ has the dimension $n \times (k_1 \cdot k_2)$ and full column rank $k_1 \cdot k_2$.

Moreover, we consider a similar reparametrization of the expectations $\theta_{i_1 i_2}$ as in Section 3.1.2.

- In doing so, we only consider the so–called *balanced case*, i.e.,
  - we additionally assume that all $k_1 \cdot k_2$ subsamples $(Y_{i_1 i_2 j},\, j = 1, \dots, n_{i_1 i_2})$ have the same sample size.
  - Hence, let $n_{i_1 i_2} = r$ for all $i_1 = 1, \dots, k_1$ and $i_2 = 1, \dots, k_2$, where $r = n/(k_1 k_2)$.
- Let $\mu \in \mathbb{R}$ and for all $i_1 \in \{1, \dots, k_1\}$ and $i_2 \in \{1, \dots, k_2\}$ let $\alpha_{i_1}^{(1)} \in \mathbb{R}$, $\alpha_{i_2}^{(2)} \in \mathbb{R}$ and $\alpha_{i_1 i_2} \in \mathbb{R}$ be real numbers, such that

$$\theta_{i_1 i_2} = \mu + \alpha_{i_1}^{(1)} + \alpha_{i_2}^{(2)} + \alpha_{i_1 i_2}\,, \qquad \forall\, i_1 = 1, \dots, k_1,\ i_2 = 1, \dots, k_2 \tag{17}$$

  and

$$\sum_{i_1=1}^{k_1} \alpha_{i_1}^{(1)} = \sum_{i_2=1}^{k_2} \alpha_{i_2}^{(2)} = \sum_{i_1=1}^{k_1} \alpha_{i_1 i_2} = \sum_{i_2=1}^{k_2} \alpha_{i_1 i_2} = 0\,. \tag{18}$$

- Then, the random sample $\mathbf{Y}$ can be written in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where
  - the design matrix $\mathbf{X}$ is given by a matrix of dimension $n \times (1 + k_1 + k_2 + k_1 k_2)$, whose entries only consist of zeros and ones and which does not have full rank.
  - Therefore, the parameter vector $\boldsymbol{\beta}$ has the following form:

$$\boldsymbol{\beta} = \left(\mu, \alpha_1^{(1)}, \dots, \alpha_{k_1}^{(1)}, \alpha_1^{(2)}, \dots, \alpha_{k_2}^{(2)}, \alpha_{11}, \dots, \alpha_{k_1 k_2}\right)^\top .$$

Remark

- The additional linear conditions (18) for the components of the parameter vector $\boldsymbol{\beta}$ ensure, in a similar way as in the model of one–factor analysis of variance considered in Section 3.1.2, that the representation (17) – (18) of the expectations $\theta_{11}, \ldots, \theta_{k_1 k_2}$ is unique.
- Here,
  - μ can be perceived as *general mean* of the expectations $\mathbb{E}\, Y_{i_1 i_2 j}$ of the sampling variables $Y_{i_1 i_2 j}$,
  - $\alpha_{i_1}^{(1)}$ is called *main effect* of the i_1-th level of the first predictor variable,
  - $\alpha_{i_2}^{(2)}$ is called *main effect* of the i_2-th level of the second predictor variable and
  - $\alpha_{i_1 i_2}$ is called *interaction* between the levels i_1 and i_2 of the level combination (i_1, i_2).

For the construction of estimators for the model parameters μ, $\alpha_{i_1}^{(1)}$, $\alpha_{i_2}^{(2)}$ and $\alpha_{i_1 i_2}$, we use the following notation: Let

$$Y_{i_1 \cdot \cdot} = \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} Y_{i_1 i_2 j}\,, \qquad Y_{\cdot\, i_2 \cdot} = \sum_{i_1=1}^{k_1} \sum_{j=1}^{r} Y_{i_1 i_2 j}\,, \qquad Y_{i_1 i_2 \cdot} = \sum_{j=1}^{r} Y_{i_1 i_2 j} \tag{19}$$

and

$$\overline{Y}_{i_1 \cdot \cdot} = \frac{1}{r k_2}\, Y_{i_1 \cdot \cdot}\,, \qquad \overline{Y}_{\cdot\, i_2 \cdot} = \frac{1}{r k_1}\, Y_{\cdot\, i_2 \cdot}\,, \qquad \overline{Y}_{i_1 i_2 \cdot} = \frac{1}{r}\, Y_{i_1 i_2 \cdot}\,, \qquad \overline{Y}_{\ldots} = \frac{1}{r k_1 k_2} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} Y_{i_1 i_2 j} \tag{20}$$

Theorem 3.3 *It holds that*

$$\mathbb{E}\, \overline{Y}_{\ldots} = \mu\,, \quad \mathbb{E}\left(\overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\ldots}\right) = \alpha_{i_1}^{(1)}\,, \quad \mathbb{E}\left(\overline{Y}_{\cdot\, i_2 \cdot} - \overline{Y}_{\ldots}\right) = \alpha_{i_2}^{(2)}\,, \quad \mathbb{E}\left(\overline{Y}_{\ldots} + \overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\cdot\, i_2 \cdot}\right) = \alpha_{i_1 i_2} \tag{21}$$

for arbitrary $i_1 = 1, \ldots, k_1$, $i_2 = 1, \ldots, k_2$, i.e., $\overline{Y}_{\ldots}$, $\overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\ldots}$, $\overline{Y}_{\cdot\, i_2 \cdot} - \overline{Y}_{\ldots}$ and $\overline{Y}_{\ldots} + \overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\cdot\, i_2 \cdot}$ define unbiased estimators for the model parameters μ, $\alpha_{i_1}^{(1)}$, $\alpha_{i_2}^{(2)}$ and $\alpha_{i_1 i_2}$, respectively.

Proof It follows from the definition of $\overline{Y}_{\ldots}$ in (20) that

$$\begin{aligned} \mathbb{E}\, \overline{Y}_{\ldots} &= \frac{1}{r k_1 k_2} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} \mathbb{E}\, Y_{i_1 i_2 j} = \frac{1}{k_1 k_2} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \theta_{i_1 i_2} \\ &= \mu + \frac{1}{k_1 k_2} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \left(\alpha_{i_1}^{(1)} + \alpha_{i_2}^{(2)} + \alpha_{i_1 i_2}\right) = \mu\,, \end{aligned}$$

where the last equality follows from the reparametrization conditions (18). The remaining three parts of the statement in (21) can be proved analogously. □

Remark

- The conditions (18), i.e., the assumption that the parameter vector $\boldsymbol{\beta}$ belongs to a linear subspace of $\mathbb{R}^{1+k_1+k_2+k_1 k_2}$, play a fundamental role in the proof of Theorem 3.3.
- Here, the conclusions of Theorem 3.3 can be interpreted as unbiasedness of the considered estimators with respect to this restricted parameter space.
- However, if we allow that $\boldsymbol{\beta}$ is an *arbitrary* vector of dimension $1 + k_1 + k_2 + k_1 k_2$, there is *no* LS–estimator for $\boldsymbol{\beta}$, which is unbiased at the same time, cf. the discussion at the end of Section 3.2.1.

The following result contains a *decomposition of sums of squared differences*, cf. also Theorems 2.9 and 3.1.

Theorem 3.4 *It holds that*

$$\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j}-\overline{Y}_{...}\big)^2 = rk_2\sum_{i_1=1}^{k_1}\big(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{...}\big)^2 + rk_1\sum_{i_2=1}^{k_2}\big(\overline{Y}_{\cdot i_2\cdot}-\overline{Y}_{...}\big)^2 + \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\big)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot i_2\cdot}+\overline{Y}_{...}\big)^2. \tag{22}$$

Proof Using the notation introduced in (19) and (20), we get that

$$\begin{aligned}
&\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(Y_{i_1i_2j}-\overline{Y}_{...}\Big)^2\\
&= \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(\big(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{...}\big)+\big(\overline{Y}_{\cdot i_2\cdot}-\overline{Y}_{...}\big)+\big(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\big)+\big(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot i_2\cdot}+\overline{Y}_{...}\big)\Big)^2\\
&= \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{...}\Big)^2+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(\overline{Y}_{\cdot i_2\cdot}-\overline{Y}_{...}\Big)^2+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\Big)^2\\
&\quad+\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(\overline{Y}_{i_1i_2\cdot}-\overline{Y}_{i_1\cdot\cdot}-\overline{Y}_{\cdot i_2\cdot}+\overline{Y}_{...}\Big)^2+R,
\end{aligned}$$

where it can be shown in a similar way as in the proof of Theorem 3.1 that the sum R of the mixed products is equal to zero. □

Remark

- The sum of squares on the left–hand side of (22) can be perceived as a measure for the (total) *variability of the sampling variables* $\{Y_{i_1i_2j},\, i_1=1,\ldots,k_1,\, i_2=1,\ldots,k_2,\, j=1,\ldots,r\}$.
- The first and second sum of squares on the right–hand side of (22) are measures for the variability *between* the levels of the first and second predictor variable, respectively, while the third sum of squares on the right–hand side of (22) is a measure for the variability *within* the pairs of levels (i_1,i_2) of the two predictor variables, the so–called *residual variance*.
- The fourth sum of squares on the right–hand side of (22) is a measure for the interactions *between* the components of the pairs of levels (i_1,i_2) of the two predictor variables.
- By similar considerations as in the proof of Theorem 2.5 it can be shown that a suitably normalized version of the residual variance is an unbiased estimator for the variance σ^2 of the error terms.
- In particular, it holds that $\mathbb{E}\,S^2=\sigma^2$, where

$$S^2=\frac{1}{k_1k_2(r-1)}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\Big(Y_{i_1i_2j}-\overline{Y}_{i_1i_2\cdot}\Big)^2.$$

3.2 Estimation of Model Parameters

Now we return to the analysis of the linear model with arbitrary design matrix $\mathbf{X}$ given in (1) – (3). In this section, we assume that

- $\mathrm{rk}(\mathbf{X})=r<m$, i.e., $\mathbf{X}$ has *not* full column rank and that
- $\boldsymbol{\beta}\in\mathbb{R}^m$ is an *arbitrary* m–dimensional vector, i.e., at first we do *not* consider any *additional conditions* of type (12) or (18).

3.2.1 LS–Estimator for $\boldsymbol{\beta}$

We first recall the following formula for the rank of quadratic matrices.

Lemma 3.2 *Let* $\mathbf{A}$ *be an arbitrary* $n \times n$ *matrix. Then it holds that*

$$\operatorname{rk}(\mathbf{A}) = n - \dim \ker(\mathbf{A})\,, \tag{23}$$

where $\ker(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{o}\}$ *and* $\dim \ker(\mathbf{A})$ *denotes the dimension of* $\ker(\mathbf{A}) \subset \mathbb{R}^n$.

Moreover, the following property of the rank of products of matrices is useful, which immediately follows from Lemma 3.2.

Lemma 3.3 *Let* $m, n, r \in \mathbb{N}$ *be arbitrary natural numbers and let* $\mathbf{A}, \mathbf{B}$ *be arbitrary* $m \times n$ *and* $n \times r$ *matrices. Then it holds that*

$$\operatorname{rk}(\mathbf{AB}) \leq \min\{\operatorname{rk}(\mathbf{A}), \operatorname{rk}(\mathbf{B})\}\,. \tag{24}$$

Remark

- As we now assume that the design matrix $\mathbf{X}$ does not have full rank, the $m \times m$ matrix $\mathbf{X}^\top\mathbf{X}$ is *not* invertible because Lemma 3.3 implies that $\operatorname{rk}\big(\mathbf{X}^\top\mathbf{X}\big) \leq \operatorname{rk}(\mathbf{X}) < m$.
- Therefore, the normal equation (2.9), i.e.,

  $$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}\,, \tag{25}$$

  does not have a uniquely determined solution.
- In order to specify the solution set of (25), we need the notion of the generalized inverse of a matrix.

Definition An $m \times n$ matrix $\mathbf{A}^-$ is called *generalized inverse* of the $n \times m$ matrix $\mathbf{A}$ if

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}\,. \tag{26}$$

In order to show that there always is a solution $\mathbf{A}^-$ of (26), we use the following general representation formula, which we state without proof at this point.

Lemma 3.4 *Let* $\mathbf{A}$ *be an* $n \times m$ *matrix with* $n \geq m$ *and* $\operatorname{rk}(\mathbf{A}) = r \leq m$. *Then there are invertible* $n \times n$ *and* $m \times m$ *matrices* $\mathbf{P}$ *and* $\mathbf{Q}$, *such that*

$$\mathbf{PAQ} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \qquad \textit{and} \qquad \mathbf{A} = \mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1}\,. \tag{27}$$

By use of Lemma 3.4 one can show how solutions $\mathbf{A}^-$ of (26) can be found.

- Let $\mathbf{P}$ and $\mathbf{Q}$ be matrices with the properties considered in Lemma 3.4 and let $\mathbf{B}$ be an arbitrary $m \times n$ matrix with

  $$\mathbf{B} = \mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{pmatrix}\mathbf{P}\,, \tag{28}$$

  where $\mathbf{R}$, $\mathbf{S}$ and $\mathbf{T}$ are arbitrary matrices with dimensions $r \times (n-r)$, $(m-r) \times r$ and $(m-r) \times (n-r)$, respectively.

- Then (27) and (28) imply that

$$\begin{aligned}\mathbf{ABA} &= \mathbf{P}^{-1}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}\mathbf{Q}\begin{pmatrix}\mathbf{I}_r & \mathbf{R}\\ \mathbf{S} & \mathbf{T}\end{pmatrix}\mathbf{P}\mathbf{P}^{-1}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}\\ &= \mathbf{P}^{-1}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{I}_r & \mathbf{R}\\ \mathbf{S} & \mathbf{T}\end{pmatrix}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}\\ &= \mathbf{A}\,,\end{aligned}$$

  i.e., the matrix $\mathbf{B}$ given in (28) is a generalized inverse of $\mathbf{A}$.

- Let $k \in \{r, \ldots, m\}$ be an arbitrary natural number. Let

$$\mathbf{R} = \mathbf{0}\,, \qquad \mathbf{S} = \mathbf{0} \qquad \text{and} \qquad \mathbf{T} = \begin{pmatrix}\mathbf{I}_{k-r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\,, \tag{29}$$

  then it is $\mathrm{rk}(\mathbf{B}) = k$.

Altogether, we obtain the following result.

Lemma 3.5 *Let $\mathbf{A}$ be an $n \times m$ matrix with $n \geq m$ and $\mathrm{rk}(\mathbf{A}) = r \leq m$. Let furthermore $\mathbf{B}$ be the $m \times n$ matrix given in (28) – (29), for each $k \in \{r, \ldots, m\}$. Then it holds that $\mathrm{rk}(\mathbf{B}) = k$ and $\mathbf{A}^- = \mathbf{B}$ is a solution of (26).*

Moreover, the following properties of the generalized inverse are useful.

Lemma 3.6

- *Let $\mathbf{A}$ be an arbitrary $n \times m$ matrix with $n \geq m$ and let $\big(\mathbf{A}^\top\mathbf{A}\big)^-$ be a generalized inverse of the symmetric $m \times m$ matrix $\mathbf{A}^\top\mathbf{A}$.*
- *Then the transposed matrix $\big(\big(\mathbf{A}^\top\mathbf{A}\big)^-\big)^\top$ is a generalized inverse of $\mathbf{A}^\top\mathbf{A}$ as well.*
- *Besides, it holds that*

$$\mathbf{A}^\top\mathbf{A}\big(\mathbf{A}^\top\mathbf{A}\big)^-\mathbf{A}^\top = \mathbf{A}^\top\,. \tag{30}$$

Proof

- By definition of the generalized inverse, we have $\mathbf{A}^\top\mathbf{A}\big(\mathbf{A}^\top\mathbf{A}\big)^-\mathbf{A}^\top\mathbf{A} = \mathbf{A}^\top\mathbf{A}$.
- From this equation and from the symmetry of the matrix $\mathbf{A}^\top\mathbf{A}$ it follows that

$$\mathbf{A}^\top\mathbf{A} \;=\; \big(\mathbf{A}^\top\mathbf{A}\big)^\top \;=\; \Big(\mathbf{A}^\top\mathbf{A}\big(\mathbf{A}^\top\mathbf{A}\big)^-\mathbf{A}^\top\mathbf{A}\Big)^\top \;=\; \mathbf{A}^\top\mathbf{A}\Big(\big(\mathbf{A}^\top\mathbf{A}\big)^-\Big)^\top\mathbf{A}^\top\mathbf{A}\,,$$

  i.e., the transposed matrix $\big(\big(\mathbf{A}^\top\mathbf{A}\big)^-\big)^\top$ is a generalized inverse of $\mathbf{A}^\top\mathbf{A}$ as well.
- In order to prove (30), the second part of the statement, we consider the matrix

$$\mathbf{B} = \mathbf{A}^\top\mathbf{A}\big(\mathbf{A}^\top\mathbf{A}\big)^-\mathbf{A}^\top - \mathbf{A}^\top\,.$$

- Then it holds that

$$\begin{aligned}\mathbf{B}\mathbf{B}^\top &= \Big(\mathbf{A}^\top\mathbf{A}(\mathbf{A}^\top\mathbf{A})^-\mathbf{A}^\top - \mathbf{A}^\top\Big)\Big(\mathbf{A}^\top\mathbf{A}(\mathbf{A}^\top\mathbf{A})^-\mathbf{A}^\top - \mathbf{A}^\top\Big)^\top \\ &= \mathbf{A}^\top\mathbf{A}(\mathbf{A}^\top\mathbf{A})^-\mathbf{A}^\top\mathbf{A}\Big((\mathbf{A}^\top\mathbf{A})^-\Big)^\top\mathbf{A}^\top\mathbf{A} \\ &\quad -\mathbf{A}^\top\mathbf{A}(\mathbf{A}^\top\mathbf{A})^-\mathbf{A}^\top\mathbf{A} - \mathbf{A}^\top\mathbf{A}\Big((\mathbf{A}^\top\mathbf{A})^-\Big)^\top\mathbf{A}^\top\mathbf{A} + \mathbf{A}^\top\mathbf{A} \\ &= \mathbf{A}^\top\mathbf{A} - \mathbf{A}^\top\mathbf{A} - \mathbf{A}^\top\mathbf{A} + \mathbf{A}^\top\mathbf{A} = \mathbf{0}\,.\end{aligned}$$

- Therefore, we get that $\mathbf{B} = \mathbf{0}$. □

By use of the generalized inverse $(\mathbf{X}^\top\mathbf{X})^-$ of $\mathbf{X}^\top\mathbf{X}$ and its properties (considered in Lemma 3.6), the solution set of the normal equation (25) can be specified.

Theorem 3.5 *The general solution $\boldsymbol{\beta}$ of the normal equation $\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{Y}$ has the form*

$$\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)\mathbf{z}\,, \tag{31}$$

where $(\mathbf{X}^\top\mathbf{X})^-$ is an arbitrary solution of

$$\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{X}^\top\mathbf{X} \tag{32}$$

and $\mathbf{z} \in \mathbb{R}^m$ is an arbitrary m–dimensional vector.

Proof

- By plugging (31) into the left–hand side of the normal equation (25), one sees
  - that for each $\mathbf{z} \in \mathbb{R}^m$ equation (31) gives a solution of (25),
  - as it holds that

$$\begin{aligned}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{X}\Big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)\mathbf{z}\Big) \\ &= \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \;=\; \mathbf{X}^\top\mathbf{Y}\,,\end{aligned}$$

    where the last equality follows from Lemma 3.6.
- Let now $\widetilde{\boldsymbol{\beta}}$ be an arbitrary solution and $\boldsymbol{\beta}$ be a solution of the form (31) of equation (25).
  - Then subtraction on each side of (25) yields

$$\mathbf{X}^\top\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{o}\,. \tag{33}$$

  - Hence, for a $\mathbf{z} \in \mathbb{R}^m$ it holds that

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= \boldsymbol{\beta} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \overset{(31)}{=} (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)\mathbf{z} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \\ &= (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)(\mathbf{z} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})) + (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\overset{(33)}{=} (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)(\mathbf{z} - (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}))\,.\end{aligned}$$

  - This means that $\widetilde{\boldsymbol{\beta}}$ is a solution of the form (31) of equation (25) as well. □

Example (one–factor analysis of variance)

- *Recall*: In the reparametrized model of one–factor analysis of variance (cf. case 2 of the example considered in Section 3.1.2) the design matrix is given by the $n \times (k+1)$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{34}$$

  and the parameter vector $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_k)^\top$.
- One can easily see that in this case

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & n_1 & n_2 & n_3 & \dots & n_{k-1} & n_k \\ n_1 & n_1 & 0 & 0 & \dots & 0 & 0 \\ n_2 & 0 & n_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ n_k & 0 & 0 & 0 & \dots & 0 & n_k \end{pmatrix} \tag{35}$$

  and that a generalized inverse of $\mathbf{X}^\top \mathbf{X}$ is given by

$$(\mathbf{X}^\top \mathbf{X})^- = \begin{pmatrix} \frac{1}{n} & 0 & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{n} & \frac{1}{n_1} & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{n} & 0 & \frac{1}{n_2} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\frac{1}{n} & 0 & 0 & 0 & \dots & 0 & \frac{1}{n_k} \end{pmatrix}. \tag{36}$$

- Therefore, the normal equation (25), i.e., $\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}$, has the following form:

$$n\mu + \sum_{i=1}^{k} n_i \alpha_i \;=\; Y_{\cdot\cdot}\,, \qquad n_i \mu + n_i \alpha_i \;=\; Y_{i\cdot}\,, \qquad \forall\, i = 1, \dots, k\,.$$

- If we only consider solutions of this system of equations which are in the *restricted parameter space* $\Theta \subset \mathbb{R}^{k+1}$, where

$$\Theta = \left\{ \boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_k) : \sum_{i=1}^{k} n_i \alpha_i = 0 \right\}, \tag{37}$$

we obtain the (uniquely determined) solution $\widehat{\boldsymbol{\beta}} = (\widehat{\mu}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_k)$ with

$$\widehat{\mu} \;=\; \overline{Y}_{\cdot\cdot}\,, \qquad \widehat{\alpha}_i \;=\; \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\,, \quad \forall\, i = 1, \ldots, k\,. \tag{38}$$

- One can easily see that the solution $\widehat{\boldsymbol{\beta}}$ of the normal equation (25) which is given in (38)
  - has the form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$, where the generalized inverse $(\mathbf{X}^\top\mathbf{X})^-$ is given by (36), and
  - is an unbiased estimator for $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)$ with respect to the restricted parameter space Θ, which has the form given in (37).
- Without the additional condition considered in (37), there is no LS–estimator which is unbiased at the same time, cf. Theorem 3.8.

Now we consider the linear model with general design matrix $\mathbf{X}$ again, which is given in (1) – (3). In particular, we consider the solutions of the normal equation (25) discussed in Theorem 3.5 and show that the mean squared error $e(\boldsymbol{\beta})$ given in (2.8) is minimized for $\mathbf{z} = \mathbf{o}$.

Theorem 3.6 *Let $(\mathbf{X}^\top\mathbf{X})^-$ be a generalized inverse of $\mathbf{X}^\top\mathbf{X}$. Then the sample function*

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \tag{39}$$

minimizes the mean squared error $e(\boldsymbol{\beta})$, i.e., $\overline{\boldsymbol{\beta}}$ is an LS–estimator for $\boldsymbol{\beta}$.

Proof

- For each m–dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ it holds that

$$\begin{aligned} n\,e(\boldsymbol{\beta}) \;&=\; (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \;=\; \big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}} + \mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\big)^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}} + \mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\big) \\ &=\; \big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\big)^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\big) + (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\geq\; \big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\big)^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\big) = n\,e\big(\overline{\boldsymbol{\beta}}\big) \end{aligned}$$

- because

$$(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \Big(\mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\Big)^\top\Big(\mathbf{X}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\Big) \geq 0$$

and

$$(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{X}^\top\big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}\big) = (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\big(\mathbf{X}^\top - \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\mathbf{Y} = 0\,,$$

where the last equality follows from Lemma 3.6. □

3.2.2 Expectation Vector and Covariance Matrix of the LS–Estimator $\overline{\boldsymbol{\beta}}$

The model assumptions (3) for $\varepsilon_1, \ldots, \varepsilon_n$ and the general calculation rules for the expectation and covariance of real–valued random variables imply that the expectation vector and the covariance matrix of the LS–estimator $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ have the following form.

Theorem 3.7 *It holds that*

$$\mathbb{E}\,\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \tag{40}$$

and

$$\mathrm{Cov}\,\overline{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top. \tag{41}$$

Proof

- It follows from $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{o}$ that

$$\mathbb{E}\,\overline{\boldsymbol{\beta}} = \mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}\right) = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbb{E}\,\mathbf{Y} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\,.$$

- Moreover, it holds that

$$\begin{aligned}
\operatorname{Cov}\left(\overline{\beta}_i,\, \overline{\beta}_j\right) &= \operatorname{Cov}\Big(\sum_{\ell=1}^{n}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{i\ell}Y_\ell,\, \sum_{r=1}^{n}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{jr}Y_r\Big) \\
&= \sum_{\ell=1}^{n}\sum_{r=1}^{n}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{i\ell}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{jr}\operatorname{Cov}\left(Y_\ell,\, Y_r\right) \\
&= \sigma^2\sum_{\ell=1}^{n}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{i\ell}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{j\ell} \\
&= \sigma^2\sum_{\ell=1}^{n}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)_{i\ell}\big(\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\big)_{\ell j} \\
&= \sigma^2\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\big)_{ij}
\end{aligned}$$

  for arbitrary $i, j \in \{1, \ldots, m\}$. □

Together with Lemma 3.3, Theorems 3.5 and 3.7 imply that there is *no* LS–estimator for $\boldsymbol{\beta}$ which is additionally unbiased. In particular, the LS–estimator $\overline{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ given in (39) is biased.

Theorem 3.8 *If* $\operatorname{rk}(\mathbf{X}) < m$*, there is no unbiased LS–estimator for* $\boldsymbol{\beta}$*.*

Proof

- Due to $\operatorname{rk}(\mathbf{X}) < m$, it follows from Lemma 3.3 that $\operatorname{rk}(\mathbf{X}^\top\mathbf{X}) < m$ and

$$\operatorname{rk}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big) < m\,.$$

  – Hence, there is a $\boldsymbol{\beta} \neq \mathbf{o}$ with $(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{o}$, i.e., the equation

$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{42}$$

    does *not* hold for each $\boldsymbol{\beta} \in \mathbb{R}^m$.
  – So because of (40), the LS–estimator $\overline{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ given in (39) is biased.
- As (42) does not hold for each $\boldsymbol{\beta} \in \mathbb{R}^m$, one additionally obtains that for each arbitrary but fixed $\mathbf{z} \in \mathbb{R}^m$ the equation

$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta} - \mathbf{z}) = \boldsymbol{\beta} - \mathbf{z}$$

  or equivalently

$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \big(\mathbf{I}_m - (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)\mathbf{z} = \boldsymbol{\beta}$$

  does not hold for each $\boldsymbol{\beta} \in \mathbb{R}^m$.
- Due to Theorem 3.5, this means that there is no LS–estimator for $\boldsymbol{\beta}$ which is additionally unbiased. □

3.2.3 Estimable Functions

- In Section 3.2.2 we have shown that if the design matrix $\mathbf{X}$ does not have full rank, there is no unbiased LS–estimator for $\boldsymbol{\beta}$ in the linear model without additional conditions.
- Hence, instead of the vector $\boldsymbol{\beta}$, one considers a class of (real–valued) linear functions $\mathbf{a}^\top\boldsymbol{\beta}$ of the parameter vector $\boldsymbol{\beta}$, for which unbiased LS–estimators can be constructed.
- In other words: Instead of the (vectorial) linear transformation $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ of the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ one considers a class of (real–valued) linear functions $\mathbf{c}^\top\mathbf{Y}$ of $\mathbf{Y}$, which can be perceived as estimators for $\mathbf{a}^\top\boldsymbol{\beta}$.
- This leads to the following conception.

Definition

- Let $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$ be an arbitrary m–dimensional vector.
- The linear function $\mathbf{a}^\top\boldsymbol{\beta}$ of the parameter vector $\boldsymbol{\beta}$ is called *estimable without bias* or an *estimable function* if there is an n–dimensional vector $\mathbf{c} = (c_1, \ldots, c_n)^\top$ such that

$$\mathbb{E}\left(\mathbf{c}^\top\mathbf{Y}\right) = \mathbf{a}^\top\boldsymbol{\beta}\,, \qquad \forall\,\boldsymbol{\beta} \in \mathbb{R}^m\,. \tag{43}$$

Example (one–factor analysis of variance)

- For the reparametrized model of one–factor analysis of variance with parameter vector $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top \in \mathbb{R}^{k+1}$ one can show that for example $\alpha_1 - \alpha_2$ is an estimable function as defined by (43).
- This is true because for

$$\mathbf{a}^\top = (0, 1, -1, 0, \ldots, 0) \qquad \text{and} \qquad \mathbf{c}^\top = (\underbrace{0, \ldots, 0}_{n_1 - 1}, 1, -1, 0, \ldots, 0)$$

  it holds that

$$\mathbb{E}\left(\mathbf{c}^\top\mathbf{Y}\right) = \mathbb{E}\left(Y_{1n_1} - Y_{21}\right) = (\mu + \alpha_1) - (\mu + \alpha_2) = \alpha_1 - \alpha_2 = \mathbf{a}^\top\boldsymbol{\beta}$$

  for each $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_k)^\top \in \mathbb{R}^{k+1}$.
- In a similar way it can be shown that $\mu + \alpha_i$ and $\alpha_i - \alpha_{i'}$ are estimable functions of $\boldsymbol{\beta}$ for $i = 1, \ldots, k$ and $i, i' = 1, \ldots, k$ with $i \neq i'$, respectively.

Example (two–factor analysis of variance with balanced subsamples)

- For the model of two–factor analysis of variance with balanced subsamples, introduced in Section 3.1.3, the normal equation (25) has the following form:

$$\begin{aligned}
rk_1k_2\mu + rk_2\sum_{i_1=1}^{k_1}\alpha_{i_1}^{(1)} + rk_1\sum_{i_2=1}^{k_2}\alpha_{i_2}^{(2)} + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\alpha_{i_1i_2} &= Y_{\cdots} \\
rk_2\mu + rk_2\alpha_{i_1}^{(1)} + r\sum_{i_2=1}^{k_2}\alpha_{i_2}^{(2)} + r\sum_{i_2=1}^{k_2}\alpha_{i_1i_2} &= Y_{i_1\cdot\cdot} \qquad \forall\, i_1 = 1, \ldots, k_1 \\
rk_1\mu + r\sum_{i_1=1}^{k_1}\alpha_{i_1}^{(1)} + rk_1\alpha_{i_2}^{(2)} + r\sum_{i_1=1}^{k_1}\alpha_{i_1i_2} &= Y_{\cdot i_2\cdot} \qquad \forall\, i_2 = 1, \ldots, k_2 \\
r\mu + r\alpha_{i_1}^{(1)} + r\alpha_{i_2}^{(2)} + r\alpha_{i_1i_2} &= Y_{i_1i_2\cdot} \qquad \forall\, i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2
\end{aligned}$$

- In consideration of the additional condition (18), this system of equations can be solved uniquely. In other words: If only parameter vectors

$$\boldsymbol{\beta} = \left(\mu, \alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_{k_2}^{(2)}, \alpha_{11}, \ldots, \alpha_{k_1 k_2}\right)^\top$$

from the restricted parameter space

$$\Theta = \left\{\boldsymbol{\beta} : \sum_{i_1=1}^{k_1} \alpha_{i_1}^{(1)} = \sum_{i_2=1}^{k_2} \alpha_{i_2}^{(2)} = \sum_{i_1=1}^{k_1} \alpha_{i_1 i_2} = \sum_{i_2=1}^{k_2} \alpha_{i_1 i_2} = 0\right\}$$

are considered, one obtains the unique solution

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\mu}, \widehat{\alpha}_1^{(1)}, \ldots, \widehat{\alpha}_{k_1}^{(1)}, \widehat{\alpha}_1^{(2)}, \ldots, \widehat{\alpha}_{k_2}^{(2)}, \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{k_1 k_2}\right)^\top \tag{44}$$

of the normal equation, where

$$\widehat{\mu} = \overline{Y}_{\ldots}, \quad \widehat{\alpha}_{i_1}^{(1)} = \overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\ldots}, \quad \widehat{\alpha}_{i_2}^{(2)} = \overline{Y}_{\cdot i_2 \cdot} - \overline{Y}_{\ldots}, \quad \widehat{\alpha}_{i_1 i_2} = \overline{Y}_{\ldots} + \overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\cdot i_2 \cdot} \tag{45}$$

for arbitrary $i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2$.

- It can be shown that the solution $\widehat{\boldsymbol{\beta}}$ of the normal equation given in (44) – (45) has the form

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

where $(\mathbf{X}^\top \mathbf{X})^-$ is a generalized inverse of $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}$ is the design matrix of the model of two–factor analysis of variance with balanced subsamples.

- *Remark*: The sample function $\widehat{\boldsymbol{\beta}}$ given in (44) – (45) was already discussed in Theorem 3.3, where we have shown that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$ with respect to the parameter space Θ.
- Furthermore, one can show that the linear functions $\mu + \alpha_{i_1}^{(1)} + \alpha_{i_2}^{(2)} + \alpha_{i_1 i_2}$ of the parameter vector $\boldsymbol{\beta}$ are estimable without bias as defined in (43) (without taking into account the additional conditions (18)) for arbitrary $i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2$.
- In the model of two–factor analysis of variance without interactions, i.e., $\alpha_{i_1 i_2} = 0$ for arbitrary $i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2$, also the linear functions $\alpha_{i_1}^{(1)} - \alpha_{i_1'}^{(1)}$ and $\alpha_{i_2}^{(2)} - \alpha_{i_2'}^{(2)}$ are estimable without bias for arbitrary $i_1, i_1' = 1, \ldots, k_1$ with $i_1 \neq i_1'$ and for arbitrary $i_2, i_2' = 1, \ldots, k_2$ with $i_2 \neq i_2'$, respectively.

The following lemma, which is an extension of Lemma 3.6, is needed to derive two general *criteria for linear functions* $\mathbf{a}^\top \boldsymbol{\beta}$ *of the parameter vector* $\boldsymbol{\beta}$ *to be estimable without bias.*

Lemma 3.7 *Let* $\left(\mathbf{X}^\top \mathbf{X}\right)^-$ *be a generalized inverse of* $\mathbf{X}^\top \mathbf{X}$. *Then it holds that*

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{X}. \tag{46}$$

Proof In Lemma 3.6 we have shown that

- the transposed matrix $\left(\left(\mathbf{X}^\top \mathbf{X}\right)^-\right)^\top$ is a generalized inverse of $\mathbf{X}^\top \mathbf{X}$ as well and that $\mathbf{X}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^- \mathbf{X}^\top = \mathbf{X}^\top$.
- Hence, it is $\mathbf{X}^\top \mathbf{X}\left(\left(\mathbf{X}^\top \mathbf{X}\right)^-\right)^\top \mathbf{X}^\top = \mathbf{X}^\top$.
- This leads to (46) by swapping columns and rows. □

Theorem 3.9 *Let* $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$ *be an arbitrary vector. The linear function* $\mathbf{a}^\top \boldsymbol{\beta}$ *of the parameter vector* $\boldsymbol{\beta}$ *is estimable without bias if and only if one of the following conditions is fulfilled:*

1. *There is a* $\mathbf{c} \in \mathbb{R}^n$*, such that*
$$\mathbf{a}^\top = \mathbf{c}^\top \mathbf{X}\,. \tag{47}$$
2. *The vector* $\mathbf{a}$ *fulfills the following system of equations:*
$$\mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{a}^\top\,. \tag{48}$$

Proof

- Let $\mathbf{a}^\top \boldsymbol{\beta}$ be an estimable function of the parameter vector $\boldsymbol{\beta}$.
  - Then it follows from (43) that
$$\mathbf{a}^\top \boldsymbol{\beta} = \mathbb{E}\left(\mathbf{c}^\top \mathbf{Y}\right) = \mathbf{c}^\top \mathbb{E}\,\mathbf{Y} = \mathbf{c}^\top \mathbf{X}\boldsymbol{\beta} \qquad \text{and hence} \qquad \left(\mathbf{c}^\top \mathbf{X} - \mathbf{a}^\top\right)\boldsymbol{\beta} = 0$$
    for each $\boldsymbol{\beta} \in \mathbb{R}^m$.
  - Therefore, we get that $\mathbf{a}^\top = \mathbf{c}^\top \mathbf{X}$.
- Conversely, let $\mathbf{c}$ be a vector satisfying condition (47).
  - Then
$$\mathbb{E}\left(\mathbf{c}^\top \mathbf{Y}\right) = \mathbf{c}^\top \mathbb{E}\,\mathbf{Y} = \mathbf{c}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{a}^\top \boldsymbol{\beta}\,, \qquad \forall\, \boldsymbol{\beta} \in \mathbb{R}^m\,.$$
  - Thus, also the sufficiency of condition (47) is proved.
- In order to show the necessity of condition (48), we use the result of Lemma 3.7.
  - Let $\mathbf{a}^\top \boldsymbol{\beta}$ be an estimable function of the parameter vector $\boldsymbol{\beta}$.
  - Then it follows from (47) and (46) that
$$\mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{c}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{c}^\top \mathbf{X} = \mathbf{a}^\top\,.$$
- In order to show the sufficiency of condition (48), one only has to observe
  - that (48) implies $\mathbf{a}^\top = \mathbf{c}^\top \mathbf{X}$ for $\mathbf{c}^\top = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top$.
  - Now the first part of the statement implies that $\mathbf{a}^\top \boldsymbol{\beta}$ is estimable without bias. □

Remark

- If the design matrix $\mathbf{X}$ has full rank, i.e., $(\mathbf{X}^\top \mathbf{X})^- = (\mathbf{X}^\top \mathbf{X})^{-1}$, then condition (48) is obviously fulfilled for each $\mathbf{a} \in \mathbb{R}^m$.
- In this case, *every* linear function of the parameter vector $\boldsymbol{\beta}$ is estimable without bias, which has already been shown in Theorem 2.2.

In the case that the design matrix $\mathbf{X} = (x_{ij})$ does not have full rank, we show

- how the second part of the statement in Theorem 3.9 implies that the following linear functions $\mathbf{a}^\top \boldsymbol{\beta}$ of $\boldsymbol{\beta}$ are estimable without bias.
- In doing so, the vector (of weights) $\mathbf{c}$ of the linear unbiased estimator $\mathbf{c}^\top \mathbf{Y}$ for $\mathbf{a}^\top \boldsymbol{\beta}$ can be chosen as in the proof of Theorem 3.9, i.e.,
$$\mathbf{c}^\top = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top\,. \tag{49}$$

Theorem 3.10 *The following linear functions of the parameter vector* $\boldsymbol{\beta}$ *are estimable without bias:*

1. *the components* $\sum_{j=1}^m x_{1j}\beta_j, \ldots, \sum_{j=1}^m x_{nj}\beta_j$ *of the expectation vector* $\mathbb{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$*,*
2. *each linear function of estimable functions,*

3. *the components* $\beta'_1, \ldots, \beta'_m$ *of the so–called projected parameter vector* $\boldsymbol{\beta}' = (\beta'_1, \ldots, \beta'_m)^\top$, *where*

$$\boldsymbol{\beta}' = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \,. \tag{50}$$

Proof

- Let $\mathbf{a}_i = (x_{i1}, \ldots, x_{im})^\top$ for each $i \in \{1, \ldots, n\}$. Then it holds that

$$\begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{X} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{X} = \begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_n^\top \end{pmatrix},$$

  where the last but one equality follows from Lemma 3.7. Now Theorem 3.9 implies that every linear combination $\sum_{j=1}^m x_{1j}\beta_j, \ldots, \sum_{j=1}^m x_{nj}\beta_j$ is an estimable function.

- In order to prove the second part of the statement, we consider a (finite) family $\mathbf{a}_1^\top \boldsymbol{\beta}, \ldots, \mathbf{a}_s^\top \boldsymbol{\beta}$ of s estimable functions, which we write in the form $\mathbf{A}\boldsymbol{\beta}$, where $\mathbf{A} = \big(\mathbf{a}_1, \ldots, \mathbf{a}_s\big)^\top$ is an $s \times m$–dimensional matrix and $s \in \mathbb{N}$ is an arbitrary natural number.

- Theorem 3.9 implies that

$$\mathbf{A}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{A} \,.$$

  – Thus, for each s–dimensional vector $\mathbf{b} = (b_1, \ldots, b_s)^\top \in \mathbb{R}^s$ it holds that

$$\mathbf{b}^\top \mathbf{A}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{b}^\top \mathbf{A} \,.$$

  – Therefore, it follows by use of Theorem 3.9 that the linear function $\mathbf{b}^\top \mathbf{A}\boldsymbol{\beta}$ of the estimable functions $\mathbf{A}\boldsymbol{\beta}$ is an estimable function itself.

- In the third part of the statement the family $\mathbf{A}\boldsymbol{\beta}$ of linear functions of the parameter vector $\boldsymbol{\beta}$ is considered, where the $m \times m$ matrix $\mathbf{A}$ is given by $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X}$.

  – Hence,

$$\mathbf{A}(\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} = \mathbf{A} \,,$$

    where the last but one equality follows from the definition of the generalized inverse in (26).

  – Now Theorem 3.9 implies that the components $\beta'_1, \ldots, \beta'_m$ of the projected parameter vector

$$\boldsymbol{\beta}' = \mathbf{A}\boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^- \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}$$

    are estimable functions. □

3.2.4 Best Linear Unbiased Estimator; Gauss–Markov Theorem

- In this section we show how BLUE–estimators for estimable functions of the parameter vector $\boldsymbol{\beta}$ can be constructed.

- *Recall*: A linear unbiased estimator is called BLUE–estimator if there is no linear unbiased estimator with lower variance (BLUE = best linear unbiased estimator).

- In the theory of linear models, the following result is called *Gauss–Markov theorem*.

Theorem 3.11

- *Let $\mathbf{a}^\top\boldsymbol{\beta}$ be an estimable function of the parameter vector $\boldsymbol{\beta}$, let $(\mathbf{X}^\top\mathbf{X})^-$ be an arbitrary generalized inverse of the $m \times m$ matrix $\mathbf{X}^\top\mathbf{X}$ and let $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$.*
- *Then $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ is a BLUE–estimator for $\mathbf{a}^\top\boldsymbol{\beta}$, where*

$$\operatorname{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) = \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a}\,. \tag{51}$$

Proof

- First, we show that $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ is a linear unbiased estimator for $\mathbf{a}^\top\boldsymbol{\beta}$.
  - It is clear that
  $$\mathbf{a}^\top\overline{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$$
  is a linear function of the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$.
  - As we assume that $\mathbf{a}^\top\boldsymbol{\beta}$ is an estimable function of the parameter vector $\boldsymbol{\beta}$, Theorem 3.9 implies that there is a $\mathbf{c} \in \mathbb{R}^n$ such that
  $$\mathbf{a}^\top = \mathbf{c}^\top\mathbf{X}\,. \tag{52}$$
  - Hence, it holds that
  $$\mathbb{E}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) = \mathbf{c}^\top\mathbf{X}\mathbb{E}\,\overline{\boldsymbol{\beta}} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbb{E}\,\mathbf{Y} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{c}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^\top\boldsymbol{\beta}$$
  for each $\boldsymbol{\beta} \in \mathbb{R}^m$, where the last but one equality follows from Lemma 3.7.
  - Therefore, we have shown that $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ is a linear unbiased estimator for $\mathbf{a}^\top\boldsymbol{\beta}$.
- It follows from the calculation rules for the variance of random variables (cf. Theorem WR-4.13) that
$$\operatorname{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) = \operatorname{Var}\Big(\sum_{i=1}^m a_i\overline{\beta}_i\Big) = \sum_{i=1}^m\sum_{j=1}^m a_i a_j\,\operatorname{Cov}\left(\overline{\beta}_i, \overline{\beta}_j\right).$$
  - Moreover, we have shown in Theorem 3.7 that
  $$\operatorname{Cov}\left(\overline{\beta}_i, \overline{\beta}_j\right) = \sigma^2\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\big)_{ij}\,.$$
  - This leads to
  $$\begin{aligned}\operatorname{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) &= \sigma^2\sum_{i=1}^m\sum_{j=1}^m a_i a_j\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\big)_{ij}\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\mathbf{a}\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\mathbf{X}^\top\mathbf{c}\\ &= \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{c}\,,\end{aligned}$$
  where the last but one equality follows from (52) and the last equality follows from Lemma 3.6.
  - Applying (52) again yields the variance formula (51).
- It remains to show that the estimator $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ has minimum variance in the class of all linear unbiased estimators for $\mathbf{a}^\top\boldsymbol{\beta}$.
  - Let $\mathbf{b} \in \mathbb{R}^n$, such that $\mathbf{b}^\top\mathbf{Y}$ is a linear unbiased estimator for $\mathbf{a}^\top\boldsymbol{\beta}$. Then it holds that
  $$\mathbf{a}^\top\boldsymbol{\beta} = \mathbb{E}\left(\mathbf{b}^\top\mathbf{Y}\right) = \mathbf{b}^\top\mathbf{X}\boldsymbol{\beta}, \qquad \forall\,\boldsymbol{\beta} \in \mathbb{R}^m$$
  and thus
  $$\mathbf{b}^\top\mathbf{X} = \mathbf{a}^\top\,. \tag{53}$$

– For the covariance of $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ and $\mathbf{b}^\top\mathbf{Y}$ it holds that

$$\mathrm{Cov}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}},\, \mathbf{b}^\top\mathbf{Y}\right) = \mathrm{Cov}\left(\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y},\, \mathbf{b}^\top\mathbf{Y}\right) = \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{b} = \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a}\,,$$

where the last equality follows from (53).

– This result and the variance formula (51) imply that

$$\begin{aligned} 0 \;\leq\; \mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}} - \mathbf{b}^\top\mathbf{Y}\right) \;&=\; \mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right) + \mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right) - 2\,\mathrm{Cov}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}},\, \mathbf{b}^\top\mathbf{Y}\right) \\ &=\; \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a} + \mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right) - 2\,\sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a} \\ &=\; \mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right) - \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{a} \;=\; \mathrm{Var}\left(\mathbf{b}^\top\mathbf{Y}\right) - \mathrm{Var}\left(\mathbf{a}^\top\overline{\boldsymbol{\beta}}\right). \end{aligned}$$

□

Remark

- In the proof of Theorem 3.11 it has never explicitely been used that $\mathrm{rk}(\mathbf{X}) < m$.
- In other words: If the design matrix $\mathbf{X}$ has full rank, i.e., $\mathrm{rk}(\mathbf{X}) = m$, then $\mathbf{a}^\top\boldsymbol{\beta}$ is estimable without bias for *each* m–dimensional vector $\mathbf{a}^\top \in \mathbb{R}^m$ and $\mathbf{a}^\top\widehat{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is a BLUE–estimator for $\mathbf{a}^\top\boldsymbol{\beta}$.

The following *invariance property* of the generalized inverse $(\mathbf{X}^\top\mathbf{X})^-$ of $\mathbf{X}^\top\mathbf{X}$ implies that the BLUE–estimator $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$, considered in Theorem 3.11, does *not* depend on the specific choice of $(\mathbf{X}^\top\mathbf{X})^-$.

Lemma 3.8 *Let* $\mathbf{A}$ *and* $\mathbf{A}'$ *be arbitrary generalized inverses of the matrix* $\mathbf{X}^\top\mathbf{X}$. *Then it holds that*

$$\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}\mathbf{A}'\mathbf{X}^\top\,. \tag{54}$$

Proof

- In Lemma 3.7 we have shown that

$$\mathbf{X}\mathbf{A}\mathbf{X}^\top\mathbf{X} = \mathbf{X} = \mathbf{X}\mathbf{A}'\mathbf{X}^\top\mathbf{X} \tag{55}$$

  for arbitrary generalized inverses $\mathbf{A}$ and $\mathbf{A}'$ of the matrix $\mathbf{X}^\top\mathbf{X}$.
- If this chain of equations is multiplied by $\mathbf{A}\mathbf{X}^\top$ from the right, one obtains

$$\mathbf{X}\mathbf{A}\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}\mathbf{A}'\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top\,. \tag{56}$$

- formula (55) implies for the left–hand side of the last equality that $\mathbf{X}\mathbf{A}\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}\mathbf{A}\mathbf{X}^\top$.
- Furthermore, we have shown in Lemma 3.6 that $\mathbf{X}^\top\mathbf{X}\mathbf{A}\mathbf{X}^\top = \mathbf{X}^\top$ for each generalized inverse $\mathbf{A}$ of $\mathbf{X}^\top\mathbf{X}$.
- If this is plugged into the right–hand side of (56), one obtains (54). □

By using Lemma 3.8, we can now prove the invariance property of the BLUE–estimator $\mathbf{a}^\top\overline{\boldsymbol{\beta}}$ considered in Theorem 3.11 which was already mentioned above.

Theorem 3.12 *Let* $\mathbf{a}^\top\boldsymbol{\beta}$ *be an estimable function of the parameter vector* $\boldsymbol{\beta}$. *Then the BLUE–estimator* $\mathbf{a}^\top\overline{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ *does not depend on the choice of the generalized inverse* $(\mathbf{X}^\top\mathbf{X})^-$.

Proof

- Recall: It follows from Theorem 3.9 that for each estimable function $\mathbf{a}^\top\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ there is a $\mathbf{c} \in \mathbb{R}^n$, such that $\mathbf{a}^\top = \mathbf{c}^\top\mathbf{X}$.
- Together with Lemma 3.8 this implies that

$$\mathbf{a}^\top\overline{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \mathbf{c}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$$

  does not depend on the choice of the generalized inverse $(\mathbf{X}^\top\mathbf{X})^-$. □

Example (one–factor analysis of variance)

- For the reparametrized model of one–factor analysis of variance $\mu + \alpha_i$ and $\alpha_i - \alpha_{i'}$ are estimable functions of $\boldsymbol{\beta}$ for $i = 1, \ldots, k$ and $i, i' = 1, \ldots, k$, respectively, cf. Theorem 3.10.
- Theorem 3.11 implies that $\widehat{\mu}+\widehat{\alpha}_i$ and $\widehat{\alpha}_i-\widehat{\alpha}_{i'}$ are BLUE–estimators for $\mu+\alpha_i$ and $\alpha_i-\alpha_{i'}$, respectively, where

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = (\widehat{\mu}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_k)^\top$$

  with $\widehat{\mu} = \overline{Y}_{\cdot\cdot}$ and $\widehat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ is the solution of the normal equation (25), which was already considered in Section 3.2.1.

Example (two–factor analysis of variance with balanced subsamples)

- For the model of two–factor analysis of variance with balanced subsamples introduced in Section 3.1.3 the linear functions $\mu + \alpha^{(1)}_{i_1} + \alpha^{(2)}_{i_2} + \alpha_{i_1 i_2}$ of the parameter vector

$$\boldsymbol{\beta} = \left(\mu, \alpha^{(1)}_1, \ldots, \alpha^{(1)}_{k_1}, \alpha^{(2)}_1, \ldots, \alpha^{(2)}_{k_2}, \alpha_{11}, \ldots, \alpha_{k_1 k_2}\right)$$

  are estimable without bias for arbitrary $i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2$, cf. Theorem 3.10.
- Theorem 3.11 implies that $\widehat{\mu}+\widehat{\alpha}^{(1)}_{i_1}+\widehat{\alpha}^{(2)}_{i_2}+\widehat{\alpha}_{i_1 i_2}$ is a BLUE–estimator for $\mu+\alpha^{(1)}_{i_1}+\alpha^{(2)}_{i_2}+\alpha_{i_1 i_2}$, where

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \left(\widehat{\mu}, \widehat{\alpha}^{(1)}_1, \ldots, \widehat{\alpha}^{(1)}_{k_1}, \widehat{\alpha}^{(2)}_1, \ldots, \widehat{\alpha}^{(2)}_{k_2}, \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{k_1 k_2}\right)$$

  with

$$\widehat{\mu} = \overline{Y}_{\cdots}, \qquad \widehat{\alpha}^{(1)}_{i_1} = \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots}, \qquad \widehat{\alpha}^{(2)}_{i_2} = \overline{Y}_{\cdot i_2\cdot} - \overline{Y}_{\cdots}, \qquad \widehat{\alpha}_{i_1 i_2} = \overline{Y}_{\cdots} + \overline{Y}_{i_1 i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot i_2\cdot}$$

  is the solution of the normal equation (25) already considered in Section 3.2.3.
- Furthermore, it follows from Theorem 3.10 that in the model of two–factor analysis of variance without interactions, i.e., $\alpha_{i_1 i_2} = 0$ for arbitrary $i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2$, also $\alpha^{(1)}_{i_1} - \alpha^{(1)}_{i'_1}$ and $\alpha^{(2)}_{i_2} - \alpha^{(2)}_{i'_2}$ are estimable without bias for arbitrary $i_1, i'_1 = 1, \ldots, k_1$ with $i_1 \neq i'_1$ and $i_2, i'_2 = 1, \ldots, k_2$ with $i_2 \neq i'_2$, respectively.
- Therefore, Theorem 3.11 implies that $\widehat{\alpha}^{(1)}_{i_1} - \widehat{\alpha}^{(1)}_{i'_1}$ and $\widehat{\alpha}^{(2)}_{i_2} - \widehat{\alpha}^{(2)}_{i'_2}$ are BLUE–estimators for $\alpha^{(1)}_{i_1} - \alpha^{(1)}_{i'_1}$ and $\alpha^{(2)}_{i_2} - \alpha^{(2)}_{i'_2}$, respectively.

3.3 Normally Distributed Error Terms

In addition to the model assumptions made at the beginning of Chapter 3, we now assume again that $n > m$ and that the random error terms $\varepsilon_1, \ldots, \varepsilon_n : \Omega \to \mathbb{R}$ are independent and (identically) normally distributed, i.e., in particular, that $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ for each $i = 1, \ldots, n$.

3.3.1 Maximum–Likelihood Estimation

- In the same way as in the case of a design matrix with full column rank, which has been discussed in Section 2.2.1, Theorem 1.3 implies that the vector $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ of response variables is normally distributed with
$$\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I})\,. \tag{57}$$
- In other words: The distribution of the random vector $\mathbf{Y}$ is absolutely continuous with density
$$f_{\mathbf{Y}}(\mathbf{y}) = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\Big) \tag{58}$$
for each $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$.
- Therefore, the loglikelihood function $\log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \log f_{\mathbf{Y}}(\mathbf{y})$ has the form
$$\log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\,\frac{n}{2}\,\log(2\pi) - \frac{n}{2}\,\log(\sigma^2) - \frac{1}{2\sigma^2}\,|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|^2\,. \tag{59}$$
- In order to specify a maximum–likelihood estimator for the parameter vector $(\boldsymbol{\beta}, \sigma^2)$, we first consider the mapping
$$\mathbb{R}^m \ni \boldsymbol{\beta} \mapsto \log L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) \tag{60}$$
for arbitrary but fixed $\mathbf{y} \in \mathbb{R}^n$ and $\sigma^2 > 0$ (just as in the proof of Theorem 2.6).
- It follows from (59) and (60) that the following expression $e(\boldsymbol{\beta})$ has to be minimized, where
$$e(\boldsymbol{\beta}) = \frac{1}{n}\,|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|^2 = \frac{1}{n}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,.$$
- Hence, Theorem 3.6 implies that the LS–estimator $\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y}$ simultaneously is an ML–estimator for $\boldsymbol{\beta}$ (which does not depend on σ^2).
- Moreover, one obtains just as in the proof of Theorem 2.6 that an ML–estimator for $(\boldsymbol{\beta}, \sigma^2)$ is given by $(\overline{\boldsymbol{\beta}}, \overline{\sigma}^2)$, where
$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \qquad \text{and} \qquad \overline{\sigma}^2 = \frac{1}{n}\,\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)^\top\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)\,. \tag{61}$$

Remark

- In Section 3.2.2 we have shown that in general $\overline{\boldsymbol{\beta}}$ is not an unbiased estimator for $\boldsymbol{\beta}$.
- Similarly, $\overline{\sigma}^2$ is not an unbiased estimator for σ^2; however, an unbiased estimator for σ^2 can be derived by a simple modification of $\overline{\sigma}^2$.
- In order to show this statement, the following properties of the matrix
$$\mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top \tag{62}$$
are useful, which can be perceived as a generalization of the corresponding matrix properties derived in Lemmas 2.1 and 2.2 for the case of a design matrix $\mathbf{X}$ with full column rank.

Lemma 3.9 *Let* $\mathrm{rk}(\mathbf{X}) = r \le m$. *Then for the matrix* $\mathbf{G}$ *given in* (62) *it holds that*

1) $\mathbf{G}$ *is idempotent and symmetric,*

2) $\mathbf{GX} = \mathbf{0}$ *and* 3) $\mathrm{tr}(\mathbf{G}) = \mathrm{rk}(\mathbf{G}) = n - r$.

Proof

- Lemma 3.6 implies that

$$\begin{aligned}\mathbf{G}^2 &= \big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\\ &= \mathbf{I}-2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top+\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top}_{=\mathbf{X}^\top}\\ &= \mathbf{I}-2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top+\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top=\mathbf{G}\,.\end{aligned}$$

- In Lemma 3.6 we have shown that $\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top$ is a generalized inverse of $\mathbf{X}^\top\mathbf{X}$. Hence, it follows from Lemma 3.8 that

$$\mathbf{G}^\top=\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)^\top=\mathbf{I}-\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\mathbf{X}^\top=\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top=\mathbf{G}\,.$$

- Thus, the first part of the statement is proved. In order to prove the second part of the statement, it suffices to observe that

$$\mathbf{G}\mathbf{X}=\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\mathbf{X}=\mathbf{X}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}=\mathbf{X}-\mathbf{X}=\mathbf{0}\,,$$

  where the last but one equality follows from Lemma 3.7.
- The third part of the statement can be proved as follows:
  - Lemmas 3.3 and 3.7 imply that

$$r=\operatorname{rk}(\mathbf{X})=\operatorname{rk}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big)\le\operatorname{rk}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\le\operatorname{rk}(\mathbf{X})=r\,,$$

    i.e.,

$$\operatorname{rk}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)=r\,. \tag{63}$$

  - It follows from Lemma 3.6 that the matrix $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ is idempotent because it holds that

$$\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)=\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top}_{=\mathbf{X}^\top}=\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\,.$$

  - Since additionally $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ is symmetric, Lemma 1.3 together with (63) implies that

$$\begin{aligned}\operatorname{tr}(\mathbf{G}) &= \operatorname{tr}\big(\mathbf{I}_n-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\\ &= \operatorname{tr}(\mathbf{I}_n)-\operatorname{tr}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\\ &= n-\operatorname{rk}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)=n-r\,.\end{aligned}$$

□

By use of Lemma 3.9 one now obtains a formula for the expectation of the ML–estimator $\overline{\sigma}^2$ considered in (61).

Theorem 3.13 *It holds that*

$$\mathbb{E}\,\overline{\sigma}^2=\frac{n-r}{n}\,\sigma^2\,. \tag{64}$$

Proof Due to the properties of the matrix $\mathbf{G}=\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ derived in Lemma 3.9, it holds for the ML–estimator $\overline{\sigma}^2=\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)^\top\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)/n$ considered in (61) that

$$\begin{aligned}\mathbb{E}\,\overline{\sigma}^2 &= \frac{1}{n}\,\mathbb{E}\left(\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)^\top\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)\right)\\ &= \frac{1}{n}\,\mathbb{E}\left(\mathbf{Y}^\top\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)^\top\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\mathbf{Y}\right)\\ &= \frac{1}{n}\,\mathbb{E}\left(\mathbf{Y}^\top\mathbf{G}^\top\mathbf{G}\mathbf{Y}\right)=\frac{1}{n}\,\mathbb{E}\left(|\mathbf{G}\mathbf{Y}|^2\right)=\frac{1}{n}\,\mathbb{E}\left(|\mathbf{G}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|^2\right)\\ &= \frac{1}{n}\,\mathbb{E}\left(|\mathbf{G}\boldsymbol{\varepsilon}|^2\right)=\frac{1}{n}\,\mathbb{E}\left(\boldsymbol{\varepsilon}^\top\mathbf{G}^\top\mathbf{G}\boldsymbol{\varepsilon}\right)=\frac{1}{n}\,\mathbb{E}\left(\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\right)\\ &= \frac{\sigma^2}{n}\,\operatorname{tr}(\mathbf{G})=\frac{n-r}{n}\,\sigma^2\,.\end{aligned}$$

□

Remark By using the notation

$$S^2 = \frac{1}{n-r}\,\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big)^\top\big(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}\big) \qquad \text{and} \qquad S^2 = \frac{1}{n-r}\,\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}, \tag{65}$$

Theorem 3.13 implies that $\mathbb{E}\,S^2 = \sigma^2$, i.e., S^2 is an unbiased estimator for σ^2.

For being able to specify the distributions of the estimators $\overline{\boldsymbol{\beta}}$ and S^2, we need the notion of the degenerate multivariate normal distribution, cf. Section 1.2.5.

Theorem 3.14 *Let* $\mathrm{rk}(\mathbf{X}) = r \le m$. *Then it holds that*

$$\overline{\boldsymbol{\beta}} \sim \mathrm{N}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\big) \tag{66}$$

and

$$\frac{(n-r)S^2}{\sigma^2} \sim \chi^2_{n-r}\,, \tag{67}$$

where the random variables $\overline{\boldsymbol{\beta}}$ *and* S^2 *are independent.*

Proof

- For the estimator $\overline{\boldsymbol{\beta}}$ given in (61) it holds that

$$\overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\big(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\big) = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\varepsilon}\,,$$

  where

$$\boldsymbol{\mu} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\,, \qquad \mathbf{B} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\,, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o},\, \sigma^2\mathbf{I}_n)\,.$$

- Now the definition of the (degenerate) multivariate normal distribution implies that $\overline{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\mu},\, \mathbf{K})$, where

$$\mathbf{K} = \sigma^2\mathbf{B}\mathbf{B}^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top\,.$$

- Thus, (66) is proved. In order to prove (67), we use the identity derived in the proof of Theorem 3.13, i.e.,

$$S^2 = \frac{1}{n-r}\,\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}\,, \qquad \text{where } \mathbf{G} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top. \tag{68}$$

- As $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o},\, \sigma^2\mathbf{I}_n)$ and as we have shown in Lemma 3.9 that
  - the matrix $\mathbf{G}$ is idempotent and symmetric
  - with $\mathrm{rk}(\mathbf{G}) = n-r$,

  it follows from Theorem 1.9 that the quadratic form $(n-r)S^2/\sigma^2$ has a (central) χ^2–distribution with $n-r$ degrees of freedom, i.e., $(n-r)S^2/\sigma^2 \sim \chi^2_{n-r}$.
- Since every idempotent and symmetric matrix simultaneously is positive semidefinite and since

$$\mathbf{B}\mathbf{G} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big(\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big) = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top - (\mathbf{X}^\top\mathbf{X})^-\underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top}_{=\mathbf{X}^\top} = \mathbf{0}$$

  due to Lemma 3.6, it follows from Theorem 1.10 that the random variables $\mathbf{B}\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^\top\mathbf{G}\boldsymbol{\varepsilon}$ are independent. Thus, also the random variables $\overline{\boldsymbol{\beta}}$ and S^2 are independent. □

3.3.2 Testing Linear Hypotheses

In this section we discuss a generalized version of the test for linear forms of $\boldsymbol{\beta}$ which we considered in Section 2.2.3 for the case of a design matrix $\mathbf{X}$ with full column rank, cf. Theorem 2.11. However, we now assume that $\mathrm{rk}(\mathbf{X}) = r < m$.

- Let $s \in \{1, \ldots, m\}$, let $\mathbf{H}$ be an $s \times m$ matrix with full rank $\mathrm{rk}(\mathbf{H}) = s$ and let $\mathbf{d} \in \mathbb{R}^s$.
- The hypothesis to be tested is

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d} \qquad \text{versus} \qquad H_1 : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{d}\,, \tag{69}$$

  where we assume that the entries of the matrix $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_s)^\top$ and the components of the vector $\mathbf{d} = (d_1, \ldots, d_s)^\top$ are known.
- In order to verify the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ considered in (69), we construct a test statistic whose distribution does *not* depend on the unknown parameter vector $(\boldsymbol{\beta}, \sigma^2)$ (in a similar way as in Theorem 2.11).
- For this purpose, we introduce the following term for assuming that the components of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable without bias.

Definition The hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is called *testable* if all components $\mathbf{h}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{h}_s^\top\boldsymbol{\beta}$ of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable functions of the parameter vector $\boldsymbol{\beta}$.

Remark Theorem 3.9 implies that the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is testable if and only if

- there is an $s \times n$ matrix $\mathbf{C}$, such that

$$\mathbf{H} = \mathbf{C}\mathbf{X}\,, \tag{70}$$

or

- the matrix $\mathbf{H}$ fulfills the following equation:

$$\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X} = \mathbf{H}\,. \tag{71}$$

In order to construct a test statistic for the verification of the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ considered in (69), the following lemma is useful.

Lemma 3.10

- *Let $s \leq m$, let $\mathbf{H}$ be an $s \times m$ matrix with full rank $\mathrm{rk}(\mathbf{H}) = s$ which fulfills (70) or (71) and let $(\mathbf{X}^\top\mathbf{X})^-$ be an arbitrary generalized inverse of $\mathbf{X}^\top\mathbf{X}$.*
- *Then the $s \times s$ matrix $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top$ is positive definite (and thus invertible).*

Proof

- One can show that the symmetric $m \times m$ matrix $\mathbf{X}^\top\mathbf{X}$ with $\mathrm{rk}(\mathbf{X}^\top\mathbf{X}) = r < m$ can be represented in the form:

$$\mathbf{X}^\top\mathbf{X} = \mathbf{P}^{-1}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^{-1}\,,$$

  where the $m \times m$ matrix $\mathbf{P}$ is invertible and symmetric; cf. also Lemma 3.4.

- In the proof of Lemma 3.5 we have shown that in this case

$$\mathbf{P}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-r} \end{pmatrix}\mathbf{P} = \mathbf{PP}$$

  defines a generalized inverse of $\mathbf{X}^\top\mathbf{X}$, which obviously is positive definite.
- Now it follows from Lemma 1.8 that also the matrix $\mathbf{HPPH}^\top$ is positive definite.
- This implies that $\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top$ is positive definite for any arbitrary generalized inverse $(\mathbf{X}^\top\mathbf{X})^-$ of $\mathbf{X}^\top\mathbf{X}$ as it follows from (70) that

$$\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top = \mathbf{CX}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{C}^\top = \mathbf{CXPPX}^\top\mathbf{C}^\top = \mathbf{HPPH}^\top,$$

  where the second equality holds due to Lemma 3.8. □

Remark

- Lemma 3.10 implies that the test statistic $T_{\mathbf{H}}$ with

$$T_{\mathbf{H}} = \frac{\big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)}{sS^2} \tag{72}$$

  is well–defined, where $\overline{\boldsymbol{\beta}}$ and S^2 are the estimators for $\boldsymbol{\beta}$ and σ^2 which are given in (61) and (65), respectively.
- This test statistic is a generalization of the corresponding test statistic $T_{\mathbf{H}}$ considered in Section 2.2.3 for a design matrix $\mathbf{X}$ with full rank. The distribution of the test statistic $T_{\mathbf{H}}$ given in (72) can be specified as follows.

Theorem 3.15 *Let the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ be testable. Then, assuming that $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is true, it holds that $T_{\mathbf{H}} \sim \mathrm{F}_{s,n-r}$, i.e., the test statistic $T_{\mathbf{H}}$ given in (72) has an F–distribution with $(s, n-r)$ degrees of freedom.*

Proof Assuming that $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is true, the following statements hold.

- The definition of $\overline{\boldsymbol{\beta}}$ in (61) implies that

$$\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} - \mathbf{d} = \mathbf{H}\big((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)\big(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\big) - \mathbf{d} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\varepsilon},$$

  where

$$\boldsymbol{\mu} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{d} = \mathbf{C}\underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}}_{=\mathbf{X}}\boldsymbol{\beta} - \mathbf{d} = \mathbf{H}\boldsymbol{\beta} - \mathbf{d} = \mathbf{o},$$

  $\mathbf{B} = \mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I}_n)$.
- Hence, or the numerator $Z = \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)$ of the test statistic $T_{\mathbf{H}}$ given in (72) it holds that

$$\begin{aligned} Z &= \boldsymbol{\varepsilon}^\top\mathbf{B}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\mathbf{B}\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\big)^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}^\top\mathbf{A}\boldsymbol{\varepsilon}, \end{aligned}$$

  – where the matrix $\mathbf{A} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top$ is idempotent as due to (71) it holds that

$$\begin{aligned} \mathbf{A}^2 &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1} \\ &\quad\times \underbrace{\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{X}}_{=\mathbf{H}}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top \\ &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top = \mathbf{A}. \end{aligned}$$

  - Since $\mathbf{A}$ is also symmetric with $\mathrm{rk}(\mathbf{A}) = s$, Theorem 1.9 implies that the quadratic form Z/σ^2 has a χ^2–distribution with s degrees of freedom.
- Furthermore, it has been shown in Theorem 3.14 that $(n-r)S^2/\sigma^2 \sim \chi^2_{n-r}$ and that the random variables $\overline{\boldsymbol{\beta}}$ and S^2 are independent.
- Therefore, the random variables Z and S^2 are also independent and it holds that

$$T_{\mathbf{H}} = \frac{Z/s\sigma^2}{S^2/\sigma^2} \sim \mathrm{F}_{s,n-r}\,.$$

□

Remark

- The choice of the test statistic $T_{\mathbf{H}}$ in (72) can be motivated in the following way: Theorem 1.9 implies (in a similar way as in the proof of Theorem 3.15) that the quadratic form Z/σ^2 with

$$Z = \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)$$

  in general (i.e., without assuming that $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is true) has a noncentral χ^2–distribution $\chi^2_{s,\lambda}$ with

$$\lambda = \frac{\big(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\big)}{\sigma^2}\,.$$

- This implies that

$$\mathbb{E}\Big(\frac{Z}{\sigma^2}\Big) = \frac{d}{dt}\,\mathbb{E}\,\exp\Big(\frac{tZ}{\sigma^2}\Big)\Big|_{\,t=0} = s + \lambda\,,$$

  where the last equality follows from the formula for the moment generating function of the $\chi^2_{s,\lambda}$–distribution which has been derived in Theorem 1.8.
  - In other words: It holds that

$$\mathbb{E}\Big(\frac{Z}{s}\Big) = \sigma^2 + \frac{\big(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\big)}{s} \tag{73}$$

    and Theorem 3.13 implies that $\mathbb{E}\,(S^2) = \sigma^2$.
  - Hence, assuming that the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is true, the expectations of numerator and denominator of the test statistic $T_{\mathbf{H}}$ are equal.
- On the other hand, Lemmas 1.8 and 3.10 imply that the inverse matrix $\big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}$ is positive definite and thus,

$$\big(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\big)^{-1}\big(\mathbf{H}\boldsymbol{\beta} - \mathbf{d}\big) > 0$$

  if the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ is false. In this case, it follows from (73) that

$$\mathbb{E}\Big(\frac{Z}{s}\Big) > \sigma^2 = \mathbb{E}\,(S^2)\,. \tag{74}$$

- In general, we have $\mathbb{E}\,T_{\mathbf{H}} = \mathbb{E}\,\big(Z/s\big)\mathbb{E}\,(1/S^2)$ (due to the independence of Z and S^2) and the Jensen inequality implies that $\mathbb{E}\,(1/S^2) > 1/\mathbb{E}\,(S^2)$.
- So (74) implies that

$$\mathbb{E}\,T_{\mathbf{H}} > \frac{\mathbb{E}\,\Big(\frac{Z}{s}\Big)}{\mathbb{E}\,(S^2)} > 1$$

  in the case that H_0 is false.
- Therefore, it is reasonable to reject the null hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}$ if the test statistic $T_{\mathbf{H}}$ takes values which are significantly larger than 1.

- Thus, due to the distributional property of the test statistic $T_{\mathbf{H}}$ which has been derived in Theorem 3.15, the null hypothesis H_0 is rejected if $T_{\mathbf{H}} > \mathrm{F}_{s,n-r,1-\alpha}$.

In some cases it is more convenient to consider an *alternative representation* of the test statistic $T_{\mathbf{H}}$ given in (72). For this purpose, we define the following **s***ums of* **s***quared* **e***rrors* SSE and SSE_H by

$$SSE = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}), \qquad \text{where} \qquad \overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^- \mathbf{X}^\top \mathbf{Y}, \tag{75}$$

and

$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H), \qquad \text{where} \qquad \overline{\boldsymbol{\beta}}_H = \overline{\boldsymbol{\beta}} - (\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top\big)^{-1} \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big). \tag{76}$$

Theorem 3.16 *For the test statistic $T_{\mathbf{H}}$ given in (72) it holds that*

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/s}{SSE/(n-r)}. \tag{77}$$

Proof

- It holds that

$$\begin{aligned} SSE_H &= (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) \\ &= \big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}} + \mathbf{X}(\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}_H)\big)^\top \big(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}} + \mathbf{X}(\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}_H)\big) \\ &= (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}) + (\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}_H)^\top \mathbf{X}^\top \mathbf{X} (\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}_H) \end{aligned}$$

  because parts 1 and 2 of the statement in Lemma 3.9 imply that

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}) = \mathbf{X}^\top \big(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^- \mathbf{X}^\top\big)\mathbf{Y} = \underbrace{\mathbf{X}^\top \mathbf{G}}_{=\mathbf{0}} \mathbf{Y} = \mathbf{o}.$$

- Furthermore, it follows from (70), i.e., $\mathbf{H} = \mathbf{C}\mathbf{X}$, and from Lemmas 3.6 and 3.7 that

$$\begin{aligned} SSE_H &= (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}) \\ &\quad + \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top\big)^{-1} \underbrace{\mathbf{H}\big((\mathbf{X}^\top\mathbf{X})^-\big)^\top \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top}_{=\mathbf{H}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top} \\ &\qquad \times \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top\big)^{-1} \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big) \\ &= SSE + \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top\big)^{-1} \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big). \end{aligned}$$

□

Remark

- From the definition of $\overline{\boldsymbol{\beta}}_H$ in (76) it follows that

$$\mathbf{H}\overline{\boldsymbol{\beta}}_H = \mathbf{H}\Big(\overline{\boldsymbol{\beta}} - (\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top \big(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^- \mathbf{H}^\top\big)^{-1} \big(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\big)\Big) = \mathbf{d},$$

  i.e., the random vector $\overline{\boldsymbol{\beta}}_H$ given in (76) only takes values in the *restricted parameter space* $\Theta_H = \big\{\boldsymbol{\beta} \in \mathbb{R}^m : \mathbf{H}\boldsymbol{\beta} = \mathbf{d}\big\}$.

- Furthermore, it can easily be shown that $\overline{\boldsymbol{\beta}}_H$ minimizes the mean squared error $e(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in \Theta_H$, where

$$e(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im})\big)^2.$$

3.3.3 Confidence Regions

- For the construction of confidence regions we proceed in a similar way as in Section 2.2.4, where we considered the case that the design matrix $\mathbf{X}$ has full rank, i.e., $\mathrm{rk}(\mathbf{X}) = m$. However, in doing so we now assume that $\mathrm{rk}(\mathbf{X}) = r < m$ as we did in Section 3.3.2.
- Let $s \in \{1, \ldots, m\}$ and let $\mathbf{H}$ be an $s \times m$ matrix with full rank $\mathrm{rk}(\mathbf{H}) = s$ whose entries are known, where $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_s)^\top$.
- Then Theorem 3.15 immediately leads to the following *confidence region for the vector* $\mathbf{H}\boldsymbol{\beta}$ with confidence level $1 - \alpha \in (0, 1)$.

Theorem 3.17 *Let all components $\mathbf{h}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{h}_s^\top\boldsymbol{\beta}$ of the vector $\mathbf{H}\boldsymbol{\beta}$ be estimable functions of $\boldsymbol{\beta}$. Then the (random) ellipsoid*

$$E = \left\{ \mathbf{d} \in \mathbb{R}^s : \frac{\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\right)^\top \left(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\right)^{-1} \left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{d}\right)}{sS^2} \leq \mathrm{F}_{s,n-r,1-\alpha} \right\} \tag{78}$$

is a confidence region for $\mathbf{H}\boldsymbol{\beta}$ with confidence level $1 - \alpha \in (0, 1)$, where $\overline{\boldsymbol{\beta}}$ and S^2 are the estimators for $\boldsymbol{\beta}$ and σ^2 given in (61) *and* (65)*, respectively.*

In particular, Theorem 3.17 implies the following result.

Corollary 3.1 *For each $i \in \{1, \ldots, s\}$ the (random) interval*

$$(\underline{\theta}, \overline{\theta}) = \left(\mathbf{h}_i^\top\overline{\boldsymbol{\beta}} - \mathrm{t}_{n-r,1-\alpha/2} S \sqrt{\mathbf{h}_i^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}_i}\ ,\ \mathbf{h}_i^\top\overline{\boldsymbol{\beta}} + \mathrm{t}_{n-r,1-\alpha/2} S \sqrt{\mathbf{h}_i^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}_i} \right) \tag{79}$$

is a confidence interval for $\mathbf{h}_i^\top\boldsymbol{\beta}$ with confidence level $1 - \alpha \in (0, 1)$.

Example

- We consider the following linear model, cf. N. Ravishanker und D.K. Dey (2002) *A First Course in Linear Model Theory*, Chapman & Hall/CRC, S. 235:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix},$$

  where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3)^\top \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I})$.
- By means of Corollary 3.1, a confidence interval for $\beta_1 + \beta_2/3 + 2\beta_3/3$ with confidence level $1 - \alpha = 0.95$ shall be specified.
- As $\mathrm{rk}(\mathbf{X}) = 2 < m = 3$, we first need to check whether the function

$$\mathbf{h}^\top\boldsymbol{\beta} = \beta_1 + \beta_2/3 + 2\beta_3/3 \tag{80}$$

  of $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \beta_3)$ with $\mathbf{h}^\top = (1, 1/3, 2/3)$ is estimable without bias.
  - Due to criterion 1 in Theorem 3.9, this is the case if and only if there is a $\mathbf{c}^\top = (c_1, c_2, c_3) \in \mathbb{R}^3$, such that $\mathbf{h}^\top = \mathbf{c}^\top\mathbf{X}$, i.e., if

$$\begin{aligned} 1 &= c_1 + c_2 + c_3 \\ 1/3 &= c_1 + c_3 \\ 2/3 &= c_2\,. \end{aligned}$$

– Since this system of equations obviously is solvable, $\mathbf{h}^\top\boldsymbol{\beta}$ is estimable without bias.

- Moreover, it holds that

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and a generalized inverse of $\mathbf{X}^\top\mathbf{X}$ is given by

$$(\mathbf{X}^\top\mathbf{X})^- = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 3/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

- This yields that $\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h} = 1/2$. Thus,

$$(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & -1 & 1/2 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \overline{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} Y_2 \\ (Y_1 - 2Y_2 + Y_3)/2 \\ 0 \end{pmatrix}.$$

- Hence, one obtains that $(\mathbf{X}\overline{\boldsymbol{\beta}})^\top = \big((Y_1+Y_3)/2, Y_2, (Y_1+Y_3)/2\big)$ and

$$\mathbf{h}^\top\overline{\boldsymbol{\beta}} = (Y_1 + 4Y_2 + Y_3)/6 \qquad \text{and} \qquad S^2 = (Y_1 - Y_3)^2/2\,.$$

- Therefore, a confidence interval $(\underline{\theta}, \overline{\theta})$ for $\mathbf{h}^\top\boldsymbol{\beta}$ with confidence level $1-\alpha = 0.95$ is obtained, which has the form

$$(\underline{\theta}, \overline{\theta}) = \Big((Y_1 + 4Y_2 + Y_3)/6 - Z\,,\, (Y_1 + 4Y_2 + Y_3)/6 + Z\Big),$$

where $Z = t_{1,0.975}|Y_1 - Y_3|/2$.

By generalizing Theorem 2.12, we now derive a so–called *Scheffé confidence band*, i.e., simultaneous confidence intervals for a whole class of estimable functions of the parameter vector $\boldsymbol{\beta}$.

- Let $s \in \{1, \ldots, m\}$, let $\mathbf{H}$ once more be an $s \times m$ matrix with full rank, i.e., $\mathrm{rk}(\mathbf{H}) = s$, where $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_s)^\top$, and let all components $\mathbf{h}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{h}_s^\top\boldsymbol{\beta}$ of the vector $\mathbf{H}\boldsymbol{\beta}$ be estimable functions of $\boldsymbol{\beta}$.
- As $\mathbf{H}$ has full (row) rank, the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_s$ are linearly independent and form the basis of an s–dimensional linear subspace in $\mathbb{R}^m$, which we denote by $\mathcal{L} = \mathcal{L}(\mathbf{h}_1, \ldots, \mathbf{h}_s)$.
- Due to Theorem 3.10, the function $\mathbf{h}^\top\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ is estimable without bias for each $\mathbf{h} \in \mathcal{L}$.
- We are looking for a number $a_\gamma > 0$, such that

$$\mathbf{h}^\top\overline{\boldsymbol{\beta}} - a_\gamma Z_{\mathbf{h}} \le \mathbf{h}^\top\boldsymbol{\beta} \le \mathbf{h}^\top\overline{\boldsymbol{\beta}} + a_\gamma Z_{\mathbf{h}} \tag{81}$$

holds for each $\mathbf{h} \in \mathcal{L}$ simultaneously with the (given) probability $\gamma \in (0,1)$, where $Z_{\mathbf{h}} = S\sqrt{\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}}$ and $\overline{\boldsymbol{\beta}}$, S^2 are the estimators for $\boldsymbol{\beta}$ and σ^2 given in (61) and (65), respectively.

Theorem 3.18 *Let $a_\gamma = \sqrt{s\,\mathrm{F}_{s,n-r,\gamma}}$. Then it holds that*

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{h}\in\mathcal{L}} \frac{\big(\mathbf{h}^\top\overline{\boldsymbol{\beta}} - \mathbf{h}^\top\boldsymbol{\beta}\big)^2}{S^2\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}} \le a_\gamma^2\right) = \gamma\,. \tag{82}$$

Proof

- In a similar way as in the proof of Theorem 2.12, the Cauchy–Schwarz inequality for scalar products, cf. (65), implies that

$$\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\right)^\top \left(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\right)^{-1}\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\right) = \max_{\mathbf{x}\neq\mathbf{o}} \frac{\left(\mathbf{x}^\top\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\right)\right)^2}{\mathbf{x}^\top\left(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\right)\mathbf{x}} ,$$

  where the maximum extends over all vectors $\mathbf{x} \in \mathbb{R}^s$ with $\mathbf{x} \neq \mathbf{o}$.
- By means of Theorem 3.15 we now get that

$$\begin{aligned}
\gamma &= \mathbb{P}_{\boldsymbol{\beta}}\Big(\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\right)^\top \left(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\right)^{-1}\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\right) \leq sS^2\,\mathrm{F}_{s,n-r,\gamma}\Big) \\
&= \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{x}\neq\mathbf{o}} \frac{\left(\mathbf{x}^\top\left(\mathbf{H}\overline{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}\right)\right)^2}{\mathbf{x}^\top\left(\mathbf{H}(\mathbf{X}^\top\mathbf{X})^-\mathbf{H}^\top\right)\mathbf{x}} \leq sS^2\,\mathrm{F}_{s,n-r,\gamma}\right) \\
&= \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{x}\neq\mathbf{o}} \frac{\left((\mathbf{H}^\top\mathbf{x})^\top\overline{\boldsymbol{\beta}} - (\mathbf{H}^\top\mathbf{x})^\top\boldsymbol{\beta}\right)^2}{(\mathbf{H}^\top\mathbf{x})^\top(\mathbf{X}^\top\mathbf{X})^-(\mathbf{H}^\top\mathbf{x})} \leq sS^2\,\mathrm{F}_{s,n-r,\gamma}\right) \\
&= \mathbb{P}_{\boldsymbol{\beta}}\left(\max_{\mathbf{h}\in\mathcal{L}} \frac{\left(\mathbf{h}^\top\overline{\boldsymbol{\beta}} - \mathbf{h}^\top\boldsymbol{\beta}\right)^2}{\mathbf{h}^\top(\mathbf{X}^\top\mathbf{X})^-\mathbf{h}} \leq sS^2\,\mathrm{F}_{s,n-r,\gamma}\right).
\end{aligned}$$

□

3.4 Examples

3.4.1 F–Test for the ANOVA Null Hypothesis

- We consider the reparametrized model of *one–factor analysis of variance*, i.e., the design matrix $\mathbf{X}$ is the $n \times (k+1)$ matrix given in (13) with $\mathrm{rk}(\mathbf{X}) = k < m = k+1$, where

$$\mathbf{X} = \begin{pmatrix}
1 & 1 & 0 & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 1 & 0 & 0 & \dots & 0 & 0 \\
1 & 0 & 1 & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 0 & 1 & 0 & \dots & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
1 & 0 & 0 & 0 & \dots & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 0 & 0 & 0 & \dots & 0 & 1
\end{pmatrix} \tag{83}$$

  and the parameter vector $\boldsymbol{\beta}$ has the form $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_k)^\top$.
- It shall be tested whether the levels of the predictor variable are significant, i.e., the hypothesis to be tested is the ANOVA null hypothesis $H_0 : \alpha_1 = \dots = \alpha_k$ (against the alternative $H_1 : \alpha_i \neq \alpha_j$ for some pair $i, j \in \{1, \dots, k\}$ with $i \neq j$). For this purpose, we use the general testing approach introduced in Theorems 3.15 and 3.16.

- An equivalent formulation of the null hypothesis $H_0 : \alpha_1 = \ldots = \alpha_k$ is given by

$$H_0 : \alpha_1 - \alpha_2 = 0, \ldots, \alpha_1 - \alpha_k = 0 \qquad \text{or} \qquad H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}\,, \tag{84}$$

  where $\mathbf{H}$ is a $(k-1) \times (k+1)$ matrix with

$$\mathbf{H} = \begin{pmatrix} 0 & 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \ldots & -1 & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & -1 \end{pmatrix}. \tag{85}$$

- Obviously, $\mathbf{H}$ is a matrix with full row rank, i.e., $\mathrm{rk}(\mathbf{H}) = k-1$. Furthermore, Theorem 3.10 implies that all components $\alpha_1 - \alpha_2, \ldots, \alpha_1 - \alpha_k$ of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable functions of $\boldsymbol{\beta}$.

- In other words, the matrix $\mathbf{H}$ fulfills the requirements of Theorems 3.15 and 3.16. Thus, the test statistic

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k-1)}{SSE/(n-k)}$$

  considered in Theorem 3.16 may be used for the verification of the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$, where the sums of squares SSE and SSE_H defined in (75) and (76), respectively, can be determined as follows.

- *Recall*: In Section 3.2.1 we have shown that a generalized inverse of $\mathbf{X}^\top\mathbf{X}$ is given by (36), i.e.,

$$\big(\mathbf{X}^\top\mathbf{X}\big)^- = \begin{pmatrix} \frac{1}{n} & 0 & 0 & 0 & \ldots & 0 & 0 \\ -\frac{1}{n} & \frac{1}{n_1} & 0 & 0 & \ldots & 0 & 0 \\ -\frac{1}{n} & 0 & \frac{1}{n_2} & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -\frac{1}{n} & 0 & 0 & 0 & \ldots & 0 & \frac{1}{n_k} \end{pmatrix}. \tag{86}$$

- Together with (83) this implies that

$$\overline{\boldsymbol{\beta}} = \big(\mathbf{X}^\top\mathbf{X}\big)^-\mathbf{X}^\top\mathbf{Y} = \big(\overline{Y}_{\cdot\cdot}, \overline{Y}_{1\cdot} - \overline{Y}_{\cdot\cdot}, \ldots, \overline{Y}_{k\cdot} - \overline{Y}_{\cdot\cdot}\big)^\top$$

  and

$$\mathbf{X}\overline{\boldsymbol{\beta}} = \mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^-\mathbf{X}^\top\mathbf{Y} = \Big(\underbrace{\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{1\cdot}}_{n_1}, \ldots, \underbrace{\overline{Y}_{k\cdot}, \ldots, \overline{Y}_{k\cdot}}_{n_k}\Big)^\top.$$

- Hence, one gets for the sum of squares $SSE = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}})$ that

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(Y_{ij} - \overline{Y}_{i\cdot}\Big)^2. \tag{87}$$

Remark

- Due to the special form (83) of the design matrix $\mathbf{X}$, formula (87) can also be derived directly from the fact that $\overline{\boldsymbol{\beta}}$ is an LS–estimator. Indeed, it holds that

$$\begin{aligned} SSE &= (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}})^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{k}\Big\{\min_{x\in\mathbb{R}}\sum_{j=1}^{n_i}\big(Y_{ij}-x\big)^2\Big\} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(Y_{ij}-\overline{Y}_{i\cdot}\Big)^2 . \end{aligned}$$

- Moreover, it follows from the remark at the end of Section 3.3.2 that

$$SSE_H = (\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y}-\mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta}\in\Theta_H}\big|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\big|^2 ,$$

  where $\Theta_H = \{\boldsymbol{\beta}\in\mathbb{R}^{k+1} : \mathbf{H}\boldsymbol{\beta}=\mathbf{o}\}$ and $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta}\in\Theta_H\}\subset\mathbb{R}^n$ is the set of those n–dimensional vectors whose components are all equal.

- Then one gets that

$$SSE_H = \min_{x\in\mathbb{R}}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij}-x\big)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij}-\overline{Y}_{\cdot\cdot}\big)^2 \tag{88}$$

  because the mean $\overline{Y}_{\cdot\cdot}$ minimizes the sum of squares $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij}-x\big)^2$.

- Together with (87) this implies that

$$SSE_H - SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij}-\overline{Y}_{\cdot\cdot}\big)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(Y_{ij}-\overline{Y}_{i\cdot}\Big)^2 = \sum_{i=1}^{k} n_i\big(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\big)^2 ,$$

  where the last equality follows from the decomposition

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij}-\overline{Y}_{\cdot\cdot}\big)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(Y_{ij}-\overline{Y}_{i\cdot}\Big)^2 + \sum_{i=1}^{k} n_i\big(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\big)^2,$$

  cf. formula (9) in Theorem 3.1.

- Therefore, for the test statistic $T_{\mathbf{H}}$ considered in Theorem 3.16 it holds that

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k-1)}{SSE/(n-k)} = \frac{(n-k)\sum_{i=1}^{k} n_i\big(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\big)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(Y_{ij}-\overline{Y}_{i\cdot}\Big)^2} \sim \mathrm{F}_{k-1,n-k}\,. \tag{89}$$

3.4.2 F–Tests for the Two–Factor Analysis of Variance

Now we construct F–tests for the model of *two–factor analysis of variance* with balanced subsets, which has been introduced in Section 3.1.3, i.e.,

- the parameter vector $\boldsymbol{\beta}$ has the form

$$\boldsymbol{\beta} = \big(\mu, \alpha_1^{(1)},\ldots,\alpha_{k_1}^{(1)}, \alpha_1^{(2)},\ldots,\alpha_{k_2}^{(2)}, \alpha_{11},\ldots,\alpha_{k_1k_2}\big)^\top ,$$

- the design matrix $\mathbf{X}$ has the dimension $n\times m$, where $n = rk_1k_2$ and $m = 1+k_1+k_2+k_1k_2$,

- the entries of $\mathbf{X}$ only consist of zeros and ones and it holds that $\mathrm{rk}(\mathbf{X}) = k_1 k_2 < m$.

Significance of the Predictor Variables

We first construct a test to investigate the question whether the levels of the first predictor variable are significant. For this purpose, we verify the hypothesis that the effects

$$\alpha_{i_1}^{(1)*} = \alpha_{i_1}^{(1)} + \frac{1}{k_2} \sum_{i_2=1}^{k_2} \alpha_{i_1 i_2}$$

of the first predictor variable plus their interactions, averaged over all levels of the second predictor variable, are equal. In other words: The hypothesis to be tested is

$$H_0 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} = 0 \quad \forall\, i_1 \in \{1, \ldots, k_1\} \quad \text{vs.} \quad H_1 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} \neq 0 \quad \text{for some } i_1 \in \{1, \ldots, k_1\}, \tag{90}$$

where it is actually sufficient to consider the pair of hypotheses

$$H_0 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} = 0 \quad \forall\, i_1 \in \{2, \ldots, k_1\} \quad \text{vs.} \quad H_1 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} \neq 0 \quad \text{for some } i_1 \in \{2, \ldots, k_1\}.$$

- It can easily be shown that the null hypothesis in (90) has the form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$,
  - where

$$\mathbf{H} = \begin{pmatrix} 0 & 1 & -1 & 0 & \ldots & 0 & 0 & \ldots & 0 & \frac{1}{k_2} & \ldots & \frac{1}{k_2} & \frac{-1}{k_2} & \ldots & \frac{-1}{k_2} & & 0 & \ldots & 0 \\ 0 & 1 & 0 & -1 & & 0 & 0 & \ldots & 0 & \frac{1}{k_2} & \ldots & \frac{1}{k_2} & 0 & \ldots & 0 & & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \ddots & & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & 1 & 0 & & & -1 & 0 & \ldots & 0 & \frac{1}{k_2} & \ldots & \frac{1}{k_2} & 0 & \ldots & 0 & & \frac{-1}{k_2} & \ldots & \frac{-1}{k_2} \end{pmatrix}$$

    is a $(k_1 - 1) \times m$ matrix with full row rank $\mathrm{rk}(\mathbf{H}) = k_1 - 1$ and with blocks of rows of the lengths 1, 1, $k_1 - 1$, k_2, k_2 and $(k_1 - 1)k_2$, respectively,
  - and where all components of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable functions of $\boldsymbol{\beta}$ because of

$$\alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} = \frac{1}{k_2} \sum_{i_2=1}^{k_2} \left(\theta_{1 i_2} - \theta_{i_1 i_2} \right).$$

- In order to verify the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ we can now again use the test statistic $T_{\mathbf{H}}$ considered in Theorem 3.16, where

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1 - 1)}{SSE/(k_1 k_2 (r-1))} \sim \mathrm{F}_{k_1 - 1,\, k_1 k_2 (r-1)}$$

  with

$$SSE = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} \left(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot} \right)^2 \tag{91}$$

  and

$$SSE_H - SSE = r k_2 \sum_{i_1=1}^{k_1} \left(\overline{Y}_{i_1 \ldots} - \overline{Y}_{\ldots} \right)^2. \tag{92}$$

- The formulas (91) and (92) for the sums of squares SSE and SSE_H defined in (75) and (76), respectively, can be derived in a similar way as in Section 3.4.1.
- Indeed, the same minimization technique that has been used for the direct derivation of (87) yields (91). Moreover, the sum of squares SSE_H can be determined as follows.

– Just as before, $\Theta_H = \{\boldsymbol{\beta} \in \mathbb{R}^{1+k_1+k_2+k_1k_2} : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}\}$ is the restricted parameter space.
– Due to the special form of the matrices $\mathbf{X}$ and $\mathbf{H}$, the minimization in

$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta}\in\Theta_H} \left|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right|^2$$

with respect to the set $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \Theta_H\} \subset \mathbb{R}^{rk_1k_2}$ can (in a similar way as in formula (88)) be replaced by a minimization with respect to the set $\mathbb{R} \times \mathbb{R}_H^{k_2} \times \mathbb{R}_H^{k_1k_2} \subset \mathbb{R}^{1+k_2+k_1k_2}$ of those vectors $\mathbf{x} = (x, x_1, \ldots, x_{k_2}, x_{11}, \ldots, x_{k_1k_2}) \in \mathbb{R}^{1+k_2+k_1k_2}$ which fulfill the following conditions:

$$\sum_{i_2=1}^{k_2} x_{i_2} = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} x_{i_1i_2} = 0\,.$$

- In more detail, it holds that

$$\begin{aligned} SSE_H &= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_2}\times\mathbb{R}_H^{k_1k_2}} \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - (x + x_{i_2} + x_{i_1i_2})\big)^2 \\ &= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_2}\times\mathbb{R}_H^{k_1k_2}} \Bigg\{ \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\big)^2 + k_1k_2r\big(\overline{Y}_{\cdots} - x\big)^2 + k_2r\sum_{i_1=1}^{k_1}\big(\overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots}\big)^2 \\ &\quad + k_1r\sum_{i_2=1}^{k_2}\big(\overline{Y}_{\cdot i_2\cdot} - \overline{Y}_{\cdots} - x_{i_2}\big)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdot i_2\cdot} + \overline{Y}_{\cdots} - x_{i_1i_2}\big)^2 \Bigg\}, \end{aligned}$$

i.e.,

$$SSE_H = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\big)^2 + k_2r\sum_{i_1=1}^{k_1}\big(\overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots}\big)^2\,.$$

- Together with (91) this implies (92).

Remark

- In the same way, one can construct a test to investigate the question whether the levels of the second predictor variable are significant. For this purpose, we verify the hypothesis that the effects

$$\alpha_{i_2}^{(2)*} = \alpha_{i_2}^{(2)} + \frac{1}{k_1}\sum_{i_1=1}^{k_1}\alpha_{i_1i_2}$$

of the second predictor variable plus their interactions, averaged over all levels of the first predictor variable, are equal.

- Thus, the hypothesis to be tested is

$$H_0 : \alpha_1^{(2)*} - \alpha_2^{(2)*} = 0\,,\ \ldots\ ,\ \alpha_1^{(2)*} - \alpha_{k_2}^{(2)*} = 0 \quad \text{vs.} \quad H_1 : \alpha_1^{(2)*} - \alpha_{i_2}^{(2)*} \neq 0 \quad \text{for some } i_2 \in \{1, \ldots, k_2\}\,.$$

- In this case, one obtains the test statistic

$$T_{\mathbf{H}} = \frac{k_1k_2(r-1)rk_1\sum_{i_2=1}^{k_2}\big(\overline{Y}_{\cdot i_2\cdot} - \overline{Y}_{\cdots}\big)^2}{(k_2-1)\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\big)^2} \sim \mathrm{F}_{k_2-1,\,k_1k_2(r-1)}.$$

Interactions between both Predictor Variables

Now we construct a test to check whether there are significant interactions between the two predictor variables. For this purpose, the hypothesis

$$H_0 : \alpha^*_{11} - \alpha^*_{i_1 i_2} = 0 \quad \forall\, (i_1, i_2) \in \{1, \ldots, k_1\} \times \{1, \ldots, k_2\} \tag{93}$$

is tested, where

$$\alpha^*_{i_1 i_2} = \alpha_{i_1 i_2} - \overline{\alpha}_{i_1 \cdot} - \overline{\alpha}_{\cdot\, i_2} + \overline{\alpha}_{\cdot\cdot}$$

and

$$\overline{\alpha}_{i_1 \cdot} = \frac{1}{k_2} \sum_{i_2=1}^{k_2} \alpha_{i_1 i_2}\,, \qquad \overline{\alpha}_{\cdot\, i_2} = \frac{1}{k_1} \sum_{i_1=1}^{k_1} \alpha_{i_1 i_2}\,, \qquad \overline{\alpha}_{\cdot\cdot} = \frac{1}{k_1 k_2} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \alpha_{i_1 i_2}\,.$$

- In a similar way as before, one can show that the hypothesis considered in (93) can be written in the form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$, where
  - $\mathbf{H}$ is a $(k_1 k_2 - 1) \times m$ matrix with full row rank, i.e., $\mathrm{rk}(\mathbf{H}) = k_1 k_2 - 1$ and
  - all components of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable functions of $\boldsymbol{\beta}$ as it holds that

$$\alpha^*_{i_1 i_2} = \theta_{i_1 i_2} - \frac{1}{k_2} \sum_{i_2=1}^{k_2} \theta_{i_1 i_2} - \frac{1}{k_1} \sum_{i_1=1}^{k_1} \theta_{i_1 i_2} + \frac{1}{k_1 k_2} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \theta_{i_1 i_2}\,.$$

- For the verification of the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ we can thus consider the test statistic

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1 k_2 - 1)}{SSE/(k_1 k_2 (r-1))}$$

  considered in Theorem 3.16, where SSE is given by (91) as before while the sum of squares SSE_H results from the following considerations.
- Due to the special form of the matrices $\mathbf{X}$ and $\mathbf{H}$, the set $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \Theta_H\} \subset \mathbb{R}^{r k_1 k_2}$ in the minimization in

$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta} \in \Theta_H} \left|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right|^2$$

  can (in a similar way as before) be replaced by the set $\mathbb{R} \times \mathbb{R}_H^{k_1} \times \mathbb{R}_H^{k_2} \subset \mathbb{R}^{1+k_1+k_2}$ of those vectors

$$\mathbf{x} = (x, x_1^{(1)}, \ldots, x_{k_1}^{(1)}, x_1^{(2)}, \ldots, x_{k_2}^{(2)}) \in \mathbb{R}^{1+k_1+k_2}$$

  which fulfill the following conditions:

$$\sum_{i_1=1}^{k_1} x_{i_1}^{(1)} = \sum_{i_2=1}^{k_2} x_{i_2}^{(2)} = 0\,.$$

- Indeed, it holds that

$$SSE_H = \min_{\mathbf{x} \in \mathbb{R} \times \mathbb{R}_H^{k_1} \times \mathbb{R}_H^{k_2}} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} \big(Y_{i_1 i_2 j} - (x + x_{i_1}^{(1)} + x_{i_2}^{(2)})\big)^2$$

  and

$$\begin{aligned} SSE_H &= \min_{\mathbf{x} \in \mathbb{R} \times \mathbb{R}_H^{k_1} \times \mathbb{R}_H^{k_2}} \Bigg\{ \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} \big(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot}\big)^2 \\ &\quad + k_1 k_2 r \big(\overline{Y}_{\cdots} - x\big)^2 + k_2 r \sum_{i_1=1}^{k_1} \big(\overline{Y}_{i_1 \cdot\cdot} - \overline{Y}_{\cdots} - x_{i_1}^{(1)}\big)^2 + k_1 r \sum_{i_2=1}^{k_2} \big(\overline{Y}_{\cdot\, i_2 \cdot} - \overline{Y}_{\cdots} - x_{i_2}^{(2)}\big)^2 \\ &\quad + r \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \big(\overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot\cdot} - \overline{Y}_{\cdot\, i_2 \cdot} + \overline{Y}_{\cdots}\big)^2 \Bigg\} \\ &= \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{j=1}^{r} \big(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot}\big)^2 + r \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \big(\overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot\cdot} - \overline{Y}_{\cdot\, i_2 \cdot} + \overline{Y}_{\cdots}\big)^2\,. \end{aligned}$$

- Together with (91) this implies that

$$SSE_H - SSE = r \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \left(\overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\cdot i_2 \cdot} + \overline{Y}_{\cdots} \right)^2 . \tag{94}$$

- Therefore, for the test statistic $T_{\mathbf{H}}$ considered in Theorem 3.16 it holds that

$$\begin{aligned} T_{\mathbf{H}} &= \frac{(SSE_H - SSE)/(k_1 k_2 - 1)}{SSE/\big(k_1 k_2 (r-1)\big)} \\ &= \frac{k_1 k_2 (r-1) r \sum\limits_{i_1=1}^{k_1} \sum\limits_{i_2=1}^{k_2} \left(\overline{Y}_{i_1 i_2 \cdot} - \overline{Y}_{i_1 \cdot \cdot} - \overline{Y}_{\cdot i_2 \cdot} + \overline{Y}_{\cdots} \right)^2}{(k_1 k_2 - 1) \sum\limits_{i_1=1}^{k_1} \sum\limits_{i_2=1}^{k_2} \sum\limits_{j=1}^{r} \left(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot} \right)^2} \sim \mathrm{F}_{k_1 k_2 - 1,\, k_1 k_2 (r-1)} \, . \end{aligned}$$

3.4.3 Two–Factor Analysis of Variance with Hierarchical Classification

- Instead of the model of two–factor analysis of variance with interactions, introduced in Section 3.1.3, one sometimes considers the following model of two–factor analysis of variance with *hierarchical classification* of the pairs of levels i_1, i_2 of the two predictor variables.

- Here we consider the representation

$$\theta_{i_1 i_2} = \mu + \alpha_{i_1}^{(1)} + \alpha_{i_2|i_1}^{(2|1)} , \qquad \forall\, i_1 = 1, \ldots, k_1,\ i_2 = 1, \ldots, k_2 \tag{95}$$

  of the expectations $\theta_{i_1 i_2} = \mathbb{E}\, Y_{i_1 i_2 j}$ of the sampling variables $Y_{i_1 i_2 j}$.

- In other words: Into each of the k_1 levels of the first, i.e., superior, predictor variable k_2 levels of the second (inferior) predictor variable are embedded.

- This situation can occur, e.g., in clinical trials which are carried out in k_1 countries (superior predictor variable) and k_2 hospitals in each country (inferior predictor variable).

- Then the parameter vector $\boldsymbol{\beta}$ has the dimension $m = 1 + k_1 + k_1 k_2$ with

$$\boldsymbol{\beta} = \big(\mu, \alpha_1^{(1)}, \ldots, \alpha_{k_1}^{(1)}, \alpha_{1|1}^{(2|1)}, \ldots, \alpha_{k_2|k_1}^{(2|1)} \big)^\top$$

- and
  - μ is again perceived as *general mean* of the expectations $\mathbb{E}\, Y_{i_1 i_2 j}$ of the sampling variables $Y_{i_1 i_2 j}$,
  - $\alpha_{i_1}^{(1)}$ is called the *effect* of the i_1–th level of the superior predictor variable and
  - $\alpha_{i_2|i_1}^{(2|1)}$ is called the *effect* of the i_2–th level of the inferior predictor variable in the case that the i_1–th level of the superior predictor variable is on hand.

- Again, we only consider the balanced case, i.e., we assume that all $k_1 \cdot k_2$ subsamples $(Y_{i_1 i_2 j},\ j = 1, \ldots, n_{i_1 i_2})$ have the same sample size.

- Hence, it holds that $n_{i_1 i_2} = r$ for all $i_1 = 1, \ldots, k_1$ and $i_2 = 1, \ldots, k_2$ with $r = n/(k_1 k_2)$, the design matrix $\mathbf{X}$ has the dimension $n \times m$ with $n = r k_1 k_2$ and $m = 1 + k_1 + k_1 k_2$, and the entries of $\mathbf{X}$ only consist of zeros and ones; $\mathrm{rk}(\mathbf{X}) = k_1 k_2 < m$.

Significance of the Superior Predictor Variable

- Just in the same way as in Section 3.4.2 one can construct a test to investigate the question whether the levels of the superior predictor variable are significant. For this purpose, we verify the hypothesis that the mean effects $\alpha_{i_1}^{(1)*}$ are equal, where

$$\alpha_{i_1}^{(1)*} = \alpha_{i_1}^{(1)} + \frac{1}{k_2}\sum_{i_2=1}^{k_2} \alpha_{i_2|i_1}^{(2|1)}\,.$$

- In other words: The hypothesis to be tested is

$$H_0 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} = 0 \quad \forall\, i_1 \in \{1,\ldots,k_1\} \qquad \text{versus} \qquad H_1 : \alpha_1^{(1)*} - \alpha_{i_1}^{(1)*} \neq 0 \quad \text{for some } i_1 \in \{1,\ldots,k_1\}\,.$$

- One can show that the null hypothesis has the form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$, where $\mathbf{H}$ is a $(k_1 - 1) \times m$ matrix with full row rank, i.e., $\mathrm{rk}(\mathbf{H}) = k_1 - 1$, and all components of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable functions of $\boldsymbol{\beta}$.
- For the verification of the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ we can thus use the test statistic

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1 - 1)}{SSE/(k_1 k_2 (r-1))} \sim \mathrm{F}_{k_1-1,\, k_1 k_2 (r-1)}$$

considered in Theorem 3.16 with

$$SSE = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot}\big)^2\,, \quad SSE_H - SSE = r k_2 \sum_{i_1=1}^{k_1}\big(\overline{Y}_{i_1 \cdot\cdot} - \overline{Y}_{\cdots}\big)^2\,, \tag{96}$$

where the formulas in (96) are proved in the same way as in (91) and (92), respectively.

Significance of the Inferior Predictor Variable

- In order to check whether the levels of the inferior predictor variable are significant, one can proceed in a similar way as in the last test in Section 3.4.2 (test for significance of interactions). For this purpose, the hypothesis

$$H_0 : \alpha_{1|1}^{(2|1)*} - \alpha_{i_2|i_1}^{(2|1)*} = 0 \quad \forall\,(i_1, i_2) \in \{1,\ldots,k_1\} \times \{1,\ldots,k_2\} \tag{97}$$

is tested, where

$$\alpha_{i_2|i_1}^{(2|1)*} = \alpha_{i_2|i_1}^{(2|1)} - \overline{\alpha}_{i_1} + \overline{\alpha}\,, \qquad \overline{\alpha}_{i_1} = \frac{1}{k_2}\sum_{i_2=1}^{k_2} \alpha_{i_2|i_1}^{(2|1)}\,, \qquad \overline{\alpha} = \frac{1}{k_1 k_2}\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} \alpha_{i_2|i_1}^{(2|1)}\,.$$

- It can be shown that the hypothesis considered in (93) can be written in the form $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$, where $\mathbf{H}$ is a $k_1(k_2 - 1) \times m$ matrix with full row rank, i.e., $\mathrm{rk}(\mathbf{H}) = k_1(k_2 - 1)$, and all components of the vector $\mathbf{H}\boldsymbol{\beta}$ are estimable functions of $\boldsymbol{\beta}$.
- For the verification of the hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{o}$ one can thus use the test statistic

$$T_{\mathbf{H}} = \frac{(SSE_H - SSE)/(k_1(k_2 - 1))}{SSE/(k_1 k_2 (r-1))}$$

considered in Theorem 3.16, where the sums of squares SSE and SSE_H defined in (75) and (76), respectively, can be determined as follows.

- Just as before, it holds that

$$SSE = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1 i_2 j} - \overline{Y}_{i_1 i_2 \cdot}\big)^2\,, \tag{98}$$

and the set $\{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \Theta_H\}$ in the minimization in

$$SSE_H = (\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H)^\top(\mathbf{Y} - \mathbf{X}\overline{\boldsymbol{\beta}}_H) = \min_{\boldsymbol{\beta}\in\Theta_H} \left|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right|^2$$

can be replaced by the set $\mathbb{R} \times \mathbb{R}_H^{k_1} \subset \mathbb{R}^{1+k_1}$ of those vectors $\mathbf{x} = (x, x_1^{(1)}, \ldots, x_{k_1}^{(1)}) \in \mathbb{R}^{1+k_1}$ for which $\sum_{i_1=1}^{k_1} x_{i_1}^{(1)} = 0$ holds, leading to

$$\begin{aligned}
SSE_H &= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_1}} \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - (x + x_{i_1}^{(1)})\big)^2 \\
&= \min_{\mathbf{x}\in\mathbb{R}\times\mathbb{R}_H^{k_1}} \Bigg\{ \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\big)^2 + k_1k_2r\big(\overline{Y}_{\cdots} - x\big)^2 \\
&\quad + k_2r\sum_{i_1=1}^{k_1}\big(\overline{Y}_{i_1\cdot\cdot} - \overline{Y}_{\cdots} - x_{i_1}^{(1)}\big)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\big)^2 \Bigg\} \\
&= \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\big)^2 + r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\big)^2 .
\end{aligned}$$

- Together with (98) this implies that

$$SSE_H - SSE = r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\big)^2 .$$

- Therefore, it holds that

$$T_{\mathbf{H}} = \frac{k_1k_2(r-1)r\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\big(\overline{Y}_{i_1i_2\cdot} - \overline{Y}_{i_1\cdot\cdot}\big)^2}{k_1(k_2-1)\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{j=1}^{r}\big(Y_{i_1i_2j} - \overline{Y}_{i_1i_2\cdot}\big)^2} \sim \mathrm{F}_{k_1(k_2-1),\,k_1k_2(r-1)} .$$

4 Generalized Linear Models

- In Chapters 2 and 3 we always assumed for the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - that $\mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{o}$, i.e., $\big(\mathbb{E}\,Y_1, \ldots, \mathbb{E}\,Y_n\big)^\top = \mathbf{X}\boldsymbol{\beta}$,
  - where furthermore $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ holds if $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I})$.
- Now, we generalize this model and allow that the expectations $\mathbb{E}\,Y_1, \ldots, \mathbb{E}\,Y_n$ of the sampling variables $Y_1, \ldots, Y_n$
  - can be expressed by the components of the vector $\mathbf{X}\boldsymbol{\beta}$ by using an arbitrary monotone function $g : G \to \mathbb{R}$, the so–called *link function*, such that
$$\big(g(\mathbb{E}\,Y_1), \ldots, g(\mathbb{E}\,Y_n)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,, \tag{1}$$
  - where the domain $G \subset \mathbb{R}$ of g will be specified more precisely.
- Moreover, the (independent) sampling variables $Y_1, \ldots, Y_n$ do not have to be normally distributed since we just assume that the distributions of $Y_1, \ldots, Y_n$ belong to an *exponential family*.
- In this chapter we will always assume (as in Chapter 2) that the design matrix $\mathbf{X}$ has full column rank, i.e., $\mathrm{rk}(\mathbf{X}) = m$.

In the same way as in the linear models, which have been investigated in Chapters 2 and 3, the goal is to estimate the parameter vector $\boldsymbol{\beta}$ from the observation of the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, where we assume that the link function $g : G \to \mathbb{R}$ is known.

4.1 Definition and Basic Properties

4.1.1 Exponential Family

We assume that the sample variables $Y_1, \ldots, Y_n$ are independent (but in general *not* identically distributed),

- where their distributions belong to a one-parametric *exponential family*, i.e., their densities or probability mass functions, respectively, have the following form: For each $i \in \{1, \ldots, n\}$ it holds that
  - in the absolutely continuous case
$$f(y; \theta_i) = \exp\Big(\frac{1}{\tau^2}\,\big(y\theta_i + a(y,\tau) - b(\theta_i)\big)\Big)\,, \qquad \forall\, y \in \mathbb{R}\,, \tag{2}$$
  - in the discrete case
$$\mathbb{P}_{\theta_i}(Y_i = y) = \exp\Big(\frac{1}{\tau^2}\,\big(y\theta_i + a(y,\tau) - b(\theta_i)\big)\Big)\,, \qquad \forall\, y \in C\,, \tag{3}$$
    where $a : \mathbb{R} \times (0,\infty) \to \mathbb{R}$ and $b : \Theta \to \mathbb{R}$ are certain functions and $C \subset \mathbb{R}$ is the smallest countable subset of $\mathbb{R}$, for which it holds that $\mathbb{P}_{\theta_i}(Y_i \in C) = 1$.
- $\tau^2 > 0$ is a so–called *nuisance parameter*, which does not depend on the index i, where it is often assumed that τ^2 is known.
- Then
$$\Theta = \Big\{\theta \in \mathbb{R} : \int_{-\infty}^{\infty} \exp\Big(\frac{y\theta + a(y,\tau)}{\tau^2}\Big)\,dy < \infty\Big\} \tag{4}$$
  or
$$\Theta = \Big\{\theta \in \mathbb{R} : \sum_{y \in C} \exp\Big(\frac{y\theta + a(y,\tau)}{\tau^2}\Big) < \infty\Big\} \tag{5}$$
  is the *natural parameter space*, where we always assume that the integrability condition in (4) or (5), respectively, is fulfilled for at least two different $\theta_1, \theta_2 \in \mathbb{R}$.

Remark In the absolutely continuous case the nuisance parameter τ^2 can be seen as an additional variance parameter, while τ^2 usually is set to 1 in the discrete case.

Lemma 4.1 *The parameter space $\Theta \subset \mathbb{R}$, given in (4) and (5), respectively, is an interval in $\mathbb{R}$.*

Proof

- We only consider the absolutely continuous case since the proof of the discrete case proceeds analogously.
- One can easily see that for arbitrary $x_1, x_2 \in \mathbb{R}$ and $\alpha \in (0,1)$ we have

$$\left(e^{x_1}\right)^{\alpha} \left(e^{x_2}\right)^{1-\alpha} \;\leq\; \max_{i=1,2,} \left(e^{x_i}\right)^{\alpha} \left(e^{x_i}\right)^{1-\alpha} \;=\; \max_{i=1,2,} e^{x_i} \;\leq\; e^{x_1} + e^{x_2}\,.$$

- This and the notation $\theta = \alpha\theta_1 + (1-\alpha)\theta_2$ imply that for arbitrary $\theta_1, \theta_2 \in \Theta$ and $\alpha \in (0,1)$

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\Big(\frac{y\theta + a(y,\tau)}{\tau^2}\Big)\, dy &= \int_{-\infty}^{\infty} \Big(\exp\Big(\frac{y\theta_1 + a(y,\tau)}{\tau^2}\Big)\Big)^{\alpha} \Big(\exp\Big(\frac{y\theta_2 + a(y,\tau)}{\tau^2}\Big)\Big)^{1-\alpha} dy \\ &\leq \int_{-\infty}^{\infty} \Big(\exp\Big(\frac{y\theta_1 + a(y,\tau)}{\tau^2}\Big) + \exp\Big(\frac{y\theta_2 + a(y,\tau)}{\tau^2}\Big)\Big)\, dy < \infty\,.\end{aligned}$$

- Therefore, it also holds that $\theta \in \Theta$. □

Because of Lemma 4.1 we will always assume in this chapter that $\Theta \subset \mathbb{R}$ is an *open* interval such that the integrability condition in (4) and (5), respectively, is fulfilled for each $\theta \in \Theta$.

Lemma 4.2

- *Let the distribution of the random variable $Y : \Omega \to \mathbb{R}$ be given by (2) or (3) for an arbitrary $\theta \in \Theta$ such that*

$$\mathbb{E}\,(Y^2) < \infty \qquad \forall\, \theta \in \Theta \tag{6}$$

holds and the function $b : \Theta \to \mathbb{R}$ is twice continuously differentiable.

- *Then it holds that*

$$\mathbb{E}\,Y = b^{(1)}(\theta) \qquad \textit{and} \qquad \operatorname{Var} Y = \tau^2\, b^{(2)}(\theta)\,. \tag{7}$$

Proof

- Once more, we only treat the absolutely continuous case since the proof of the discrete case proceeds analogously. It holds that

$$\begin{aligned}\mathbb{E}\,Y &= \int_{-\infty}^{\infty} y\, \exp\Big(\frac{1}{\tau^2}\,\big(y\theta + a(y,\tau) - b(\theta)\big)\Big)\, dy \;=\; e^{-b(\theta)/\tau^2} \int_{-\infty}^{\infty} y\, \exp\Big(\frac{1}{\tau^2}\,\big(y\theta + a(y,\tau)\big)\Big)\, dy \\ &= e^{-b(\theta)/\tau^2}\, \tau^2 \int_{-\infty}^{\infty} \frac{d}{d\theta}\, \exp\Big(\frac{1}{\tau^2}\,\big(y\theta + a(y,\tau)\big)\Big)\, dy \\ &= e^{-b(\theta)/\tau^2}\, \tau^2\, \frac{d}{d\theta} \int_{-\infty}^{\infty} \exp\Big(\frac{1}{\tau^2}\,\big(y\theta + a(y,\tau)\big)\Big)\, dy \\ &= e^{-b(\theta)/\tau^2}\, \tau^2\, \frac{d}{d\theta}\, e^{b(\theta)/\tau^2} \underbrace{\int_{-\infty}^{\infty} \exp\Big(\frac{1}{\tau^2}\,\big(y\theta + a(y,\tau) - b(\theta)\big)\Big)\, dy}_{=1} \\ &= b^{(1)}(\theta)\,.\end{aligned}$$

- In a similar way we get that $\mathbb{E}\,(Y^2) = \tau^2\, b^{(2)}(\theta) + (\mathbb{E}\,Y)^2$. □

4.1.2 Link of the Parameters; Natural Link Function

- Now, we assume that the function $b : \Theta \to \mathbb{R}$ is twice continuously differentiable with $b^{(2)}(\theta) > 0$ for each $\theta \in \Theta$.
- Furthermore, let $G = \{b^{(1)}(\theta) : \theta \in \Theta\}$ and the link function $g : G \to \mathbb{R}$ be twice continuously differentiable, such that $g^{(1)}(x) \neq 0$ for each $x \in G$. The inverse function of g is denoted by $h = g^{-1}$.
- We consider the generalized linear model (GLM) given in (1), i.e., it holds that

$$\big(g(\mathbb{E}\,Y_1), \ldots, g(\mathbb{E}\,Y_n)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{8}$$

  - By using the notation $\mathbf{X} = (x_{ij})$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^\top$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$, where $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, formula (8) implies for the expectations $\mu_i = \mathbb{E}\,Y_i$ $(i = 1, \ldots, n)$ that

$$\mu_i = h(\eta_i) = h\big(\mathbf{x}_i^\top \boldsymbol{\beta}\big) \qquad \text{and thus} \qquad \boldsymbol{\mu} = \big(h(\eta_1), \ldots, h(\eta_n)\big)^\top, \tag{9}$$

    where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$.
  - Because of (7) and (8) the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ are related as follows: It holds that

$$\big(g\big(b^{(1)}(\theta_1)\big), \ldots, g\big(b^{(1)}(\theta_n)\big)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{10}$$

- Together with (9) this implies that

$$\big(b^{(1)}(\theta_1), \ldots, b^{(1)}(\theta_n)\big) = \big(h(\mathbf{x}_1^\top \boldsymbol{\beta}), \ldots, h(\mathbf{x}_n^\top \boldsymbol{\beta})\big)$$

  and, equivalently,

$$(\theta_1, \ldots, \theta_n) = \big(\psi\big(h(\mathbf{x}_1^\top \boldsymbol{\beta})\big), \ldots, \psi\big(h(\mathbf{x}_n^\top \boldsymbol{\beta})\big)\big)\,, \tag{11}$$

  where $\psi = \big(b^{(1)}\big)^{-1}$ is the inverse function of $b^{(1)}$.
- Furthermore, it is possible to express the variance $\sigma_i^2 = \operatorname{Var} Y_i$ of the sample variables Y_i as a function $\sigma_i^2(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ for each $i = 1, \ldots, n$ as it follows from Lemma 4.2 and (11) that

$$\sigma_i^2(\boldsymbol{\beta}) = \tau^2\, b^{(2)}\big(\psi\big(h(\mathbf{x}_i^\top \boldsymbol{\beta})\big)\big) \qquad \forall\, i = 1, \ldots, n\,. \tag{12}$$

Remark The link function $g : G \to \mathbb{R}$ is called *natural* if $g = \psi$. In this case it holds that $\theta_i = \psi(\mu_i)$ and, therefore, $\theta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ for each $i = 1, \ldots, n$, i.e.,

$$(\theta_1, \ldots, \theta_n)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{13}$$

4.2 Examples

4.2.1 Linear Model with Normally Error Terms

- For the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{14}$$

  with normally distributed error terms $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim \mathrm{N}(\mathbf{o}, \sigma^2\mathbf{I})$, considered in Section 2.2, it holds that

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2) \quad \text{with } \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}\,, \qquad \forall\, i = 1, \ldots, n, \tag{15}$$

  where we assume that σ^2 is known.
- Then the distribution of Y_i belongs to the one–parametric exponential family considered in Section 4.1.1 since the density $f(y; \theta_i)$ of Y_i can be written in the following form, where $\theta_i = \mu_i$ for each $i = 1, \ldots, n$:

– It holds that

$$f(y;\theta_i) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\Big(-\,\frac{1}{2\sigma^2}\,(y-\mu_i)^2\Big) \;=\; \exp\Big(\frac{1}{\tau^2}\,\big(y\theta_i + a(y,\tau) - b(\theta_i)\big)\Big) \qquad \forall\, y\in\mathbb{R}\,,$$

– where

$$\tau^2 = \sigma^2\,, \qquad a(y) = -\,\frac{y^2}{2} \qquad \text{and} \qquad b(\theta_i) = \frac{\theta_i^2}{2} \;+\; \sigma^2 \log\sqrt{2\pi\sigma^2}\,. \tag{16}$$

- Because of (15) the link function $g:\mathbb{R}\to\mathbb{R}$ fulfills $g(x)=x$ for each $x\in\mathbb{R}$.
  - Moreover, (16) yields $x = b^{(1)}(x)$ for each $x\in\mathbb{R}$.
  - Therefore, it holds that $g(x) = x = \psi(x)$ for each $x\in\mathbb{R}$, i.e., the natural link function is given by $g(x)=x$.

4.2.2 Binary Categorical Regression

- In this section we consider the case that the sample variables $Y_1,\dots,Y_n$ are Bernoulli–distributed, i.e., they can only take the values 0 and 1 with positive probability.
  - Here, we use the notation

$$\pi_i = \mathbb{P}(Y_i=1) \qquad \Big(=\mu_i \;=\; \mathbb{E}\,Y_i\Big) \qquad \forall\, i=1,\dots,n\,,$$

    where it is assumed that $0<\pi_i<1$ for each $i=1,\dots,n$.
  - In this case the probabilities $\pi_1,\dots,\pi_n$ are linked to the parameter vector $\boldsymbol{\beta}$ by using a link function $g:(0,1)\to\mathbb{R}$, i.e.,

$$\big(g(\pi_1),\dots,g(\pi_n)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{17}$$

- For each $i=1,\dots,n$ the $\mathrm{Bin}(1,\pi_i)$–distribution belongs to the exponential family introduced in Section 4.1.1, where $\theta_i = \log(\pi_i/(1-\pi_i))$.
  - Because for $y=0,1$ it holds that

$$\mathbb{P}_{\theta_i}(Y_i=y) = \pi_i^y\,(1-\pi_i)^{1-y} = \exp\Big(y\log\Big(\frac{\pi_i}{1-\pi_i}\Big) + \log(1-\pi_i)\Big) = \exp\Big(\frac{1}{\tau^2}\,\big(y\theta_i + a(y,\tau) - b(\theta_i)\big)\Big)\,,$$

  - where

$$\tau^2 = 1\,, \qquad a(y) = 0 \qquad \text{and} \qquad b(\theta_i) = \log(1+e^{\theta_i})\,. \tag{18}$$

Remark

- From (18) it follows that $(b^{(1)})^{-1}(x) = \log(x/(1-x))$ for each $x\in(0,1)$, i.e., the natural link function $g:(0,1)\to\mathbb{R}$ is given by

$$g(x) = \log\Big(\frac{x}{1-x}\Big) \qquad \forall\, x\in(0,1)\,. \tag{19}$$

  - The GLM, considered in (17), with the natural link function, given in (19), is then called (binary) *logistic regression model*.
  - In this case the dependency of the probabilities $\pi_i = \pi_i(\boldsymbol{\beta})$ of the linear combinations $\mathbf{x}_i^\top\boldsymbol{\beta}$ is given by

$$\pi_i \;=\; \frac{1}{1+\exp(-\mathbf{x}_i^\top\boldsymbol{\beta})} \qquad \forall\, i=1,\dots,n\,. \tag{20}$$

- Another (nonnatural) link function $g:(0,1)\to\mathbb{R}$, which is considered in this context, is given by

$$g = \Phi^{-1}\,, \tag{21}$$

  - where $\Phi:\mathbb{R}\to(0,1)$ denotes the distribution function of the $\mathrm{N}(0,1)$–distribution.
  - Then it holds that $\pi_i = \Phi(\mathbf{x}_i^\top\boldsymbol{\beta})$ for each $i=1,\dots,n$ and the GLM is called the model of the *probit analysis*.

4.2.3 Poisson–Distributed Sample Variables with Natural Link Function

- Now, let the sample variables $Y_1, \ldots, Y_n$ be Poisson–distributed, i.e., let $Y_i \sim \text{Poi}(\lambda_i)$ with $0 < \lambda_i < \infty$ for each $i = 1, \ldots, n$.
- The $\text{Poi}(\lambda_i)$–distribution also belongs to the exponential family introduced in Section 4.1.1, where $\theta_i = \log \lambda_i$
  - because it holds for each $y = 0, 1, \ldots$ that

$$\mathbb{P}_{\theta_i}(Y_i = y) \;=\; \frac{\lambda_i^y e^{-\lambda_i}}{y!} = \exp\big(y \log \lambda_i - \log(y!) - \lambda_i\big) = \exp\Big(\frac{1}{\tau^2}\,\big(y\theta_i + a(y,\tau) - b(\theta_i)\big)\Big)\,,$$

  - where

$$\tau^2 = 1\,, \qquad a(y) = -\log(y!) \qquad \text{and} \qquad b(\theta_i) = e^{\theta_i}\,.$$

- The natural link function $g : (0, \infty) \to \mathbb{R}$ is given by

$$g(x) = \log x \qquad \forall\, x > 0\,. \tag{22}$$

4.3 Maximum–Likelihood Estimator for $\boldsymbol{\beta}$

- Since we assumed, that the distributions of the sample variables $Y_1, \ldots, Y_n$ belong to an exponential family, it is possible to estimate the parameter vector $\boldsymbol{\beta}$ by using the maximum–likelihood method.
- In order to show this, we first discuss some properties of the loglikelihood function $\log L(\mathbf{Y}, \boldsymbol{\theta})$ of the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and its partial derivatives with respect to the components $\beta_1, \ldots, \beta_m$ of $\boldsymbol{\beta}$.

4.3.1 Loglikelihood Function and its Partial Derivatives

- From (2) – (3) and from (11) it follows that the loglikelihood function $\log L(\mathbf{Y}, \boldsymbol{\theta})$ of the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ can be written as a function $\log L(\mathbf{Y}, \boldsymbol{\beta})$ of $\boldsymbol{\beta}$.
  - From (2) – (3) it follows that

$$\log L(\mathbf{Y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{\tau^2}\,\big(Y_i\theta_i + a(Y_i,\tau) - b(\theta_i)\big)\,. \tag{23}$$

  - This and (11) imply that

$$\log L(\mathbf{Y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\tau^2}\,\Big(Y_i\psi\big(h(\mathbf{x}_i^\top\boldsymbol{\beta})\big) + a(Y_i,\tau) - b\big(\psi\big(h(\mathbf{x}_i^\top\boldsymbol{\beta})\big)\big)\Big)\,. \tag{24}$$

- For generalized linear models with natural link function, (13) and (23) imply that

$$\log L(\mathbf{Y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\tau^2}\,\big(Y_i\mathbf{x}_i^\top\boldsymbol{\beta} + a(Y_i,\tau) - b\big(\mathbf{x}_i^\top\boldsymbol{\beta}\big)\big)\,. \tag{25}$$

For the computation of the maximum–likelihood estimators, the knowledge of the so–called score function, i.e., the partial derivative of the loglikelihood function, is useful, as well as the Fisher–information matrix, which is defined as follows.

Definition For arbitrary $i, j = 1, \ldots, m$ let

$$U_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_i} \log L(\mathbf{Y}, \boldsymbol{\beta}) \qquad \text{and} \qquad I_{ij}(\boldsymbol{\beta}) = \mathbb{E}\left(U_i(\boldsymbol{\beta}) U_j(\boldsymbol{\beta})\right).$$

Then the m–dimensional random vector $\mathbf{U}(\boldsymbol{\beta}) = \big(U_1(\boldsymbol{\beta}), \ldots, U_m(\boldsymbol{\beta})\big)^\top$ is called the *score vector* and the (deterministic) $m \times m$–matrix $\mathbf{I}(\boldsymbol{\beta}) = \big(I_{ij}(\boldsymbol{\beta})\big)$ is called the *Fisher–information matrix.*

By using the notation

$$\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta}) \;=\; \frac{dh(s)}{ds}\Big|_{s=\eta_i} \;=\; \Big(\frac{dg(t)}{dt}\Big)^{-1}\Big|_{t=h(\eta_i)} \tag{26}$$

we get the following result.

Theorem 4.1 *For arbitrary $j, k = 1, \ldots, m$ it holds that*

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij}\big(Y_i - \mu_i(\boldsymbol{\beta})\big)\, \frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})} \tag{27}$$

and

$$I_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij} x_{ik} \Big(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Big)^2\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})} \tag{28}$$

or in matrix notation

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{V}^{-1}(\boldsymbol{\beta})\, \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\, (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \qquad \textit{and} \qquad \mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{V}^{-1}(\boldsymbol{\beta}) \Big(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\Big)^2 \mathbf{X}, \tag{29}$$

where

$$\mathbf{V}(\boldsymbol{\beta}) \;=\; \operatorname{diag}\big(\sigma_i^2(\boldsymbol{\beta})\big) \qquad \textit{and} \qquad \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta}) \;=\; \operatorname{diag}\Big(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Big).$$

Proof

- The loglikelihood function, given in (23) or (24), respectively, can be written in the form

$$\log L(\mathbf{Y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\tau^2}\, \ell^{(i)}(\theta_i),$$

  where $\ell^{(i)}(\theta_i) = Y_i \theta_i + a(Y_i, \tau) - b(\theta_i)$ and $\theta_i = \psi\big(h(\mathbf{x}_i^\top \boldsymbol{\beta})\big)$.
- Therefore, it holds for each $j = 1, \ldots, m$ that

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\tau^2}\, \frac{\partial \ell^{(i)}}{\partial \beta_j}(\theta_i), \tag{30}$$

  where multiple use of the chain rule yields

$$\frac{\partial \ell^{(i)}}{\partial \beta_j} \;=\; \frac{\partial \ell^{(i)}}{\partial \theta_i}\, \frac{\partial \theta_i}{\partial \mu_i}\, \frac{\partial \mu_i}{\partial \eta_i}\, \frac{\partial \eta_i}{\partial \beta_j}. \tag{31}$$

  – On the other hand, it obviously holds that $\partial \eta_i / \partial \beta_j = x_{ij}$ and from Lemma 4.2 it follows that

$$\frac{\partial \ell^{(i)}}{\partial \theta_i} \;=\; Y_i - b^{(1)}(\theta_i) \overset{\text{Lemma 4.2}}{=} Y_i - \mu_i$$

  or

$$\Big(\frac{\partial \theta_i}{\partial \mu_i}\Big)^{-1} \;=\; \frac{\partial \mu_i}{\partial \theta_i} \overset{\text{Lemma 4.2}}{=} b^{(2)}(\theta_i) \overset{\text{Lemma 4.2}}{=} \frac{1}{\tau^2}\, \sigma_i^2.$$

– This and (30) – (31) imply (27).

- In order to show (28) it is enough to notice that for arbitrary $i, j = 1, \ldots, n$

$$\mathbb{E}\big((Y_i - \mu_i)(Y_j - \mu_j)\big) = \begin{cases} \sigma_i^2 & \text{für } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

because of the independence of the sample variables $Y_1, \ldots, Y_n$.

– From this and (27) it follows that

$$\begin{aligned} I_{jk}(\boldsymbol{\beta}) &= \mathbb{E}\big(U_j(\boldsymbol{\beta})U_k(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} x_{ij}x_{ik}\Big(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Big)^2 \frac{1}{\sigma_i^4(\boldsymbol{\beta})}\,\mathbb{E}\big((Y_i - \mu_i)^2\big) \\ &= \sum_{i=1}^{n} x_{ij}x_{ik}\Big(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\Big)^2 \frac{1}{\sigma_i^2(\boldsymbol{\beta})}\,. \end{aligned}$$

– Therefore, (28) is proved. □

Corollary 4.1 *Let* $\big(g(\mathbb{E}\,Y_1), \ldots, g(\mathbb{E}\,Y_n)\big)^\top = \mathbf{X}\boldsymbol{\beta}$ *be a GLM with natural link function* $g : G \to \mathbb{R}$. *Then it holds for arbitrary* $j, k = 1, \ldots, m$ *that*

$$U_j(\boldsymbol{\beta}) = \frac{1}{\tau^2}\sum_{i=1}^{n} x_{ij}\big(Y_i - \mu_i(\boldsymbol{\beta})\big) \qquad \textit{or} \qquad \mathbf{U}(\boldsymbol{\beta}) = \frac{1}{\tau^2}\,\mathbf{X}^\top(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \tag{32}$$

and

$$I_{jk}(\boldsymbol{\beta}) = \frac{1}{\tau^4}\sum_{i=1}^{n} x_{ij}x_{ik}\sigma_i^2(\boldsymbol{\beta}) \qquad \textit{or} \qquad \mathbf{I}(\boldsymbol{\beta}) = \frac{1}{\tau^4}\,\mathbf{X}^\top\mathbf{V}(\boldsymbol{\beta})\mathbf{X}\,. \tag{33}$$

Proof Since $g : G \to \mathbb{R}$ is a natural link function, we have $\theta_i = \eta_i$ for each $i = 1, \ldots, n$. This and Lemma 4.2 imply that

$$\frac{d\mu_i}{d\eta_i} = b^{(2)}(\theta_i) = \frac{1}{\tau^2}\,\sigma_i^2\,.$$

Now, the statement follows from Theorem 4.1. □

4.3.2 Hessian Matrix

Besides the (score) vector $\mathbf{U}(\boldsymbol{\beta})$, which consists of the first partial derivatives of the loglikelihood function $\log L(\mathbf{Y}, \boldsymbol{\beta})$, also the *Hessian matrix*, i.e., the $m \times m$–matrix

$$\mathbf{W}(\boldsymbol{\beta}) = \big(W_{ij}(\boldsymbol{\beta})\big) = \left(\frac{\partial^2}{\partial\beta_i\partial\beta_j}\,\log L(\mathbf{Y}, \boldsymbol{\beta})\right),$$

consisting of the second partial derivatives of the loglikelihood function, is needed.

Theorem 4.2

- *For each GLM it holds that*

$$\mathbf{W}(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{R}(\boldsymbol{\beta}) \mathrm{diag}\big(Y_i - \mu_i(\boldsymbol{\beta})\big) \mathbf{X} - \mathbf{I}(\boldsymbol{\beta})\,, \tag{34}$$

where $\mathbf{I}(\boldsymbol{\beta})$ *is the Fisher–information matrix, given in* (29)*, and* $\mathbf{R}(\boldsymbol{\beta}) = \mathrm{diag}\big(v_i(\boldsymbol{\beta})\big)$ *is an* $(n \times n)$*–diagonal matrix with*

$$v_i(\boldsymbol{\beta}) \;=\; \frac{1}{\tau^2}\, \frac{d^2 u(s)}{ds^2}\Big|_{s=\mathbf{x}_i^\top \boldsymbol{\beta}} \qquad \textit{and} \qquad u = \psi \circ h\,.$$

- *For a GLM with natural link function it particularly holds that*

$$\mathbf{W}(\boldsymbol{\beta}) = -\mathbf{I}(\boldsymbol{\beta})\,. \tag{35}$$

Proof

- From formula (27) in Theorem 4.1 it follows that for arbitrary $j, k = 1, \ldots, m$

$$\begin{aligned} W_{jk}(\boldsymbol{\beta}) \;&=\; \frac{\partial}{\partial \beta_k}\, U_j(\boldsymbol{\beta}) \overset{(27)}{=} \frac{\partial}{\partial \beta_k} \sum_{i=1}^n x_{ij} \big(Y_i - \mu_i(\boldsymbol{\beta})\big)\, \frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})} \\ &=\; \sum_{i=1}^n x_{ij} \left(\big(Y_i - \mu_i(\boldsymbol{\beta})\big)\, \frac{\partial}{\partial \beta_k} \Big(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})} \Big) - \frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})}\, \frac{\partial \mu_i}{\partial \beta_k}\,(\boldsymbol{\beta}) \right). \end{aligned}$$

  – Here with the notation $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ we get from Lemma 4.2 that

$$\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})} \;=\; b^{(2)}\big(\psi \circ h(\eta_i)\big)\, (\psi \circ h)^{(1)}(\eta_i)\, \frac{1}{\tau^2\, b^{(2)}\big(\psi \circ h(\eta_i)\big)} \;=\; \frac{1}{\tau^2}\, (\psi \circ h)^{(1)}(\eta_i)$$

  and therefore

$$\frac{\partial}{\partial \beta_k} \Big(\frac{d\mu_i}{d\eta_i}(\boldsymbol{\beta})\, \frac{1}{\sigma_i^2(\boldsymbol{\beta})} \Big) \;=\; \frac{1}{\tau^2}\, (\psi \circ h)^{(2)}(\eta_i) x_{ik}\,.$$

  – Furthermore, it holds that

$$\frac{\partial \mu_i}{\partial \beta_k} \;=\; \frac{d\mu_i}{d\eta_i}\, \frac{\partial \eta_i}{\partial \beta_k}\,.$$

- Altogether, it follows that

$$W_{jk}(\boldsymbol{\beta}) \;=\; \sum_{i=1}^n x_{ij} x_{ik} (Y_i - \mu_i) v_i - \sum_{i=1}^n x_{ij} x_{ik} \Big(\frac{d\mu_i}{d\eta_j} \Big)^2\, \frac{1}{\sigma_i^2}\,.$$

- This and the representation formula (29) for the Fisher–information matrix $\mathbf{I}(\boldsymbol{\beta})$ yield (34).
- Since the superposition $u = \psi \circ h$ of a GLM with natural link function is the identity function, it holds in this case that $\mathbf{R}(\boldsymbol{\beta}) = \mathbf{0}$. Therefore, (35) follows from (34). □

Remark For the examples of GLM considered in Section 4.2, we get the following formulas for $\mathbf{U}(\boldsymbol{\beta})$ and $\mathbf{W}(\boldsymbol{\beta})$ from Theorems 4.1 and 4.2 or from Corollary 4.1, respectively.

1. For the linear model $\mathbb{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ with normally distributed sample variables (and with the link function $g(x) = x$), one has that $(d\boldsymbol{\mu}/d\boldsymbol{\eta})(\boldsymbol{\beta})$ is the identity matrix. Therefore, it holds that

$$\mathbf{U}(\boldsymbol{\beta}) \;=\; \frac{1}{\sigma^2}\, \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\,, \qquad \mathbf{W}(\boldsymbol{\beta}) \;=\; -\,\frac{1}{\sigma^2}\, \mathbf{X}^\top \mathbf{X}\,, \tag{36}$$

   cf. Section 2.2.

2. For the logistic regression model (with the natural link function) it holds that

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^\top(\mathbf{Y}-\boldsymbol{\pi}), \qquad \mathbf{W}(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathrm{diag}\big(\pi_i(1-\pi_i)\big)\mathbf{X}, \tag{37}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ and the probabilities π_i can be expressed by $\boldsymbol{\beta}$ (cf.(20)).

3. For Poisson–distributed sample variables with natural link function it holds that

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^\top(\mathbf{Y}-\boldsymbol{\lambda}), \qquad \mathbf{W}(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathrm{diag}\big(\lambda_i\big)\mathbf{X}, \tag{38}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$ and $\lambda_i = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$.

4.3.3 Maximum–Likelihood Equation and Numerical Approach

- In order to determine the maximum–likelihood estimator for $\boldsymbol{\beta}$, the *maximum–likelihood equation*

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{o} \tag{39}$$

is considered, which in general is nonlinear and therefore often can be solved only by using iterative methods.

- Because of Theorem 4.1 the equation (39) is equivalent to

$$\mathbf{X}^\top \mathbf{V}^{-1}(\boldsymbol{\beta}) \, \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta}) \, (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{o}. \tag{40}$$

Remark

- From Corollary 4.1 it follows that, in the case of a natural link function, (40) simplifies to:

$$\mathbf{X}^\top \, (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = \mathbf{o}. \tag{41}$$

- Since we furthermore assume that $0 < \sigma_i^2(\boldsymbol{\beta}) < \infty$ for each $i = 1, \ldots, n$ and that the design matrix $\mathbf{X}$ has full column rank, the matrix $\mathbf{W}(\boldsymbol{\beta}) = -\tau^{-4}\mathbf{X}^\top \mathbf{V}(\boldsymbol{\beta})\mathbf{X}$ of the second partial derivatives is negative definite.
- Hence, it holds that if (41) has a solution, then the solution is a uniquely determined maximum–likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

Now, we discuss the basic ideas of two numerical iteration methods for solving the maximum–likelihood equation (39). We consider a sequence of random vectors $\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \ldots : \Omega \to \mathbb{R}^m$, which converge under certain conditions to a random vector $\widehat{\boldsymbol{\beta}}$ such that $\widehat{\boldsymbol{\beta}}$ is a solution of (39).

1. *Newton's Method*
   - Let $\widehat{\boldsymbol{\beta}}_0 : \Omega \to \mathbb{R}^m$ be a suitably chosen start vector and let the iterations $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_k$ be already computed.
   - For the computation of the $(k+1)$–th iteration $\widehat{\boldsymbol{\beta}}_{k+1}$ from $\widehat{\boldsymbol{\beta}}_k$, the left hand side $\mathbf{U}(\boldsymbol{\beta})$ of the maximum–likelihood equation (39) is replaced by
     - the first two terms $\mathbf{U}(\widehat{\boldsymbol{\beta}}_k) + \mathbf{W}(\widehat{\boldsymbol{\beta}}_k)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k)$ of the Taylor series expansion of $\mathbf{U}(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_k$.
     - The $(k+1)$–th iteration $\widehat{\boldsymbol{\beta}}_{k+1}$ is also a solution of the equation

$$\mathbf{U}(\widehat{\boldsymbol{\beta}}_k) + \mathbf{W}(\widehat{\boldsymbol{\beta}}_k)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_k) = \mathbf{o}. \tag{42}$$

- If the matrix $\mathbf{W}(\widehat{\boldsymbol{\beta}}_k)$ is invertible, then it follows from (42) that

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k - \mathbf{W}^{-1}(\widehat{\boldsymbol{\beta}}_k)\mathbf{U}(\widehat{\boldsymbol{\beta}}_k)\,, \tag{43}$$

- For the constructed sequence $\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1, \ldots$ to converge to $\widehat{\boldsymbol{\beta}}$, this vector $\widehat{\boldsymbol{\beta}}$ has to be a solution of (39) and the start vector $\widehat{\boldsymbol{\beta}}_0$ has to be close enough to $\widehat{\boldsymbol{\beta}}$.

2. *Fisher–Scoring*

- Now we consider a variation of Newton's method, the so–called *scoring–method of Fisher*, where the Hessian matrix $\mathbf{W}(\boldsymbol{\beta})$ in (42) is replaced by the expectation matrix $\mathbb{E}\,\mathbf{W}(\boldsymbol{\beta})$.
  - This has the advantage that the $(m \times m)$–matrix $\mathbb{E}\,\mathbf{W}(\boldsymbol{\beta})$ is invertible.
  - From Theorems 4.1 and 4.2 it follows that

$$\begin{aligned} \mathbb{E}\,\mathbf{W}(\boldsymbol{\beta}) &\overset{(34)}{=} \mathbb{E}\left(\mathbf{X}^\top \mathbf{R}(\boldsymbol{\beta})\mathrm{diag}\big(Y_i - \mu_i(\boldsymbol{\beta})\big)\mathbf{X} - \mathbf{I}(\boldsymbol{\beta})\right) \\ &= -\mathbf{I}(\boldsymbol{\beta}) \\ &\overset{(29)}{=} -\mathbf{X}^\top \mathbf{V}^{-1}(\boldsymbol{\beta}) \left(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\right)^2 \mathbf{X}\,, \end{aligned}$$

    where the second equality follows from the identity $\mathbb{E}\,Y_i = \mu_i(\boldsymbol{\beta})$.
  - The last term is an invertible $(m \times m)$–matrix because we assumed that the design matrix $\mathbf{X}$ has full column rank and that $(d\mu_i/d\eta_i)(\boldsymbol{\beta}) \neq 0$ for each $i = 1, \ldots, n$.
- Therefore, instead of (43) the following iteration equation is considered:

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k + \big(\mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\mathbf{X}\big)^{-1} \left(\mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\Big(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\Big)(\widehat{\boldsymbol{\beta}}_k)\big(\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_k)\big)\right), \tag{44}$$

  where

$$\mathbf{Z}(\boldsymbol{\beta}) = \mathbf{V}^{-1}(\boldsymbol{\beta}) \left(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}(\boldsymbol{\beta})\right)^2 \qquad \text{and} \qquad \Big(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\Big)(\boldsymbol{\beta}) = \Big(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}\Big)^{-1}(\boldsymbol{\beta})\,.$$

- In the case of a natural link function it follows from Lemma 4.2 that

$$\frac{d\mu_i}{d\eta_i} \;=\; b^{(2)}(\theta_i) \;=\; \frac{1}{\tau^2}\,\sigma_i^2 \qquad \text{or} \qquad \mathbf{Z}(\boldsymbol{\beta}) \;=\; \frac{1}{\tau^4}\,\mathbf{V}(\boldsymbol{\beta})\,.$$

- Then the iteration equation (44) has the form:

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k + \tau^2\big(\mathbf{X}^\top \mathbf{V}(\widehat{\boldsymbol{\beta}}_k)\mathbf{X}\big)^{-1}\Big(\mathbf{X}^\top\big(\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_k)\big)\Big)\,.$$

Remark

- If the random sample $\mathbf{Y}$ in (44) is replaced by the so–called *pseudorandom variable*

$$\mathbf{Y}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta} + \Big(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\Big)(\boldsymbol{\beta})\big(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\big),$$

  the iteration equation (44) can be written in the following form:

$$\big(\mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\mathbf{X}\big)\widehat{\boldsymbol{\beta}}_{k+1} = \mathbf{X}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)\mathbf{Y}(\widehat{\boldsymbol{\beta}}_k)\,.$$

- This equation can be considered as a *weighted normal equation* for $\widehat{\boldsymbol{\beta}}_{k+1}$ with respect to the pseudo-random sample $\mathbf{Y}(\widehat{\boldsymbol{\beta}}_k)$, where the weights, i.e., the entries of the diagonal matrix $\mathbf{Z}(\widehat{\boldsymbol{\beta}}_k)$ also depend on the k-th iteration $\widehat{\boldsymbol{\beta}}_k$.

4.3.4 Asymptotic Normality of ML Estimators; Asymptotic Tests

- The notion of the convergence in distribution of random vectors is defined as follows.
  - Let $m \in \mathbb{N}$ be an arbitrary natural number and let $\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \ldots : \Omega \to \mathbb{R}^m$ be arbitrary random vectors. We say that $\{\mathbf{Z}_n\}$ converges *in distribution* to $\mathbf{Z}$ if

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{Z}_n \leq \mathbf{x}) = \mathbb{P}(\mathbf{Z} \leq \mathbf{x}) \tag{45}$$

    for each $\mathbf{x} \in \mathbb{R}^m$ with $\mathbb{P}(\mathbf{Z} = \mathbf{x}) = 0$. Notation: $\mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Z}$.

- Now, we discuss asymptotic (distributional) properties of maximum–likelihood estimators $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ or asymptotic tests if the sample size n tends to infinity.
  - Here we only consider the case of the natural link function $g : G \to \mathbb{R}$
  - and index the random sample $\mathbf{Y}$, the loglikelihood function $\log L(\mathbf{Y}, \boldsymbol{\beta})$, the score vector $\mathbf{U}(\boldsymbol{\beta})$, the Fisher–information matrix $\mathbf{I}(\boldsymbol{\beta})$ and the ML estimator $\widehat{\boldsymbol{\beta}}$ each with n.

1. *Asymptotic distributional properties*

   Under certain conditions (cf. Section VII.2.6 in Pruscha (2000)) one can show that: For each $\boldsymbol{\beta} \in \mathbb{R}^m$ with $\mathbf{x}_i^\top \boldsymbol{\beta} \in \Theta$ for $i = 1, 2, \ldots$ there exists

   - a consistent ML estimator $\widehat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$, i.e., for each $\varepsilon > 0$ it holds that

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}}\big(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| \leq \varepsilon, \mathbf{U}_n(\widehat{\boldsymbol{\beta}}_n) = \mathbf{o}\big) \;=\; 1\,, \tag{46}$$

   - a sequence $\{\boldsymbol{\Gamma}_n\}$ of invertible $(m \times m)$–matrices, which can depend on $\boldsymbol{\beta}$ and for which it holds that $\lim_{n\to\infty} \boldsymbol{\Gamma}_n = \mathbf{0}$,
   - as well as a symmetric and positive definite $(m \times m)$–matrix $\mathbf{K}(\boldsymbol{\beta})$, such that

$$\lim_{n\to\infty} \boldsymbol{\Gamma}_n^\top \mathbf{I}_n(\boldsymbol{\beta}) \boldsymbol{\Gamma}_n = \mathbf{K}^{-1}(\boldsymbol{\beta}) \tag{47}$$

   and

$$\boldsymbol{\Gamma}_n^{-1}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(\mathbf{o}, \mathbf{K}(\boldsymbol{\beta})\big) \qquad \text{or} \qquad 2\big(\log L_n(\mathbf{Y}_n, \widehat{\boldsymbol{\beta}}_n) - \log L_n(\mathbf{Y}_n, \boldsymbol{\beta})\big) \xrightarrow{\mathrm{d}} \chi^2_m\,. \tag{48}$$

2. *Asymptotic tests*

   - For large n the test statistic

$$T_n = 2\big(\log L_n(\mathbf{Y}_n, \widehat{\boldsymbol{\beta}}_n) - \log L_n(\mathbf{Y}_n, \boldsymbol{\beta}_0)\big)$$

     can be considered for the construction of an asymptotic test for the pair of hypotheses

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad \text{vs.} \qquad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

     Because of (48), H_0 is rejected if $T_n > \chi^2_{m,1-\alpha}$.
   - The null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{o}$ is particularly interesting. If it is rejected, more specific hypotheses can be tested, e.g., for each $i = 1, \ldots, m$ the hypothesis $H_0 : \beta_i = 0$.

Remark

- If $\mathbf{I}_n(\boldsymbol{\beta})$ is positive definite for each sufficiently large n and if

$$\lim_{n\to\infty} \mathbf{I}_n^{-1}(\boldsymbol{\beta}) = \mathbf{0}\,, \tag{49}$$

  then we can put $\boldsymbol{\Gamma}_n = \mathbf{I}_n^{-1/2}$ in (47) and (48), which implies that $\mathbf{K}(\boldsymbol{\beta})$ is the identity matrix.
- In (37) we have already shown that in the logistic regression model it holds that $\mathbf{I}_n(\boldsymbol{\beta}) = \mathbf{X}^\top \mathrm{diag}\big(\pi_i(1-\pi_i)\big)\mathbf{X}$.
  - Since we assume that $0 < \pi_i(\boldsymbol{\beta}) < 1$ for each $i = 1, 2, \ldots$ and that the design matrix $\mathbf{X}$ has full column rank, the matrix $\mathbf{I}_n(\boldsymbol{\beta})$ is positive definite in this case.
  - If furthermore $\inf_{i\geq 1} \pi_i(1-\pi_i) > 0$ and if the entries x_{ij} of the design matrix $\mathbf{X}$ are chosen in such a way that $\lim_{n\to\infty}\big(\mathbf{X}^\top\mathbf{X}\big)^{-1} = \mathbf{0}$, then (49) also holds.
- Now, let $\mathbf{K}(\boldsymbol{\beta})$ be the identity matrix. Because of (47) and (48), $H_0 : \beta_i = 0$ is rejected if

$$\frac{\big|\big(\widehat{\boldsymbol{\beta}}_n\big)_i\big|}{\sqrt{\big(\mathbf{I}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\big)_{ii}}} \;>\; z_{1-\alpha/2}\,, \tag{50}$$

  where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$–quantile of the $\mathrm{N}(0,1)$–distribution.

4.4 Weighted LS Estimator for Categorical Regression

Instead of the maximum-likelihood approach to estimate the parameter vector $\boldsymbol{\beta}$, discussed in Section 4.3, we now consider a *weighted LS estimator* for $\boldsymbol{\beta}$ for the categorical regression model.

4.4.1 Estimation of the Expectation Vector

Recall (cf. Section 4.2.2): In the binary categorical regression model all sample variables $Y_1, \ldots, Y_n$ are Bernoulli–distributed, i.e., they can only take the values 0 and 1 with positive probability.

- We use the notation

$$\pi_i = \mathbb{P}(Y_i = 1) \qquad \Big(= \mu_i \;=\; \mathbb{E}\, Y_i\Big) \qquad \forall\, i = 1, \ldots, n\,,$$

  where it is assumed that $0 < \pi_i < 1$ for each $i = 1, \ldots, n$.
- In this case the probabilities $\pi_1, \ldots, \pi_n$ are linked to the parameter vector $\boldsymbol{\beta}$ by using a link function $g : (0,1) \to \mathbb{R}$, i.e.,

$$\big(g(\pi_1), \ldots, g(\pi_n)\big)^\top = \mathbf{X}\boldsymbol{\beta}\,. \tag{51}$$

- In order to estimate the vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ or $\mathbf{g}(\boldsymbol{\pi}) = \big(g(\pi_1), \ldots, g(\pi_n)\big)^\top$, we assume that we are able to observe $n_i > 0$ independent and identically distributed "copies" $Y_{i1}, \ldots, Y_{in_i}$ of Y_i for each $i = 1, \ldots, n$. The total sample size is then equal to $\sum_{i=1}^n n_i$.
- For each $i = 1, \ldots, n$

$$\widehat{\pi}_i \;=\; \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \tag{52}$$

  is a natural estimator for π_i.
- This leads to the estimators $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_n)^\top$ or $\mathbf{g}(\widehat{\boldsymbol{\pi}}) = \big(g(\widehat{\pi}_1), \ldots, g(\widehat{\pi}_n)\big)^\top$ for $\boldsymbol{\pi}$ or $g(\boldsymbol{\pi})$, respectively.

One can easily see that the estimator $\widehat{\boldsymbol{\pi}}$ is unbiased for $\boldsymbol{\pi}$ and that its covariance matrix $\mathbf{K}(\widehat{\boldsymbol{\pi}}) = \big(\mathrm{Cov}\,(\widehat{\pi}_i, \widehat{\pi}_j)\big)$ has the following form.

Lemma 4.3 *It holds that*

$$\mathbb{E}\,\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}\,, \qquad \mathrm{Var}\,\widehat{\pi}_i = \pi_i(1-\pi_i)/n_i \tag{53}$$

and

$$\mathbf{K}(\widehat{\boldsymbol{\pi}}) = \mathrm{diag}\big(\mathrm{Var}\,\widehat{\pi}_i\big)\,. \tag{54}$$

Proof The statement follows directly from the fact that the random variables $n_1\widehat{\pi}_1, \ldots, n_n\widehat{\pi}_n$ are independent and binomially distributed with $n_i\widehat{\pi}_i \sim \mathrm{B}(n_i, \pi_i)$ for each $i = 1, \ldots, n$. □

Furthermore, the following *central limit theorem* implies that the estimator $\mathbf{g}(\widehat{\boldsymbol{\pi}}) = \big(g(\widehat{\pi}_1), \ldots, g(\widehat{\pi}_n)\big)^\top$ is asymptotically normally distributed.

Theorem 4.3 *If $n_i \to \infty$ for each $i = 1, \ldots, n$, such that*

$$\frac{\sum_{j=1}^n n_j}{n_i} \to \lambda_i \in [1, \infty) \qquad \forall\, i = 1, \ldots, n\,, \tag{55}$$

then it holds that

$$\Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2}\big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{g}(\boldsymbol{\pi})\big) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{o}, \mathbf{K})\,, \tag{56}$$

where

$$\mathbf{K} = \mathrm{diag}(\alpha_i) \qquad \textit{and} \qquad \alpha_i = \lambda_i (g^{(1)}(\pi_i))^2 \pi_i(1-\pi_i)\,. \tag{57}$$

Proof

- Since we assume that the link function $g : (0,1) \to \mathbb{R}$ is twice continuously differentiable, by Taylor series expansion we get that for each $i = 1, \ldots, n$

$$g(\widehat{\pi}_i) - g(\pi_i) \;=\; g^{(1)}(\pi_i)\big(\widehat{\pi}_i - \pi_i\big) + g^{(2)}(Z_i)\big(\widehat{\pi}_i - \pi_i\big)^2 \;=\; g^{(1)}(\pi_i)\big(\widehat{\pi}_i - \pi_i\big) + R_i\,,$$

  where $R_i = g^{(2)}(Z_i)\big(\widehat{\pi}_i - \pi_i\big)^2$ and $Z_i : \Omega \to \mathbb{R}$ is a random variable taking values between $\widehat{\pi}_i$ and π_i.
- From the central limit theorem for sums of independent and identically distributed random variables (cf. Theorem WR-5.16) it follows that

$$n_i^{1/2}\big(\widehat{\pi}_i - \pi_i\big) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \pi_i(1-\pi_i)) \qquad \forall\, i = 1, \ldots, n\,. \tag{58}$$

- Since $\widehat{\pi}_i - \pi_i \to 0$ and therefore also $Z_i - \pi_i \to 0$ or $g^{(2)}(Z_i) \to g^{(2)}(\pi_i)$ with probability 1, it also holds that $n_i^{1/2} R_i \xrightarrow{\mathrm{P}} 0$ or

$$\Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} R_i = \left(\frac{\sum_{j=1}^n n_j}{n_i}\right)^{1/2} n_i^{1/2} R_i \xrightarrow{\mathrm{P}} 0\,.$$

- Altogether, Slutsky's theorem (cf. Theorems WR-5.9 und WR-5.11) implies

$$\begin{aligned}\Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2}\Big(g(\widehat{\pi}_i) - g(\pi_i)\Big) \;&=\; \left(\frac{\sum_{j=1}^n n_j}{n_i}\right)^{1/2} g^{(1)}(\pi_i) n_i^{1/2}\big(\widehat{\pi}_i - \pi_i\big) + \Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} R_i \\ &\xrightarrow{\mathrm{d}} \; \mathrm{N}\big(0, \lambda_i (g^{(1)}(\pi_i))^2 \pi_i(1-\pi_i)\big)\,.\end{aligned}$$

- As the random variables $g(\widehat{\pi}_1) - g(\pi_1), \ldots, g(\widehat{\pi}_i) - g(\pi_i)$ are independent, the statement is proved. □

4.4.2 Asymptotic Normality of the LS–Estimator

The following approach for the estimation of the parameter vector $\boldsymbol{\beta}$ is motivated by the form of the asymptotic covariance matrix $\mathbf{K}$ in Theorem 4.3.

- In a similar way as is Section 2.1 we now consider the method of least squares for getting an estimator $\widehat{\boldsymbol{\beta}}$ for the unknown regression coefficients $\beta_1, \ldots, \beta_m$.
- A random vector $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)^\top$ shall be determined, such that the *weighted squared error*

$$e(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\big(g(\widehat{\pi}_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\big)^2}{\widehat{\sigma}_{ii}^2} \tag{59}$$

  gets minimal for $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$, where $\widehat{\sigma}_{ii}^2 = \big(\sum_{j=1}^{n} n_j/n_i\big)(g^{(1)}(\widehat{\pi}_i))^2 \widehat{\pi}_i(1-\widehat{\pi}_i)$ and it is assumed that the weights $\widehat{\sigma}_{ii}^2$ are positive.

Remark

- The weighted sum $e(\boldsymbol{\beta})$ of the squared residuals $\big(g(\widehat{\pi}_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\big)^2$ in (59) can be written as follows: By using the notation $\widehat{\mathbf{K}} = \operatorname{diag}\big(\widehat{\sigma}_{ii}^2\big)$ it holds that

$$e(\boldsymbol{\beta}) = \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\big)^\top \widehat{\mathbf{K}}^{-1} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\big)\,. \tag{60}$$

- In the same way as in the proof of Theorem 2.1 one can show that the weighted squared error $e(\boldsymbol{\beta})$ is minimal if and only if $\boldsymbol{\beta}$ is a solution of the following normal equation:

$$\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{g}(\widehat{\boldsymbol{\pi}})\,. \tag{61}$$

- Since the matrix $\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X}$ is invertible, (61) has the uniquely determined solution

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{g}(\widehat{\boldsymbol{\pi}})\,. \tag{62}$$

Now we show that the weighted LS–estimator $\widehat{\boldsymbol{\beta}}$ in (62) is asymptotically normally distributed if the (sub–) sample sizes n_i grow unboundedly for each $i = 1, \ldots, n$.

Here we need the following vectorial versions of Slutsky's theorem (cf. Theorems WR–5.9 and WR–5.11) and of the "continuous mapping theorem" (cf. Theorem WR–5.12).

Lemma 4.4

- *Let $m \in \mathbb{N}$, let $\mathbf{Y}, \mathbf{Y}_n, \mathbf{Z}_n : \Omega \to \mathbb{R}^m$ be arbitrary random vectors over the very same probability space and let $c \in \mathbb{R}^m$.*
- *If $\mathbf{Y}_n \xrightarrow{\mathrm{d}} \mathbf{Y}$ and $\mathbf{Z}_n \xrightarrow{\mathrm{d}} c$, then $\mathbf{Y}_n + \mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Y} + c$ and $\mathbf{Y}_n^\top \mathbf{Z}_n \xrightarrow{\mathrm{d}} c^\top \mathbf{Y}$.*

Lemma 4.5

- *Let $m \in \mathbb{N}$, let $\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \ldots : \Omega \to \mathbb{R}^m$ be arbitrary random vectors and let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be a continuous function.*
- *Then it holds that $\varphi(\mathbf{Z}_n) \xrightarrow{\mathrm{d}} \varphi(\mathbf{Z})$ provided that $\mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Z}$.*

The *proofs* of Lemmas 4.4 and 4.5 are similar to the proofs of Theorems WR–5.9, WR–5.11 and WR–5.12. Therefore, they are omitted.

Theorem 4.4 *If $n_i \to \infty$ for each $i = 1, \ldots, n$, such that*

$$\frac{\sum_{j=1}^n n_j}{n_i} \to \lambda_i \in [1, \infty) \qquad \forall\, i = 1, \ldots, n\,, \tag{63}$$

then it holds that

$$\Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} \big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(\mathbf{o}, \big(\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X}\big)^{-1}\big)\,, \tag{64}$$

where $\mathbf{K} = \mathrm{diag}(\alpha_i)$ is the diagonal matrix considered in Theorem 4.3.

Proof

- It follows from the definition of $\widehat{\boldsymbol{\beta}}$ in (62) that

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{g}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{\beta} = (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\big) \\ &= (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{g}(\boldsymbol{\pi})\big)\,, \end{aligned}$$

  where in the last equality we used that $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$; cf. (51).
- In Theorem 4.3 we have already shown that

$$\Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{g}(\boldsymbol{\pi})\big) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{o}, \mathbf{K})\,,$$

  where the asymptotic covariance matrix $\mathbf{K}$ is given in (57).
- Moreover, it holds that $\widehat{\mathbf{K}} \xrightarrow{\mathrm{P}} \mathbf{K}$ if $n_i \to \infty$ for each $i = 1, \ldots, n$.
- Altogether with
  - Slutsky's theorem (cf. Lemma 4.4),
  - the "continuous mapping theorems" (cf. Lemma 4.5) as well as
  - Theorem 1.3 about linear transformations of normally distributed random vectors,

  it follows that

$$\begin{aligned} \Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} \big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) &= \Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} (\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{g}(\boldsymbol{\pi})\big) \\ &= \Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} \underbrace{(\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \big((\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1}\big)^{-1}}_{\xrightarrow{\mathrm{P}}\, \mathbf{I}} \\ &\qquad \times (\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1} \big(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{g}(\boldsymbol{\pi})\big) \\ &\xrightarrow{\mathrm{d}} \mathrm{N}\big(\mathbf{o}, \big((\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1}\big)\, \mathbf{K}\, \big((\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{-1}\big)^\top\big) \\ &= \mathrm{N}\big(\mathbf{o}, \big(\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X}\big)^{-1}\big)\,. \end{aligned}$$

□

Remark

- If $n_1 + \ldots + n_n$ is a large number, the test statistic

$$T = \Big(\sum\nolimits_{j=1}^n n_j\Big)^{1/2} \widehat{\beta}_i \Big/ \sqrt{\widetilde{k}_{ii}}$$

  can be considered for the construction of an asymptotic test for the pair of hypotheses $H_0 : \beta_i = 0$ vs. $H_1 : \beta \neq 0$, where $\widetilde{k}_{ii}$ is the i-th diagonal entry of the matrix $\widetilde{\mathbf{K}} = \big(\mathbf{X}^\top \widehat{\mathbf{K}}^{-1} \mathbf{X}\big)^{-1}$.
- Because of Theorem 4.4, H_0 is rejected if $|T| > z_{1-\alpha/2}$.

4.4.3 Evaluation of the Goodness of Fit

An important problem is the choice of a suitable design matrix $\mathbf{X}$ in order to fit the model $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$ in the best possible way to given data. It is possible to answer this question by using the following result.

Theorem 4.5 *Under the conditions of Theorem* 4.3 *it holds that*

$$\sum_{j=1}^{n} n_j \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^\top \widehat{\mathbf{K}}^{-1} \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \xrightarrow{\mathrm{d}} \chi^2_{n-m} \,. \tag{65}$$

Proof

- In Theorem 4.3 we have already shown that the random vector $\left(\sum_{j=1}^{n} n_j\right)^{1/2} \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\boldsymbol{\beta}\right)$ is approximately $\mathrm{N}(\mathbf{o}, \mathbf{K})$ distributed.
- Since $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$ and $\widehat{\mathbf{K}} \xrightarrow{\mathrm{P}} \mathbf{K}$, the statement follows with
  - Slutsky's Theorem (cf. Lemma 4.4),
  - the "continuous mapping theorems" (cf. Lemma 4.5) as well as
  - Theorem 1.9 about quadratic forms of normally distributed random vectors. □

Remark

- Because of Theorem 4.5 the quantity $\sum_{j=1}^{n} n_j \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^\top \widehat{\mathbf{K}}^{-1} \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ can be seen as a measure for the goodness of fit of the model $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$ to given data.
- The goodness of fit is appraised as sufficiently good if

$$\sum_{j=1}^{n} n_j \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^\top \widehat{\mathbf{K}}^{-1} \left(\mathbf{g}(\widehat{\boldsymbol{\pi}}) - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \; < \; \chi^2_{n-m,1-\alpha} \,.$$

- On the other hand, the dimensions of $\mathbf{X}$ should be as small as possible, which means in particular that for each $i = 1, \ldots, m$ the null hypothesis of the test $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ should be clearly rejected.

5 Goodness–of–Fit Tests

- In this chapter the sample variables are denoted by $X_1, \ldots, X_n$, where we will assume from now on that $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ is a sequence of independent *and* identically distributed random variables.
- The assumptions, we made so far regarding the distribution P of the sample variables $X_1, \ldots, X_n$, either have been *strictly qualitative* (discrete or absolutely continuous distribution) or *parametric*, where in the latter case it has been assumed
  - that P belongs to a parametric family $\{P_\theta, \theta \in \Theta\}$ of distributions with $\Theta \subset \mathbb{R}^m$ for some integer $m \geq 1$,
  - and that merely the parameter vector $\theta = (\theta_1, \ldots, \theta_m)^\top$ or some of its components, respectively, are unknown.
- In the following, we discuss so–called *goodness–of–fit tests*.
  - To begin with, we consider a test to verify the hypothesis $H_0 : P = P_0$ that the distribution P of the sample variables is equal to a given (hypothetical) distribution P_0.
  - Afterwards, we construct tests to check if P belongs to a given (parametric) *class* of distributions $\{P_\theta, \theta \in \Theta\}$.

5.1 Kolmogorov–Smirnov Test

5.1.1 Empirical Distribution Function; KS Test Statistic

- There are different tests suggested in literature to verify the hypothesis $H_0 : P = P_0$ that the distribution P of the independent and identically distributed random variables $X_1, \ldots, X_n$ is equal to a given distribution P_0.
- This kind of null hypothesis is considered for the *Kolmogorov–Smirnov test*, which is based on the analysis of the *empirical distribution function* $\widehat{F}_n : \mathbb{R} \times \mathbb{R}^n \to [0,1]$, introduced in Section I–1.5, where

$$\widehat{F}_n(t; x_1, \ldots, x_n) = \frac{\#\{i : 1 \leq i \leq n,\, x_i \leq t\}}{n}\,. \tag{1}$$

- The sample function $T_n : \mathbb{R}^n \to [0, \infty)$ is considered with

$$T_n(x_1, \ldots, x_n) = \sqrt{n}\, \sup_{t \in \mathbb{R}} |\, \widehat{F}_n(t; x_1, \ldots, x_n) - F_0(t)|\,. \tag{2}$$

- In Section I–1.5.3 we have already shown that the distribution of the *KS test statistic* $T_n(X_1, \ldots, X_n)$ does not depend on P_0 if it is assumed that the distribution function $F_0 : \mathbb{R} \to [0,1]$, which corresponds to P_0, is continuous, cf. Theorem I–1.19.
- Let $s_{n,1-\alpha}$ be the $(1-\alpha)$–quantile of the distribution of $T_n(X_1, \ldots, X_n)$ under an arbitrary continuous distribution function F_0. The Kolmogorov–Smirnov test rejects the null hypothesis $H_0 : P = P_0$ if

$$T_n(x_1, \ldots, x_n) > s_{n,1-\alpha}\,. \tag{3}$$

Remark

- The quantiles $s_{n,1-\alpha}$ can be determined using Monte Carlo simulation, where the distribution function F_0 of the standard–uniform distribution on $[0,1]$ can be taken as a basis, cf Corollary I–1.3.
- If it is not assumed that F_0 is continuous, then the decision rule, considered in (3), provides a test whose level can be smaller than α.
- However, if it is possible to determine the quantile $s'_{n,1-\alpha}$ of $T_n(X_1, \ldots, X_n)$ under F_0, e.g., using MC simulation, then $T_n(x_1, \ldots, x_n) > s'_{n,1-\alpha}$ is a test, which taps the full level α even if F_0 is discontinuous.

5.1.2 Asymptotic Distribution

We now analyze the asymptotic distribution of the KS test statistic $T_n(X_1, \ldots, X_n)$, introduced in (2), as $n \to \infty$. To begin with, we provide some auxiliary tools.

In particular we need the following *continuity theorem* for characteristic functions of random vectors, which is a multidimensional generalization of Theorem WR–5.20 and which we will state without proof.

Lemma 5.1 *Let $m \in \mathbb{N}$ and let $\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2, \ldots : \Omega \to \mathbb{R}^m$ be arbitrary random vectors with characteristic functions $\varphi_{\mathbf{Z}_n}$ and $\varphi_{\mathbf{Z}}$. It holds that $\mathbf{Z}_n \xrightarrow{\mathrm{d}} \mathbf{Z}$ if and only if*

$$\lim_{n\to\infty} \varphi_{\mathbf{Z}_n}(\mathbf{t}) = \varphi_{\mathbf{Z}}(\mathbf{t}) \qquad \forall\, \mathbf{t} \in \mathbb{R}^m\,. \tag{4}$$

Moreover, we need a *multivariate central limit theorem* for sums of independent and identically distributed random vectors,

- whose proof can be reduced, using Lemma 5.1, to the corresponding central limit theorem for real–valued random variables (cf. Theorem WR–5.16).
- In literature this approach is sometimes called the *Cramèr–Wold device*.

Lemma 5.2

- *Let $m \in \mathbb{N}$ and let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots : \Omega \to \mathbb{R}^m$ be a sequence of independent and identically distributed random vectors with expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ and covariance matrix $\mathbf{K}$.*
- *Then it holds that*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{(\mathbf{Z}_1 + \ldots + \mathbf{Z}_n) - n\boldsymbol{\mu}}{\sqrt{n}} \le \mathbf{x}\Big) = \Phi_{\mathbf{K}}(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbb{R}^m\,, \tag{5}$$

where $\Phi_{\mathbf{K}} : \mathbb{R}^m \to [0,1]$ denotes the distribution function of the $\mathrm{N}(\mathbf{o}, \mathbf{K})$–distribution.

Proof

- Let $\mathbf{Z}_n = (Z_{n1}, \ldots, Z_{nm})^\top$. Because of Lemma 5.1 the convergence in distribution stated in (5) is equivalent to

$$\lim_{n\to\infty} \varphi_n(\mathbf{t}) = \varphi(\mathbf{t}) \qquad \forall\, \mathbf{t} \in \mathbb{R}^m, \tag{6}$$

  - where $\varphi_n(\mathbf{t})$ is the characteristic function of $(\mathbf{Z}_1 + \ldots + \mathbf{Z}_n - n\boldsymbol{\mu})/\sqrt{n}$ with

$$\varphi_n(\mathbf{t}) = \mathbb{E}\,\exp\Big(\mathrm{i} \sum_{j=1}^{m} t_j\, \frac{(Z_{1j} + \ldots + Z_{nj}) - n\mu_j}{\sqrt{n}}\Big)$$

  - and $\varphi(\mathbf{t})$ is the characteristic function of the $\mathrm{N}(\mathbf{o}, \mathbf{K})$–distribution with

$$\varphi(\mathbf{t}) = \exp\Big(-\frac{1}{2}\,\mathbf{t}^\top \mathbf{K} \mathbf{t}\Big)\,. \tag{7}$$

- Furthermore, one can easily see that

$$\varphi_n(\mathbf{t}) = \mathbb{E}\,\exp\Big(\mathrm{i} \sum_{k=1}^{n} \frac{\sum_{j=1}^{m} t_j (Z_{kj} - \mu_j)}{\sqrt{n}}\Big) \qquad \forall\, \mathbf{t} = (t_1, \ldots, t_m)^\top \in \mathbb{R}^m \tag{8}$$

  and

$$\mathbb{E}\Big(\sum_{j=1}^{m} t_j (Z_{kj} - \mu_j)\Big) = 0\,, \qquad \mathrm{Var}\Big(\sum_{j=1}^{m} t_j (Z_{kj} - \mu_j)\Big) = \mathbf{t}^\top \mathbf{K} \mathbf{t} \qquad \forall\, k \in \mathbb{N}\,. \tag{9}$$

- If $\mathbf{t}^\top \mathbf{K}\mathbf{t} = 0$, then it follows from (9)
  - that $\sum_{j=1}^m t_j(Z_{kj} - \mu_j) = 0$ with probability 1 for arbitrary $k = 1, \ldots, n$ and $n \geq 1$.
  - This and (7) – (8) imply that $\varphi_n(\mathbf{t}) = 1 = \varphi(\mathbf{t})$ for each $n \geq 1$, i.e., (6).
- Now let $\mathbf{t}^\top \mathbf{K}\mathbf{t} > 0$.
  - From (8) it follows that $\varphi_n(\mathbf{t})$ is equal to the value of the characteristic function of the real–valued random variable $\sum_{k=1}^n \sum_{j=1}^m t_j(Z_{kj} - \mu_j)/\sqrt{n}$ at 1.
  - Moreover, it follows from (7) that $\varphi(\mathbf{t})$ is the value of the characteristic function of the one–dimensional normal distribution $\mathrm{N}(0, \mathbf{t}^\top \mathbf{K}\mathbf{t})$ at 1.
  - On the other hand, Theorem WR–5.16, i.e., the (1–dimensional) central limit theorem for sums of independent and identically distributed (real–valued) random variables, implies that for $n \to \infty$

$$\sum_{k=1}^{n} \frac{\sum_{j=1}^{m} t_j(Z_{kj} - \mu_j)}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathbf{t}^\top \mathbf{K}\mathbf{t})\,. \tag{10}$$

  - This, (7) – (8) and Theorem WR-5.20, i.e., the continuity theorem for characteristic functions of real–valued random variables, imply the validity of (6). □

The following limit theorem, already mentioned in Section I–1.5.3, provides an approximation formula for the distribution function of $T_n(X_1, \ldots, X_n)$ for a large sample size n.

Theorem 5.1 *Let the distribution function $F_0 : \mathbb{R} \to [0,1]$ be continuous. Assuming that $H_0 : P = P_0$ is true, it holds that*

$$\lim_{n\to\infty} \mathbb{P}\big(T_n(X_1, \ldots, X_n) \leq x\big) = K(x) \qquad \forall\, x \in \mathbb{R}\,,$$

where $K : \mathbb{R} \to [0,1]$ is the distribution function of the so–called Kolmogorov distribution with

$$K(x) = \begin{cases} 1 - 2\sum\limits_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2x^2)\,, & \text{if } x > 0, \\ 0\,, & \text{if } x \leq 0. \end{cases} \tag{11}$$

Proof

- We only sketch the idea of the proof since the full proof of Theorem 5.1 (cf., e.g., A. van der Vaart and J. Wellner (1996)) exceeds the scope of these lecture notes
  - as it requires profound tools from the theory of stochastic processes.
  - In particular, the term of convergence in distribution in *function spaces* as well as a so–called *functional* central limit theorem is needed,
  - which can be seen as an (infinite dimensional) generalization of the classical central limit theorems for sums of real–valued random variables (cf. Section WR–5.3) or of finite–dimensional random vectors (cf. Lemma 5.2).
- As the distribution of $T_n(X_1, \ldots, X_n)$ does not depend on F_0 (cf. Theorem I–1.19), we can w.l.o.g assume that F_0 is the distribution function of the uniform distribution on $[0,1]$, i.e., $F_0(t) = t$ for each $t \in [0,1]$.
  - In order to analyze the asymptotic distribution of $T_n(X_1, \ldots, X_n)$ for $n \to \infty$, we use the abbreviating notation

$$B_n(t) = \sqrt{n}\,\big(\,\widehat{F}_n(t; X_1, \ldots, X_n) - F_0(t)\big) \qquad \forall\, t \in [0,1]\,, \tag{12}$$

  - where the family of random variables $\{B_n(t),\, t \in [0,1]\}$ is a stochastic process, which is called an *empirical process* in literature.

- For arbitrary $t_1, \ldots, t_m \in [0,1]$ it then holds that

$$\sqrt{n}\big(B_n(t_1), \ldots, B_n(t_m)\big) = \sum_{i=1}^{n} \big(Y_i(t_1) - t_1, \ldots, Y_i(t_m) - t_m\big),$$

where

$$Y_i(t_j) = \begin{cases} 1\,, & \text{if } X_i \le t_j, \\ 0\,, & \text{if } X_i > t_j. \end{cases}$$

  - For each $n \ge 1$ the random vector $\sqrt{n}\big(B_n(t_1), \ldots, B_n(t_m)\big)$ can be written as a sum of n independent and identically distributed random vectors with expectation vector $\mathbf{o}$, whose covariance matrix $\mathbf{K} = (\sigma_{ij}^2)$ is given by $\sigma_{ij}^2 = \min\{t_i, t_j\} - t_i t_j$.
  - From Lemma 5.2 it now follows that for $n \to \infty$

$$\big(B_n(t_1), \ldots, B_n(t_m)\big) \xrightarrow{\mathrm{d}} \big(B(t_1), \ldots, B(t_m)\big)\,, \tag{13}$$

    where $\big(B(t_1), \ldots, B(t_m)\big)$ is a normally distributed random vector with $\big(B(t_1), \ldots, B(t_m)\big) \sim \mathrm{N}(\mathbf{o}, \mathbf{K})$.
  - This and the continuous mapping theorem for random vectors (cf. Lemma 4.5) imply that

$$\max_{i=1,\ldots,m} \sqrt{n}\big|\widehat{F}_n(t_i; X_1, \ldots, X_n) - F_0(t_i)\big| \xrightarrow{\mathrm{d}} \max_{i=1,\ldots,m} \big|B(t_i)\big|\,. \tag{14}$$

- It is easy to see that the distribution $\mathrm{N}(\mathbf{o}, \mathbf{K})$ of the random vector $\big(B(t_1), \ldots, B(t_m)\big)$
  - can be considered as the finite-dimensional distribution of the so–called *Brownian bridge process* $\{B(t),\, t \in [0,1]\}$ with $B(t) = X(t) - tX(1)$, where $\{X(t),\, t \in [0,1]\}$ is a (standard–) Wiener–process,
  - i.e., $\{X(t),\, t \in [0,1]\}$ is a stochastic process with continuous trajectories and independent increments, such that $X(0) = 0$ and $X(t_2) - X(t_1) \sim \mathrm{N}(0, t_2 - t_1)$ for arbitrary $t_1, t_2 \in [0,1]$ with $t_1 < t_2$, cf. Section 2.4 of the lecture notes "Elementare Wahrscheinlichkeitsrechnung und Statistik".
- Using the theory of convergence in distribution in function spaces as well as a corresponding functional central limit theorem, it is possible to show that not only the "finite–dimensional" convergences (13) and (14) hold but also

$$\big(B_n(t),\, t \in [0,1]\big) \xrightarrow{\mathrm{d}} \big(B(t),\, t \in [0,1]\big) \tag{15}$$

and

$$\max_{t \in [0,1]} \sqrt{n}\big|\widehat{F}_n(t; X_1, \ldots, X_n) - F_0(t)\big| \xrightarrow{\mathrm{d}} \max_{t \in [0,1]} \big|B(t)\big|\,. \tag{16}$$

- Furthermore, one can show that the distribution function of the maximum $\max_{t \in [0,1]}\big|B(t)\big|$ of the Brownian bridge $\{B(t),\, t \in [0,1]\}$ is given by (11). □

Remark

- Because of Theorem 5.1 the hypothesis $H_0 : P = P_0$ is rejected for a sufficiently large sample size (as a rule of thumb it holds that $n > 40$, cf. the remark at the end of Section I–1.5.3) if

$$T_n(x_1, \ldots, x_n) > \xi_{1-\alpha}\,,$$

- where $\xi_{1-\alpha}$ denotes the $(1-\alpha)$–quantile of the Kolmogorov–distribution, given in (11), i.e., $\xi_{1-\alpha}$ is a solution of the equation $K(\xi_{1-\alpha}) = 1 - \alpha$.

5.1.3 Pointwise and Uniform Consistency

In this section we consider some properties of the Kolmogorov–Smirnov test.

In order to show the (pointwise) consistency of the KS test, we need the *Glivenko–Cantelli theorem* (cf. Theorem I–1.18), i.e.,

$$\mathbb{P}_{F_0}\Big(\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\big|\widehat{F}_n(t;X_1,\ldots,X_n)-F_0(t)\big|=0\Big)=1\,. \tag{17}$$

Theorem 5.2 *Let the distribution function $F_0:\mathbb{R}\to[0,1]$ be continuous. Then the Kolmogorov–Smirnov test is pointwise consistent for each distribution function F of the sample variables with $F\neq F_0$, i.e., it holds that*

$$\lim_{n\to\infty}\mathbb{P}_F\big(T_n(X_1,\ldots,X_n)>s_{n,1-\alpha}\big)=1\,. \tag{18}$$

Proof

- From (17) it follows that for each $F\neq F_0$

$$\mathbb{P}_F\Big(\lim_{n\to\infty}\sup_{t\in\mathbb{R}}\big|\widehat{F}_n(t;X_1,\ldots,X_n)-F_0(t)\big|>0\Big)=1\,.$$

- This implies that $T_n(X_1,\ldots,X_n)\to\infty$ with probability 1 under $F\neq F_0$.
- Since $s_{n,1-\alpha}\to\xi_{1-\alpha}<\infty$ for $n\to\infty$, where $\xi_{1-\alpha}$ is the $(1-\alpha)$–quantile of the Kolmogorov distribution, given in (11), it also holds that $T_n(X_1,\ldots,X_n)-(s_{n,1-\alpha}-\xi_{1-\alpha})\xrightarrow{\text{a.s.}}\infty$ and therefore

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}_F\big(T_n(X_1,\ldots,X_n)>s_{n,1-\alpha}\big) &= \lim_{n\to\infty}\mathbb{P}_F\big(T_n(X_1,\ldots,X_n)-(s_{n,1-\alpha}-\xi_{1-\alpha})>\xi_{1-\alpha}\big)\\ &= \lim_{n\to\infty}\mathbb{P}_F\big(T_n(X_1,\ldots,X_n)>\xi_{1-\alpha}\big)\;=\;1\,.\end{aligned}$$

□

Remark

- As an strengthening of Theorem 5.2 one can show that the KS test is also *uniformly consistent* if the *Kolmogorov distance*

$$d_K(\Delta_n;F_0)=\inf_{F\in\Delta_n}\sup_{t\in\mathbb{R}}|F(t)-F_0(t)| \tag{19}$$

  of the family Δ_n of alternative distribution functions and the (hypothetical) distribution function F_0 does not converge too fast to 0 as the sample size n increases.
- In this context, we need the following strengthening of Glivenko–Cantelli's theorem, which is called *inequality of Dworetsky–Kiefer–Wolfowitz* in literature and which we state without proof.

Lemma 5.3 *For arbitrary $c>0$ and $n\geq 1$ it holds that*

$$\mathbb{P}_F\Big(\sup_{t\in\mathbb{R}}\big|\widehat{F}_n(t;X_1,\ldots,X_n)-F(t)\big|>c\Big)\leq C\exp\big(-2nc^2\big)\,, \tag{20}$$

where $C\leq 2$ is an universal constant, which does not depend on F.

Remark

- From Lemma 5.3 it follows that for each $\varepsilon>0$ there is a $c'>0$ which does not depend on F and which fulfills

$$\inf_{n\geq 1}\mathbb{P}_F\Big(\sqrt{n}\sup_{t\in\mathbb{R}}\big|\widehat{F}_n(t;X_1,\ldots,X_n)-F(t)\big|\leq c'\Big)\geq 1-\varepsilon\,. \tag{21}$$

- In order to see this, it is sufficient to choose the threshold value c in (20) for $\varepsilon \in (0,1)$ in such a way that $c = c'/\sqrt{n}$, where

$$c' = \sqrt{-\frac{1}{2}\,\log\Big(\frac{\varepsilon}{C}\Big)}\,.$$

- Since c' does not depend on F, it also holds that for each $\varepsilon > 0$ there is a $c' > 0$ such that

$$\inf_{n\geq 1} \mathbb{P}_{F_n}\Big(\sqrt{n}\,\sup_{t\in\mathbb{R}}\big|\widehat{F}_n(t;X_1,\ldots,X_n) - F_n(t)\big| \leq c'\Big) \geq 1-\varepsilon\,, \tag{22}$$

  where $\{F_n\}$ is an arbitrary sequence of distribution functions.

With these tools it is possible to show the uniform consistence property of the KS test for the case that the Kolmogorov distance $d_K(\Delta_n; F_0)$ of the family Δ_n of the alternative distribution functions and the (hypothetical) distribution function F_0 does not converge too fast to 0 as the sample size n increases.

Theorem 5.3 *If there is a sequence $\{\delta_n\}$ of positive numbers with $\delta_n \to \infty$ such that*

$$\sqrt{n}\,d_K(\Delta_n; F_0) \geq \delta_n \qquad \forall\, n \geq 1\,, \tag{23}$$

then it holds that

$$\lim_{n\to\infty} \inf_{F\in\Delta_n} \mathbb{P}_F\big(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\big) = 1\,. \tag{24}$$

Proof

- Let $\{\delta_n\}$ be a sequence of positive numbers with $\delta_n \to \infty$, which fulfills (23), and let $\{F_n\}$ be an arbitrary sequence of distribution functions such that for $n \geq 1$

$$F_n \in \Delta_n \qquad \text{and therefore} \qquad \sqrt{n}\,d_K(F_n; F_0) \geq \delta_n\,, \tag{25}$$

  where $d_K(F_n; F_0) = \sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)|$.
- It is sufficient to show that

$$\lim_{n\to\infty} \mathbb{P}_{F_n}\big(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\big) = 1\,. \tag{26}$$

  – From the triangle inequality it follows that

$$d_K(F_n, F_0) \leq d_K(F_n, \widehat{F}_n) + d_K(\widehat{F}_n, F_0)\,.$$

  – This and (25) imply that

$$T_n(X_1,\ldots,X_n) \geq \delta_n - \sqrt{n}\,d_K(F_n, \widehat{F}_n)\,.$$

  – Therefore, it holds that

$$\mathbb{P}_{F_n}\big(T_n(X_1,\ldots,X_n) > s_{n,1-\alpha}\big) \geq \mathbb{P}_{F_n}\big(\sqrt{n}\,d_K(F_n, \widehat{F}_n) < \delta_n - s_{n,1-\alpha}\big)\,. \tag{27}$$

- Since $s_{n,1-\alpha} \to \xi_{1-\alpha} < \infty$ and therefore $\delta_n - s_{n,1-\alpha} \to \infty$ for $n \to \infty$, formulas (22) and (27) imply the validity of (26) . □

Remark

- In particular, condition (23) is fulfilled if $d_K(\Delta_n; F_0) \geq \delta$ for each $n \geq 1$ and $\delta > 0$ is a constant, which does not depend on n.

- If $\sqrt{n}\; d_K(F_n; F_0) \geq \delta_n$ and $\delta_n > s_{n,1-\alpha}$, then (27) implies that for each $n \geq 1$ holds the following (non–asymptotic) lower threshold for the *power* of the KS test:

$$\mathbb{P}_{F_n}\big(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}\big) \geq 1 - 2\exp\Big(-2(\delta_n - s_{n,1-\alpha})^2\Big)\,, \tag{28}$$

  where we have obtained a lower bound of the right–hand side of the inequality in (27) by applying Lemma 5.3 (for $F = F_n$).
- However, notice that for a given (finite) sample size $n < \infty$ it is possible that the "rejection probability" $\mathbb{P}_{F_0}\big(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}\big)$ is smaller than α.

On the other hand, the (asymptotic) power of the KS test can become arbitrarily small, i.e., arbitrarily close to α, if the Kolmogorov distance $d_K(\Delta_n; F_0)$ of the family Δ_n of alternative distribution functions and the (hypothetical) distribution function F_0 converges sufficiently fast to 0 as the sample size n increases.

Theorem 5.4

- *Let $\{F_n\}$ be an arbitrary sequence of continuous distribution functions such that*

$$\lim_{n\to\infty} \sqrt{n}\; d_K(F_n; F_0) \;=\; 0\,. \tag{29}$$

- *Then it holds that*

$$\limsup_{n\to\infty} \mathbb{P}_{F_n}\big(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}\big) \leq \alpha\,. \tag{30}$$

Proof

- From the triangle inequality it follows that

$$\mathbb{P}_{F_n}\big(T_n(X_1, \ldots, X_n) > s_{n,1-\alpha}\big) \leq \mathbb{P}_{F_n}\big(\sqrt{n}\; d_K(\widehat{F}_n; F_n) + \sqrt{n}\; d_K(F_n; F_0) > s_{n,1-\alpha}\big)\,.$$

- From this and (29) the validity of (30) follows because the distribution of $\sqrt{n}\; d_K(\widehat{F}_n; F_n)$ under F_n does not depend on n. □

5.2 χ^2–Goodness–of–Fit Test

We now discuss an asymptotic goodness–of–fit test, where a test statistic is considered, which is approximately χ^2–distributed for a large sample size. However, in this context the hypothesis

$$H_0 : P = P_0 \qquad \Big(\text{versus} \quad H_1 : P \neq P_0\Big), \tag{31}$$

analyzed in Section 5.1, is usually not considered since we "coarsen" the model of the random sample $(X_1, \ldots, X_n)$ by use of aggregation.

5.2.1 Aggregation; Pearson–Statistic

- For a (sufficiently large) natural number r we partition the range of the random variables $X_1, \ldots, X_n$ into r classes $(a_1, b_1], \ldots, (a_r, b_r]$ with

$$-\infty \leq a_1 < b_1 = a_2 < b_2 = \ldots = a_r < b_r \leq \infty\,.$$

- Instead of the sample variables $X_1, \ldots, X_n$ we consider the "class sizes" $Z_1, \ldots, Z_r$, where

$$Z_j = \#\{i :\, 1 \leq i \leq n,\, a_j < X_i \leq b_j\} \qquad \forall\, j = 1, \ldots, r\,. \tag{32}$$

To begin with, we show that the random vector $(Z_1, \ldots, Z_r)$ has a multinomial distribution with parameters $n \geq 1$ and

$$\mathbf{p} = (p_1, \ldots, p_{r-1})^\top \in [0,1]^{r-1}\,, \qquad \text{where } p_j = \mathbb{P}(a_j < X_1 \leq b_j) \qquad \forall\, j = 1, \ldots, r-1\,.$$

Lemma 5.4 *For arbitrary natural numbers $k_1, \ldots, k_r \geq 0$ with $k_1 + \ldots + k_r = n$ it holds that*

$$\mathbb{P}(Z_1 = k_1, \ldots, Z_r = k_r) = \frac{n!}{k_1! \cdot \ldots \cdot k_r!}\, p_1^{k_1} \ldots p_r^{k_r}\,, \tag{33}$$

where $p_r = 1 - (p_1 + \ldots + p_{r-1})$.

Proof

- Since the random variables $X_1, \ldots, X_n$ are independent and identically distributed, it holds that

$$\mathbb{P}\big(X_1 \in (a_{i_1}, b_{i_1}], \ldots, X_n \in (a_{i_n}, b_{i_n}]\big) = \prod_{j=1}^{n} \mathbb{P}(a_{i_j} < X_1 \leq b_{i_j}) = p_1^{k_1} \ldots p_r^{k_r} \tag{34}$$

  for each sequence of intervals $(a_{i_1}, b_{i_1}], \ldots, (a_{i_n}, b_{i_n}]$, which contains k_1-times the interval $(a_1, b_1], \ldots,$ k_r-times the interval $(a_r, b_r]$.
- The statement (33) follows from summation of the probabilities, considered in (34), over all permutations of sequences $(a_{i_1}, b_{i_1}], \ldots, (a_{i_n}, b_{i_n}]$ of this kind. □

Remark

- We denote the multinomial distribution with the parameters $n \geq 1$ and $\mathbf{p} = (p_1, \ldots, p_{r-1})^\top \in [0,1]^{r-1}$ by $\mathrm{M}_{r-1}(n, \mathbf{p})$. It is easy to see that for $r = 2$ the multinomial distribution $\mathrm{M}_1(n, p_1)$ coincides with the binomial distribution $\mathrm{Bin}(n, p_1)$.
- Instead of analyzing the test problem (31), we verify the hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$ (against the alternative $H_1 : \mathbf{p} \neq \mathbf{p}_0$) for a given (hypothetical) parameter vector

$$\mathbf{p}_0 = (p_{01}, \ldots, p_{0,r-1})^\top \in (0,1)^{r-1} \qquad \text{with} \quad \sum_{i=1}^{r-1} p_{0i} < 1.$$

  – We thus partition the family Δ of all considered distributions of the sample variables $X_1, \ldots, X_n$ into the subsets

$$\Delta_0 = \{P :\ \mathbb{P}_P(a_j < X_1 \leq b_j) = p_{0j}\ \forall j = 1, \ldots, r-1\} \qquad \text{and} \qquad \Delta_1 = \Delta \setminus \Delta_0\,. \tag{35}$$

  – In this context we consider the sample function $T_n : \mathbb{R}^n \to [0, \infty)$ with

$$T_n(x_1, \ldots, x_n) = \sum_{j=1}^{r} \frac{\big(Z_j(x_1, \ldots, x_n) - np_{0j}\big)^2}{np_{0j}}\,, \tag{36}$$

  – where $Z_j(x_1, \ldots, x_n)$ denotes the number of the sample values $x_1, \ldots, x_n$, which belong to $(a_j, b_j]$.
- Assuming $H_0 : \mathbf{p} = \mathbf{p}_0$ is true, it holds that $\mathbb{E}\, Z_j(X_1, \ldots, X_n) = np_{0j}$ for each $j \in \{1, \ldots, r\}$.
  – Therefore, it makes sense to reject the hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$ if $T_n(x_1, \ldots, x_n)$ is significantly larger than 0.
  – In order to make a decision, we need knowledge of the distribution of the test statistic $T_n(X_1, \ldots, X_n)$, introduced in (36), which is called the *Pearson–statistic*.

5.2.2 Asymptotic Distribution

We show that $T_n(X_1,\ldots,X_n)$ converges in distribution to a χ^2–distribution with $r-1$ degrees of freedom if $n\to\infty$. This is the base of the *χ^2-goodness–of–fit test*, which has been introduced by Karl Pearson (1857–1936).

Theorem 5.5 *For each $P\in\Delta_0$ it holds that*

$$\lim_{n\to\infty}\mathbb{P}_P\big(T_n(X_1,\ldots,X_n)>\chi^2_{r-1,1-\alpha}\big)=\alpha\,,\qquad\forall\,\alpha\in(0,1)\,,\tag{37}$$

where $\chi^2_{r-1,1-\alpha}$ denotes the $(1-\alpha)$–quantile of the χ^2–distribution with $r-1$ degrees of freedom.

Proof

- In Lemma 5.4 we have shown that the random vector $\mathbf{Z}_n=(Z_{n1},\ldots,Z_{nr})^\top$, given in (32), where $Z_{nj}=Z_j(X_1,\ldots,X_n)$, has a multinomial distribution under $P\in\Delta_0$ with parameters $n\in\mathbb{N}$ and

$$\mathbf{p}_0=(p_{01},\ldots,p_{0,r-1})^\top\in[0,1]^{r-1}\,,\qquad\text{where }\ p_{0j}=\mathbb{P}_P(a_j<X_1\le b_j)\quad\forall\,j=1,\ldots,r-1\,.$$

  - This in particular implies that for arbitrary $i,j\in\{1,\ldots,r\}$

$$\mathbb{E}\,Z_{ni}=np_{0i}\,,\qquad\mathrm{Cov}\,(Z_{ni},Z_{nj})=\begin{cases}-np_{0i}p_{0j}\,, & \text{if } i\ne j,\\ np_{0i}(1-p_{0i})\,, & \text{if } i=j.\end{cases}\tag{38}$$

  - Moreover, it follows from (32) that $Z_j=\sum_{i=1}^n\mathbb{1}_{\{a_j<X_i\le b_j\}}$, i.e., $\mathbf{Z}_n$ can be written as a sum of n independent and identically distributed random vectors, where $\mathbb{1}_{\{a_j<X_i\le b_j\}}$ is the indicator of the event $\{a_j<X_i\le b_j\}$.

- With the notation

$$\mathbf{Z}'_n=\Big(\frac{Z_{n1}}{\sqrt{n}}-\sqrt{n}p_{01},\ \ldots,\ \frac{Z_{n,r-1}}{\sqrt{n}}-\sqrt{n}p_{0,r-1}\Big)^\top\tag{39}$$

  it therefore follows from Lemma 5.2 that for $n\to\infty$ it holds that

$$\mathbf{Z}'_n\xrightarrow{\mathrm{d}}\mathbf{Z}'\sim\mathrm{N}(\mathbf{o},\mathbf{K})\,,\tag{40}$$

  - where the $(r-1)$–dimensional random vector $\mathbf{Z}'$ has a (nondegenerate) multivariate normal distribution,
  - whose covariance matrix $\mathbf{K}=(\sigma^2_{ij})$ is given by

$$\sigma^2_{ij}=\begin{cases}-p_{0i}p_{0j}\,, & \text{if } i\ne j,\\ p_{0i}(1-p_{0i})\,, & \text{if } i=j.\end{cases}\tag{41}$$

- It is easy to see that $\mathbf{K}$ is invertible, where the entries a_{ij} of the inverse matrix $\mathbf{A}=\mathbf{K}^{-1}$ are given by

$$a_{ij}=\begin{cases}\dfrac{1}{p_{0r}}\,, & \text{if } i\ne j,\\ \dfrac{1}{p_{0i}}+\dfrac{1}{p_{0r}}\,, & \text{if } i=j.\end{cases}\tag{42}$$

  - From (40) and the properties of linear transformations of normally distributed random vectors (cf. Theorem 1.3) it now follows with Lemma 4.5 that $\mathbf{A}^{1/2}\mathbf{Z}'_n\xrightarrow{\mathrm{d}}\mathrm{N}(\mathbf{o},\mathbf{I}_{r-1})$, where $\mathbf{I}_{r-1}$ is the $(r-1)\times(r-1)$–dimensional identity matrix.
  - The repeated application of Lemma 4.5 yields

$$\big(\mathbf{A}^{1/2}\mathbf{Z}'_n\big)^\top\big(\mathbf{A}^{1/2}\mathbf{Z}'_n\big)\xrightarrow{\mathrm{d}}\chi^2_{r-1}\,.$$

- Now it is sufficient to note that

$$\left(\mathbf{A}^{1/2}\mathbf{Z}_n'\right)^\top\left(\mathbf{A}^{1/2}\mathbf{Z}_n'\right) = T_n(X_1,\ldots,X_n)\,.$$

  - It namely holds that

$$\begin{aligned}\left(\mathbf{A}^{1/2}\mathbf{Z}_n'\right)^\top\left(\mathbf{A}^{1/2}\mathbf{Z}_n'\right) &= \left(\mathbf{Z}_n'\right)^\top\mathbf{A}\mathbf{Z}_n' \\ &= n\sum_{j=1}^{r-1}\frac{1}{p_{0j}}\Big(\frac{Z_{nj}}{n}-p_{0j}\Big)^2+\frac{n}{p_{0r}}\sum_{i=1}^{r-1}\sum_{j=1}^{r-1}\Big(\frac{Z_{ni}}{n}-p_{0i}\Big)\Big(\frac{Z_{nj}}{n}-p_{0j}\Big)\,,\end{aligned}$$

  - where the second summand of the last term can be written in the form

$$\frac{n}{p_{0r}}\sum_{i=1}^{r-1}\sum_{j=1}^{r-1}\Big(\frac{Z_{ni}}{n}-p_{0i}\Big)\Big(\frac{Z_{nj}}{n}-p_{0j}\Big)=\frac{n}{p_{0r}}\Big(\sum_{j=1}^{r-1}\Big(\frac{Z_{nj}}{n}-p_{0j}\Big)\Big)^2=\frac{n}{p_{0r}}\Big(\frac{Z_{nr}}{n}-p_{0r}\Big)^2\,,$$

  - because it obviously holds that $\sum_{j=1}^{r-1}Z_{nj}=n-Z_{nr}$ and $\sum_{j=1}^{r-1}p_{0j}=1-p_{0r}$. □

Remark

- In order to practically use the *χ^2-goodness–of–fit test* for the verification of the hypothesis $H_0:\mathbf{p}=\mathbf{p}_0$, first the value of the test statistic $T_n(x_1,\ldots,x_n)$, defined in (36), has to be computed.
  - For sufficiently large sample sizes n the hypothesis $H_0:\mathbf{p}=\mathbf{p}_0$ is rejected if

$$T_n(x_1,\ldots,x_n) > \chi^2_{r-1,1-\alpha}\,,$$

  - where $\chi^2_{r-1,1-\alpha}$ denotes the $(1-\alpha)$–quantile of the χ^2–distribution with $(r-1)$ degrees of freedom.
- A rule of thumb for n being sufficiently large, is the validity of the inequality $np_{0,j}\geq a$ for each $j\in\{1,\ldots,r\}$ and for a constant $a>0$.
  - In literature there are different opinions about the required size of $a>0$, which range from $a=2$ to $a=5$. Some authors even demand $a=10$.
  - Other authors think that for a large number of classes (about $r\geq 10$) the approximation is sufficiently good, even for $a=1$.

5.2.3 Goodness–of–Fit; Local Alternatives

It is not difficult to show the following (pointwise) consistency of the χ^2–goodness–of–fit test.

Theorem 5.6 *The χ^2–goodness–of–fit test is pointwise consistent for each vector $\mathbf{p}=(p_1,\ldots,p_{r-1})^\top\in[0,1]^{r-1}$ with $\mathbf{p}\neq\mathbf{p}_0$, i.e., it holds that*

$$\lim_{n\to\infty}\mathbb{P}_{\mathbf{p}}\big(T_n(X_1,\ldots,X_n)>\chi^2_{r-1,1-\alpha}\big)=1\,. \tag{43}$$

Proof

- From $\mathbf{p}\neq\mathbf{p}_0$ it follows that

$$p_j\neq p_{0,j} \tag{44}$$

  for some $j\in\{1,\ldots,r-1\}$.
- Furthermore, it follows from the strong law of large numbers (cf. Theorem WR–5.15) that $Z_{nj}/n\xrightarrow{\text{a.s.}}p_j$ for $n\to\infty$ under $\mathbb{P}_{\mathbf{p}}$.

- This and (44) imply that under $\mathbb{P}_{\mathbf{p}}$

$$T_n(X_1,\ldots,X_n) \geq n\Big(\frac{Z_{nj}}{n} - p_{0,j}\Big)^2 \xrightarrow{\text{a.s.}} \infty\,.$$

- Hence, the validity of (43) is shown. □

Remark

- Instead of considering a fixed vector $\mathbf{p} \neq \mathbf{p}_0$, there are also local alternatives $\mathbf{p}_n = (p_{n1},\ldots,p_{n,r-1})^\top$ possible of the form

$$p_{nj} = p_{0j} + \frac{h_j}{\sqrt{n}} \qquad \forall\, j = 1,\ldots,r-1\,, \tag{45}$$

which may depend on the sample size n, where

$$\sum_{j=1}^{r} h_j = 0\,. \tag{46}$$

- Then one can show that for $n \to \infty$ the asymptotic power of the χ^2–goodness–of–fit test vs. such alternatives may be smaller than 1.

In order to prove the statement, we need the following estimate, which is called the *Berry–Esseen theorem* in literature.

Lemma 5.5 *Let $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ be a sequence of independent and identically distributed random variables with $\mathbb{E}\left(|Y_1|^3\right) < \infty$. If $\mathbb{E}\,Y_1 = 0$ and $\operatorname{Var} Y_1 = 1$, then it holds for each $n \geq 1$*

$$\sup_{x\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{Y_1+\ldots+Y_n}{\sqrt{n}} \leq x\Big) - \Phi(x)\Big| \leq C\,\frac{\mathbb{E}\left(|Y_1|^3\right)}{\sqrt{n}}\,, \tag{47}$$

where $\Phi : \mathbb{R} \to [0,1]$ denotes the distribution function of the $N(0,1)$–distribution and $C < \infty$ is a universal constant, which does not depend on the distribution of the random variables $Y_1, Y_2, \ldots$.

Theorem 5.7 *Let $\{\mathbf{p}_n\}$ be a sequence of vectors, which is given by (45) and (46).*

- *Then it holds for each $x \geq 0$ that*

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big(T_n(X_1,\ldots,X_n) \leq x\big) = F_{r-1,\lambda}(x)\,, \tag{48}$$

where $F_{r-1,\lambda} : \mathbb{R} \to [0,1]$ is the distribution function of the noncentral χ^2–distribution with $r-1$ degrees of freedom, whose noncentrality parameter λ is given by

$$\lambda = \sum_{j=1}^{r} \frac{h_j^2}{p_{0j}}\,. \tag{49}$$

- *If $h_j \neq 0$ for some $j = 1,\ldots,r$, then the power of the χ^2–goodness–of–fit test converges, in the case of the local alternatives $\{\mathbf{p}_n\}$, to a limit, which is larger than α and smaller than 1, i.e.,*

$$\alpha < \lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big(T_n(X_1,\ldots,X_n) > \chi^2_{r-1,1-\alpha}\big) < 1\,. \tag{50}$$

Proof

- The proof of the first part is analogous to the proof of Theorem 5.5 since due to (45) and (46) it holds that

$$\mathbf{Z}'_n = \sqrt{n}\Big(\frac{Z_{n1}}{n} - p_{n1},\ \ldots,\ \frac{Z_{n,r-1}}{n} - p_{n,r-1}\Big)^\top + \mathbf{h}\,, \qquad \text{where} \quad \mathbf{h} = (h_1,\ldots,h_{r-1})^\top. \tag{51}$$

  for the random vector $\mathbf{Z}'_n$, introduced in (39).

- Because of (51) one can show in a similar way as in the proof of the multivariate central limit theorem in Lemma 5.2 that

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big(\mathbf{Z}'_n \le \mathbf{x}\big) = F_{\mathrm{N}(\mathbf{h},\mathbf{K})}(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbb{R}^{r-1}\,. \tag{52}$$

- In this context it is sufficient to notice that one can show, by using Berry–Esseen's theorem in Lemma 5.5 that (analogously to formula (10) in the proof of Lemma 5.2)
  - for arbitrary $\mathbf{t} = (t_1,\ldots,t_{r-1})^\top \in \mathbb{R}^{r-1}$ and $x \in \mathbb{R}$ it holds that

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\Big(n^{-1/2}\sum_{j=1}^{r-1} t_j\big(Z_{nj} - np_{nj}\big) \le x\Big) = F_{\mathrm{N}(0,\mathbf{t}^\top\mathbf{K}\mathbf{t})}(x)\,,$$

  - where $\mathbf{K}$ is the covariance matrix, introduced in (41).

- In the same way as in the proof of Theorem 5.5 we now get from (52) that

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big(\mathbf{A}^{1/2}\mathbf{Z}'_n \le \mathbf{x}\big) = F_{\mathrm{N}(\mathbf{A}^{1/2}\mathbf{h},\mathbf{I}_{r-1})}(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbb{R}^{r-1}\,.$$

  - From this and from the definition of the noncentral χ^2–distribution in Section 1.3.2 it follows that

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{p}_n}\big((\mathbf{A}^{1/2}\mathbf{Z}'_n)^\top(\mathbf{A}^{1/2}\mathbf{Z}'_n) \le x\big) = F_{r-1,\lambda}(x) \qquad \forall\, x \in \mathbb{R}\,,$$

  - where $\mathbf{A}$ is the inverse matrix $\mathbf{A} = \mathbf{K}^{-1}$ in (42) and the noncentrality parameter λ is given by

$$\lambda = (\mathbf{A}^{1/2}\mathbf{h})^\top(\mathbf{A}^{1/2}\mathbf{h}) = \mathbf{h}^\top\mathbf{A}\mathbf{h} = \sum_{j=1}^{r}\frac{h_j^2}{p_{0j}}\,.$$

- Thus, (48) is proved and because of $\alpha < 1 - F_{r-1,\lambda}(\chi^2_{r-1,1-\alpha}) < 1$ also (50) is valid. □

5.3 Pearson–Fisher Test

- The null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$, considered in Section 5.2, is in fact a compound hypothesis since it is equivalent to the hypothesis

$$H_0 :\ P \in \Delta_0\,,$$

  where Δ_0 is the subset of distributions of the sample variables, which has been introduced in (35).

- If it shall be verified whether the distribution P of the independent and identically distributed sample variables $X_1,\ldots,X_n$ belongs to a given (parametric) *class of distributions* $\{P_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\}$ with $\Theta \subset \mathbb{R}^m$, then we can proceed in a similar way as in the case of the χ^2–goodness–of–fit test, discussed in Section 5.2.

- The sample function $T_n : \mathbb{R}^n \to [0,\infty)$, which has been considered in the definition of the Pearson–statistic $T_n(X_1,\ldots,X_n)$ in (36), is replaced by a modified sample function $\widehat{T}_n : \mathbb{R}^n \to [0,\infty)$.

5.3.1 Pearson–Fisher Test Statistic

- In the same way as in Section 5.2.1, we "coarsen" the model, i.e.,
  - we partition the range of the sample variables $X_1, \ldots, X_n$ into r classes $(a_1, b_1], \ldots, (a_r, b_r]$ with $-\infty \leq a_1 < b_1 = a_2 < b_2 = \ldots = a_r < b_r \leq \infty$, where r is a (sufficiently large) natural number.
  - Instead of the sample variables $X_1, \ldots, X_n$, we once more consider the "class sizes" $Z_1, \ldots, Z_r$, which have been introduced in (32), where

$$Z_j = \#\{i : 1 \leq i \leq n,\, a_j < X_i \leq b_j\} \qquad \forall\, j = 1, \ldots, r\,.$$

- According to Lemma 5.4 it holds that $(Z_1, \ldots, Z_r) \sim \mathrm{M}_{r-1}(n, \mathbf{p})$, where we now assume
  - that the parameter $\mathbf{p} = (p_1, \ldots, p_{r-1})^\top \in [0,1]^{r-1}$ of the multinomial distribution $M_{r-1}(n, \mathbf{p})$
  - is a (known) function $\boldsymbol{\theta} \mapsto \mathbf{p}(\boldsymbol{\theta})$ of the (unknown) parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top \in \Theta \subset \mathbb{R}^m$ with $m < r - 1$.
- The hypothesis to be tested is $H_0 : \mathbf{p} \in \{\mathbf{p}(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\}$.
  - In order to be able to proceed with the verification of this hypothesis in a similar way as in Section 5.2, one first has to determine an estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)^\top$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$.
  - This also provides an estimator $(\widehat{p}_1, \ldots, \widehat{p}_r) = (p_1(\widehat{\boldsymbol{\theta}}), \ldots, p_r(\widehat{\boldsymbol{\theta}}))$ for the probabilities $(p_1, \ldots, p_r) = (p_1(\boldsymbol{\theta}), \ldots, p_r(\boldsymbol{\theta}))$, where

$$p_j(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(a_j < X_1 \leq b_j) \qquad \forall\, j = 1, \ldots, r\,.$$

Definition The random variable $\widehat{T}_n(X_1, \ldots, X_n)$, which is given by the sample function $\widehat{T}_n : \mathbb{R}^n \to [0, \infty)$ with

$$\widehat{T}_n(x_1, \ldots, x_n) = \sum_{j=1}^{r} \frac{\big(Z_j(x_1, \ldots, x_n) - n\widehat{p}_j(x_1, \ldots, x_n)\big)^2}{n\widehat{p}_j(x_1, \ldots, x_n)} \tag{53}$$

is called *Pearson–Fisher statistic.*

Remark

- If the mapping $\boldsymbol{\theta} \mapsto \mathbf{p}(\boldsymbol{\theta})$ is continuous and $\widehat{\boldsymbol{\theta}}$ is a (weakly) consistent estimator for $\boldsymbol{\theta}$,
  - then it follows from the law of large numbers (cf. Theorem WR–5.15) that for arbitrary $j \in \{1, \ldots, r\}$ and $\boldsymbol{\theta} \in \Theta$

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}} \left| \frac{1}{n}\, Z_j(X_1, \ldots, X_n) - \widehat{p}_j(X_1, \ldots, X_n) \right| = 0\,.$$

  - Therefore, it is reasonable to reject the null hypothesis $H_0 : \mathbf{p} \in \{\mathbf{p}(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\}$ if $\widehat{T}_n(x_1, \ldots, x_n)$ is significantly larger than 0.
- In order to be able to make this decision,
  - we first of all discuss conditions for the mapping $\boldsymbol{\theta} \mapsto \mathbf{p}(\boldsymbol{\theta})$ which enable the construction of a sequence of consistent (maximum–likelihood) estimators $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ that are asymptotically normally distributed.
  - Then we determine the (asymptotic limit) distribution of the test statistic $\widehat{T}_n(X_1, \ldots, X_n)$, introduced in (53), for $n \to \infty$.

5.3.2 Multivariate Central Limit Theorem for ML Estimators

In a similar way as in Section I–2.4.2, where the case $m = 1$ has been considered, it is possible to derive a *multivariate central limit theorem* for consistent sequences of maximum–likelihood estimators for the parameter vector $\boldsymbol{\theta}$.

In this context we need the following *regularity conditions*.

- The family $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ either consists only of discrete distributions or only of absolutely continuous distributions, where $\Theta \subset \mathbb{R}^m$ is an open set.
- It holds that
$$P_{\boldsymbol{\theta}} \neq P_{\boldsymbol{\theta}'} \qquad \text{if and only if} \qquad \boldsymbol{\theta} \neq \boldsymbol{\theta}' \,.$$
- The set $B = \{x \in \mathbb{R} : L(x;\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta} \in \Theta$, where the likelihood function $L(x;\boldsymbol{\theta})$ is given by
$$L(x;\boldsymbol{\theta}) = \begin{cases} p(x;\boldsymbol{\theta}) & \text{in the discrete case,} \\ f(x;\boldsymbol{\theta}) & \text{in the absolutely continuous case} \end{cases}$$
and $p(x;\boldsymbol{\theta})$ or $f(x;\boldsymbol{\theta})$ is the probability function or density of $P_{\boldsymbol{\theta}}$, respectively.
- Furthermore, let the mapping $\boldsymbol{\theta} \to L(x;\boldsymbol{\theta})$ for each $x \in B$ be three times continuously differentiable and suppose that for each $x \in B$ it holds that
$$\frac{\partial^k}{\partial\theta_{i_1} \ldots \partial\theta_{i_k}} \int\limits_B L(x;\boldsymbol{\theta})\, dx = \int\limits_B \frac{\partial^k}{\partial\theta_{i_1} \ldots \partial\theta_{i_k}} L(x;\boldsymbol{\theta})\, dx \qquad \forall\, k \in \{1,2,3\},\ i_1, \ldots, i_k \in \{1, \ldots, m\},\ \boldsymbol{\theta} \in \Theta, \tag{54}$$
where the integrals have to be replaced by the corresponding sums in the discrete case.
- For each $\boldsymbol{\theta}_0 \in \Theta$, a constant $c_{\boldsymbol{\theta}_0} > 0$ and a measurable function $g_{\boldsymbol{\theta}_0} : B \to [0, \infty)$ exist such that for each triple $(i_1, i_2, i_3) \in \{1, \ldots, m\}^3$
$$\Big| \frac{\partial^3}{\partial\theta_{i_1}\partial\theta_{i_2}\partial\theta_{i_3}} \log L(x;\boldsymbol{\theta}) \Big| \leq g_{\boldsymbol{\theta}_0}(x) \qquad \forall\, x \in B, \quad \forall\, \boldsymbol{\theta} \in \Theta \text{ with } |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < c_{\boldsymbol{\theta}_0} \tag{55}$$
and
$$\mathbb{E}_{\boldsymbol{\theta}_0} g_{\boldsymbol{\theta}_0}(X_1) < \infty \,. \tag{56}$$

Remark

- *Recall* :
  - In general, the maximum–likelihood estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$ for $\boldsymbol{\theta}$ is defined as the solution of the following optimization problem (cf. Section I–2.2.2).
  - In this context $\widehat{\boldsymbol{\theta}} : \mathbb{R}^n \to \Theta \subset \mathbb{R}^m$ is a sample function with
$$L(x_1, \ldots, x_n; \boldsymbol{\theta}) \leq L(x_1, \ldots, x_n; \widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)) \qquad \forall\, (x_1, \ldots, x_n) \in \mathbb{R}^n,\ \boldsymbol{\theta} \in \Theta \tag{57}$$
and
$$L(x_1, \ldots, x_n; \boldsymbol{\theta}) = \begin{cases} p(x_1;\boldsymbol{\theta}) \ldots p(x_n;\boldsymbol{\theta}) & \text{in the discrete case,} \\ f(x_1;\boldsymbol{\theta}) \ldots f(x_n;\boldsymbol{\theta}) & \text{in the absolutely continuous case.} \end{cases}$$
- Under the above mentioned regularity conditions one can show that for arbitrary $x_1, \ldots, x_n \in \mathbb{R}$ the estimate $\widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)$ fulfills the following system of equations:
$$\frac{\partial}{\partial\theta_i} L\big(x_1, \ldots, x_n; \widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)\big) = 0 \qquad \forall\, i = 1, \ldots, m \,. \tag{58}$$

- To formulate the multivariate central limit theorem, we need the notion of the *Fisher–information matrix*, which has already been introduced in Section 4.3.1.
  - For each $\boldsymbol{\theta} \in \Theta$ the $m \times m$ matrix $\mathbf{I}(\boldsymbol{\theta}) = (I_{ij}(\boldsymbol{\theta}))$ is considered with

$$I_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\Big(\frac{\partial}{\partial\theta_i} \log L(X_1;\boldsymbol{\theta}) \frac{\partial}{\partial\theta_j} \log L(X_1;\boldsymbol{\theta})\Big)\,, \tag{59}$$

  - where it is assumed that the expectation in (59) exists for arbitrary $i, j \in \{1, \ldots, m\}$ (and is a finite real number).

As a generalization of Theorem I–2.11, where the 1–dimensional case has been considered, it is possible to derive the following multivariate central limit theorem for weakly consistent sequences of maximum–likelihood estimators $\{\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n),\, n \geq 1\}$ for the parameter vector $\boldsymbol{\theta}$ which fulfill the system of equations (58).

Theorem 5.8

- *Let the Fisher–information matrix* $\mathbf{I}(\boldsymbol{\theta})$ *be positive definite (and therefore invertible) for each* $\boldsymbol{\theta} \in \Theta$ *and let* $\{\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n),\, n \geq 1\}$ *be a weakly consistent sequence of maximum–likelihood estimators for* $\boldsymbol{\theta}$.
- *Then it holds for* $n \to \infty$ *that*

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n) - \boldsymbol{\theta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(\mathbf{o}, \mathbf{I}^{-1}(\boldsymbol{\theta})\big)\,. \tag{60}$$

The *proof* of Theorem 5.8 proceeds in a similar way as the proof of Theorem I–2.11. It is therefore omitted, cf. for instance E.L. Lehmann und G. Casella (1998) *The Theory of Point Estimation*, Springer–Verlag, New York.

5.3.3 Fisher–Information Matrix and Central Limit Theorem in the Coarsened Model

- We now return to the "coarsened" model, already considered in Section 5.3.1.
  - Here we assume that $L : \mathbb{R} \times \Theta \to (0, 1)$ is the likelihood function with

$$L(x;\boldsymbol{\theta}) = p_j(\boldsymbol{\theta})\,, \qquad \text{if } x \in (a_j, b_j], \tag{61}$$

  - where the probabilities $p_j(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(a_j < X_1 \leq b_j)$ are positive and smaller than 1.
- Furthermore, we assume that the regularity conditions, formulated in Section 5.3.2, are fulfilled for the likelihood function given in (61).

Lemma 5.6 *Then it holds for the Fisher–information matrix* $\mathbf{I}(\boldsymbol{\theta})$ *that*

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})^\top \mathbf{C}(\boldsymbol{\theta})\,, \tag{62}$$

where

$$\mathbf{C}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial p_1(\boldsymbol{\theta})/\partial\theta_1}{\sqrt{p_1(\boldsymbol{\theta})}} & \frac{\partial p_1(\boldsymbol{\theta})/\partial\theta_2}{\sqrt{p_1(\boldsymbol{\theta})}} & \cdots & \frac{\partial p_1(\boldsymbol{\theta})/\partial\theta_m}{\sqrt{p_1(\boldsymbol{\theta})}} \\ \frac{\partial p_2(\boldsymbol{\theta})/\partial\theta_1}{\sqrt{p_2(\boldsymbol{\theta})}} & \frac{\partial p_2(\boldsymbol{\theta})/\partial\theta_2}{\sqrt{p_2(\boldsymbol{\theta})}} & \cdots & \frac{\partial p_2(\boldsymbol{\theta})/\partial\theta_m}{\sqrt{p_2(\boldsymbol{\theta})}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial p_r(\boldsymbol{\theta})/\partial\theta_1}{\sqrt{p_r(\boldsymbol{\theta})}} & \frac{\partial p_r(\boldsymbol{\theta})/\partial\theta_2}{\sqrt{p_r(\boldsymbol{\theta})}} & \cdots & \frac{\partial p_r(\boldsymbol{\theta})/\partial\theta_m}{\sqrt{p_r(\boldsymbol{\theta})}} \end{pmatrix}. \tag{63}$$

Proof

- Because of (61) it holds for each $x \in \mathbb{R}$ that

$$\log L(x;\boldsymbol{\theta}) = \sum_{j=1}^{r} \mathbb{1}_{\{a_j < x \le b_j\}} \log p_j(\boldsymbol{\theta})\,.$$

- From this it follows for the entries $I_{ij}(\boldsymbol{\theta})$ of $\mathbf{I}(\boldsymbol{\theta})$ that

$$\begin{aligned} I_{ij}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\Big(\frac{\partial}{\partial\theta_i}\log L(X_1;\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log L(X_1;\boldsymbol{\theta})\Big) = \sum_{k=1}^{r}\Big(\frac{\partial}{\partial\theta_i}\log p_k(\boldsymbol{\theta})\Big)\Big(\frac{\partial}{\partial\theta_j}\log p_k(\boldsymbol{\theta})\Big)p_k(\boldsymbol{\theta}) \\ &= \sum_{k=1}^{r}\Big(\frac{\partial}{\partial\theta_i}p_k(\boldsymbol{\theta})\Big)\big(p_k(\boldsymbol{\theta})\big)^{-1}\Big(\frac{\partial}{\partial\theta_j}p_k(\boldsymbol{\theta})\Big) = \Big(\mathbf{C}(\boldsymbol{\theta})^{\top}\mathbf{C}(\boldsymbol{\theta})\Big)_{ij}\,. \end{aligned}$$

□

Thus, Theorem 5.8 implies the following result.

Corollary 5.1 *If the matrix* $\mathbf{I}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})^{\top}\mathbf{C}(\boldsymbol{\theta})$*, given in* (63)*, is positive definite for each* $\boldsymbol{\theta} \in \Theta$*, then it holds that*

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}(X_1,\ldots,X_n) - \boldsymbol{\theta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(\mathbf{o}, \big(\mathbf{C}(\boldsymbol{\theta})^{\top}\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\big) \tag{64}$$

for each weakly consistent sequence $\{\widehat{\boldsymbol{\theta}}(X_1,\ldots,X_n),\, n \ge 1\}$ *of maximum–likelihood estimators for* $\boldsymbol{\theta}$ *which are obtained from observations of the "coarsened" model.*

Remark

- From (61) it follows for the likelihood function $L(x_1,\ldots,x_n;\boldsymbol{\theta})$ that

$$L(x_1,\ldots,x_n;\boldsymbol{\theta}) = \prod_{j=1}^{r} p_j(\boldsymbol{\theta})^{Z_j(x_1,\ldots,x_n)}\,,$$

  or for the loglikelihood function $\log L(x_1,\ldots,x_n;\boldsymbol{\theta})$ that

$$\log L(x_1,\ldots,x_n;\boldsymbol{\theta}) = \sum_{j=1}^{r} Z_j(x_1,\ldots,x_n)\log p_j(\boldsymbol{\theta})\,. \tag{65}$$

- Each maximum–likelihood estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}\big(Z_1(x_1,\ldots,x_n),\ldots,Z_r(x_1,\ldots,x_n)\big)$ for $\boldsymbol{\theta}$ which is obtained from the coarsened data $Z_1(x_1,\ldots,x_n),\ldots,Z_r(x_1,\ldots,x_n)$ satisfies the system of equations

$$\frac{\partial \log L(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_i} = 0\,, \qquad \forall\, i = 1,\ldots,m\,. \tag{66}$$

  because of the above-mentioned regularity conditions.
- Here it follows from (65) that for arbitrary $i = 1,\ldots,m$ and $\boldsymbol{\theta} \in \Theta$

$$\frac{\partial \log L(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_i} = \sum_{j=1}^{r} \frac{Z_j(x_1,\ldots,x_n)}{p_j(\boldsymbol{\theta})}\,\frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_i}$$

  or

$$\frac{\partial \log L(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_i} = \sum_{j=1}^{r} \frac{Z_j(x_1,\ldots,x_n) - np_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})}\,\frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_i}\,, \tag{67}$$

  where the last equality is due to the fact that

$$\sum_{j=1}^{r} \frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_i} = 0\,, \qquad \forall\, i = 1,\ldots,m\,. \tag{68}$$

5.3.4 Asymptotic Distribution of the Pearson–Fisher Statistic

The following theorem is the basis for the χ^2- goodness–of–fit test of Pearson-Fisher. Here we always assume that

- the likelihood function of the coarsened model, considered in (61), fulfills the regularity conditions of Section 5.3.2,
- the Fisher–information matrix $I(\boldsymbol{\theta})$, given in (62), is positive definite and $\{\widehat{\boldsymbol{\theta}}_n\} = \{\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n),\, n \geq 1\}$ is a weakly consistent sequence of ML estimators for $\boldsymbol{\theta}$ which is obtained by considering the coarsened model.

Theorem 5.9

- *Let $\widehat{T}_n(X_1, \ldots, X_n)$ be the Pearson–Fisher test statistic given in* (53)*, i.e.,*

$$\widehat{T}_n(X_1, \ldots, X_n) = \sum_{j=1}^{r} \frac{\big(Z_j(X_1, \ldots, X_n) - n\widehat{p}_j(X_1, \ldots, X_n)\big)^2}{n\widehat{p}_j(X_1, \ldots, X_n)}\,, \tag{69}$$

  where $\widehat{p}_j(X_1, \ldots, X_n) = p_j\big(\widehat{\boldsymbol{\theta}}(X_1, \ldots, X_n)\big)$.

- *Then it holds that*

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}}\big(\widehat{T}_n(X_1, \ldots, X_n) > \chi^2_{r-1-m,1-\alpha}\big) = \alpha\,, \qquad \forall\, \alpha \in (0,1) \tag{70}$$

  for each $\boldsymbol{\theta} \in \Theta$, where $\chi^2_{r-1-m,1-\alpha}$ denotes the $(1-\alpha)$–quantile of the χ^2–distribution with $r-1-m$ degrees of freedom.

It is possible to give a full proof of Theorem 5.9 by reinterpreting the χ^2–goodness–of–fit test of Pearson–Fisher as a *likelihood–ratio test*, cf. for instance Section 4.7 in H. Pruscha (2000) *Vorlesungen über mathematische Statistik*, Teubner–Verlag, Stuttgart.

However, as this method of proof is rather complex, we only show a derivation of Theorem 5.9, which is partly *heuristic*.

- Let $\mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), \ldots, p_r(\boldsymbol{\theta}))^\top$ and $\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) = \big(\widetilde{Z}_{n1}(\boldsymbol{\theta}), \ldots, \widetilde{Z}_{nr}(\boldsymbol{\theta})\big)^\top$ with

$$\widetilde{Z}_{nj}(\boldsymbol{\theta}) = \frac{Z_j(X_1, \ldots, X_n) - np_j(\boldsymbol{\theta})}{\sqrt{np_j(\boldsymbol{\theta})}}\,, \qquad j = 1, \ldots, r\,. \tag{71}$$

- Since $\mathbb{E}\,\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) = \mathbf{o}$ and since it is possible to write $\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta})$ as a sum of n independent and identically distributed random vectors, it follows from the multivariate central limit theorem (in the same way as in the proof of Theorem 5.5) that

$$\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \widetilde{\mathbf{Z}}(\boldsymbol{\theta}) \sim \mathrm{N}(\mathbf{o}, \mathbf{B}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}))\,, \tag{72}$$

  where

$$\mathbf{B}(\boldsymbol{\theta}) = \begin{pmatrix} 1/\sqrt{p_1(\boldsymbol{\theta})} & 0 & \ldots & 0 \\ 0 & 1/\sqrt{p_2(\boldsymbol{\theta})} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1/\sqrt{p_r(\boldsymbol{\theta})} \end{pmatrix}, \qquad \mathbf{K}(\boldsymbol{\theta}) = \begin{cases} -p_i(\boldsymbol{\theta})p_j(\boldsymbol{\theta})\,, & \text{if } i \neq j, \\ p_i(\boldsymbol{\theta})(1 - p_j(\boldsymbol{\theta}))\,, & \text{if } i = j. \end{cases}$$

– Thus, for the covariance matrix $\mathbf{B}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta})$ in (72) we get that

$$\mathbf{B}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta})\mathbf{B}(\boldsymbol{\theta}) = \mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\,, \qquad \text{where } \mathbf{q}(\boldsymbol{\theta}) = \big(\sqrt{p_1(\boldsymbol{\theta})},\ldots,\sqrt{p_r(\boldsymbol{\theta})}\big)^\top. \tag{73}$$

– For the matrix $\mathbf{C}(\boldsymbol{\theta})$, introduced in (63), it holds because of (68) that $\mathbf{q}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta}) = \mathbf{o}$ and thus

$$\begin{aligned}
&\Big(\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big)\Big(\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)^\top \\
&= \Big(\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big)\mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1} \\
&= \big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}.
\end{aligned}$$

– From (72) and (73) it now follows that

$$\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\widetilde{\mathbf{Z}}(\boldsymbol{\theta}) \sim \mathrm{N}\big(\mathbf{o}\,,\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\big)\,. \tag{74}$$

- Furthermore, Corollary 5.1 and the Taylor series expansion yield

$$\sqrt{n}\,\mathbf{B}(\boldsymbol{\theta})\big(\mathbf{p}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{p}(\boldsymbol{\theta})\big) = \sqrt{n}\,\mathbf{C}(\boldsymbol{\theta})\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) + o\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) \xrightarrow{\mathrm{d}} \mathrm{N}\big(\mathbf{o}, \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\big)\,.$$

– This and (74) imply that

$$\sqrt{n}\,\mathbf{B}(\boldsymbol{\theta})\big(\mathbf{p}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{p}(\boldsymbol{\theta})\big) \xrightarrow{\mathrm{d}} \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\widetilde{\mathbf{Z}}(\boldsymbol{\theta})\,. \tag{75}$$

– Moreover, it follows from (69) and (71) that

$$\begin{aligned}
\widehat{T}_n(X_1,\ldots,X_n) &= \sum_{j=1}^{r}\big(\widetilde{Z}_{nj}(\widehat{\boldsymbol{\theta}}_n)\big)^2 \\
&= \sum_{j=1}^{r}\left(\widetilde{Z}_{nj}(\boldsymbol{\theta}) + \widetilde{Z}_{nj}(\boldsymbol{\theta})\Big(\sqrt{\frac{p_j(\boldsymbol{\theta})}{p_j(\widehat{\boldsymbol{\theta}}_n)}} - 1\Big) - \frac{\sqrt{n}}{\sqrt{p_j(\widehat{\boldsymbol{\theta}}_n)}}\big(p_j(\widehat{\boldsymbol{\theta}}_n) - p_j(\boldsymbol{\theta})\big)\right)^2 \\
&= \sum_{j=1}^{r}\left(\widetilde{Z}_{nj}(\boldsymbol{\theta}) - \frac{\sqrt{n}}{\sqrt{p_j(\widehat{\boldsymbol{\theta}}_n)}}\big(p_j(\widehat{\boldsymbol{\theta}}_n) - p_j(\boldsymbol{\theta})\big) + o(1)\right)^2,
\end{aligned}$$

– where the last equality follows from the convergence

$$\widetilde{Z}_{nj}(\boldsymbol{\theta})\Big(\sqrt{\frac{p_j(\boldsymbol{\theta})}{p_j(\widehat{\boldsymbol{\theta}}_n)}} - 1\Big) \xrightarrow{\mathrm{P}} 0\,, \qquad \forall\, j = 1,\ldots,r,$$

which is due to $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\theta}$ and the continuous mapping theorem for random vectors (cf. Lemma 4.5).

- In other words: With the notation $\varphi(z_1,\ldots,z_r) = \sum_{j=1}^{r} z_j^2$ it holds that

$$\widehat{T}_n(X_1,\ldots,X_n) = \varphi\Big(\widetilde{\mathbf{Z}}_n(\boldsymbol{\theta}) - \sqrt{n}\,\mathbf{B}(\boldsymbol{\theta})\big(\mathbf{p}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{p}(\boldsymbol{\theta})\big) + o(1)\Big)\,. \tag{76}$$

– Together with (72) and (75), the asymptotic approximation formula (76) suggests the *conjecture* that for $n \to \infty$

$$\widehat{T}_n(X_1,\ldots,X_n) \xrightarrow{\mathrm{d}} \varphi\Big(\Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)\widetilde{\mathbf{Z}}(\boldsymbol{\theta})\Big)\,. \tag{77}$$

– However, the convergence in (77) *does not directly* follow from (72), (75) and (76), but needs a *separate proof*, which is omitted here.

- We now show that
$$\varphi\Big(\Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)\widetilde{\mathbf{Z}}(\boldsymbol{\theta})\Big) \sim \chi^2_{r-1-m}\,. \tag{78}$$

- In (72) and (73) we have shown that
$$\widetilde{\mathbf{Z}}(\boldsymbol{\theta}) \sim \mathrm{N}\big(\mathbf{o}, \mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\big)\,, \qquad \text{where } \mathbf{q}(\boldsymbol{\theta}) = \big(\sqrt{p_1(\boldsymbol{\theta})}, \ldots, \sqrt{p_r(\boldsymbol{\theta})}\big)^\top.$$
  - Furthermore, it follows from $\mathbf{q}^\top(\boldsymbol{\theta})\mathbf{q}(\boldsymbol{\theta}) = 1$ that
$$\begin{aligned}\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big)^2 &= \mathbf{I}_r - 2\mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta}) + \mathbf{q}(\boldsymbol{\theta})\underbrace{\mathbf{q}^\top(\boldsymbol{\theta})\mathbf{q}(\boldsymbol{\theta})}_{=1}\mathbf{q}^\top(\boldsymbol{\theta}) \\ &= \mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\,,\end{aligned}$$
  - i.e., the covariance matrix
$$\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})) = \begin{pmatrix} 1-p_1(\boldsymbol{\theta}) & -\sqrt{p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})} & -\sqrt{p_1(\boldsymbol{\theta})p_3(\boldsymbol{\theta})} & \ldots & -\sqrt{p_1(\boldsymbol{\theta})p_r(\boldsymbol{\theta})} \\ -\sqrt{p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})} & 1-p_2(\boldsymbol{\theta}) & -\sqrt{p_2(\boldsymbol{\theta})p_3(\boldsymbol{\theta})} & \ldots & -\sqrt{p_2(\boldsymbol{\theta})p_r(\boldsymbol{\theta})} \\ \vdots & \vdots & \vdots & & \vdots \\ -\sqrt{p_1(\boldsymbol{\theta})p_r(\boldsymbol{\theta})} & -\sqrt{p_2(\boldsymbol{\theta})p_r(\boldsymbol{\theta})} & -\sqrt{p_3(\boldsymbol{\theta})p_r(\boldsymbol{\theta})} & \ldots & 1-p_r(\boldsymbol{\theta}) \end{pmatrix}$$
    of the random vector $\widetilde{\mathbf{Z}}(\boldsymbol{\theta})$ is symmetric and idempotent.
  - This and the last part of Theorem 1.4 imply the representation formula
$$\widetilde{\mathbf{Z}}(\boldsymbol{\theta}) \stackrel{\mathrm{d}}{=} \Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big)\boldsymbol{\varepsilon}\,, \tag{79}$$
    where $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{o}, \mathbf{I}_r)$.

- Moreover, also the matrix $\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})$ is symmetric and idempotent and from $\mathbf{q}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta}) = \mathbf{o}$ it follows that
$$\Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big) = \mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta}) - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\,.$$
  - This implies that the matrix $\mathbf{R} = \Big(\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big)\Big(\mathbf{I}_r - \mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big)$ is symmetric and idempotent as well.
  - From (79) and Theorem 1.9 we now get that
$$\varphi\Big(\big(\mathbf{I}_r - \mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\big)\widetilde{\mathbf{Z}}(\boldsymbol{\theta})\Big) \stackrel{\mathrm{d}}{=} \varphi(\mathbf{R}\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^\top\mathbf{R}\boldsymbol{\varepsilon} \sim \chi^2_{\mathrm{rk}(\mathbf{R})}\,.$$
  - Due to Lemma 1.3 it holds for the rank $\mathrm{rk}(\mathbf{R})$ of the symmetric and idempotent matrix $\mathbf{R}$ that
$$\begin{aligned}\mathrm{rk}(\mathbf{R}) &= \mathrm{tr}(\mathbf{R}) \\ &= \mathrm{tr}(\mathbf{I}_r) - \mathrm{tr}\Big(\mathbf{C}(\boldsymbol{\theta})\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\Big) - \mathrm{tr}\Big(\mathbf{q}(\boldsymbol{\theta})\mathbf{q}^\top(\boldsymbol{\theta})\Big) \\ &= \mathrm{tr}(\mathbf{I}_r) - \mathrm{tr}\Big(\big(\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\big)^{-1}\mathbf{C}^\top(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\Big) - \mathrm{tr}\Big(\mathbf{q}^\top(\boldsymbol{\theta})\mathbf{q}(\boldsymbol{\theta})\Big) \\ &= r - m - 1\,.\end{aligned}$$

- This proves the validity of (78).

Remark For the practical usage of the χ^2- goodness–of–fit test of Pearson–Fisher one can proceed in a similar way as described in Section 5.2.2 in order to verify the hypothesis $H_0 : P \in \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$.

- First of all, a ML estimation $\widehat{\boldsymbol{\theta}}(x_1,\ldots,x_n) = (\widehat{\theta}_1(x_1,\ldots,x_n),\ldots,\widehat{\theta}_m(x_1,\ldots,x_n))^\top$ for $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)^\top$ is determined by solving the system of equations (66).
- Then the value of the test statistic $T_n(x_1,\ldots,x_n)$, defined in (53), is calculated.
- For sufficiently large sample sizes n the hypothesis $H_0 : P \in \{P_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\}$ is rejected if

$$T_n(x_1,\ldots,x_n) > \chi^2_{r-1-m,1-\alpha}\,,$$

- where $\chi^2_{r-1-m,1-\alpha}$ denotes the $(1-\alpha)$–quantile of the χ^2–distribution with $(r-1-m)$ degrees of freedom.

5.4 Examples

5.4.1 Pearson–Fisher Test for Poisson–Distribution

- By observing the (independent and identically distributed) sample variables $X_1,\ldots,X_n$ it shall be tested if the distribution P of X_i belongs to the family of Poisson–distributions.
  - So let $\Theta = (0,\infty)$ with $\boldsymbol{\theta} = \lambda$ and let $\{P_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\} = \{\,\mathrm{Poi}(\lambda),\, \lambda > 0\}$ be the family of Poisson–distributions.
  - We consider the following r classes $\{0\},\{1\},\ldots,\{r-2\}$ and $\{r-1,r,r+1,\ldots\}$, i.e.

$$(a_1,b_1] = (-\infty,0], \quad (a_2,b_2] = (0,1], \quad \ldots \quad (a_{r-1},b_{r-1}] = (r-3,r-2], \quad (a_r,b_r] = (r-2,\infty]\,.$$

  - The probabilities $p_j(\lambda) = \mathbb{P}_\lambda(a_j < X_1 \le b_j)$ then are given by

$$p_j(\lambda) = \frac{\lambda^{j-1}}{(j-1)!}\, e^{-\lambda} \quad \forall\, j = 1,\ldots,r-1 \qquad \text{and} \qquad p_r(\lambda) = \sum_{i=r}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}\, e^{-\lambda}. \tag{80}$$

- According to (66), every maximum–likelihood estimate $\widehat{\lambda}$ for λ which is obtained from grouped data satisfies the equation

$$\sum_{j=1}^{r} Z_j(x_1,\ldots,x_n)\, \frac{\frac{d}{d\lambda}\, p_j(\lambda)}{p_j(\lambda)} = 0\,. \tag{81}$$

  - Here it follows from (80) that

$$\frac{\frac{d}{d\lambda}\, p_j(\lambda)}{p_j(\lambda)} = \frac{j-1}{\lambda} - 1 \quad \forall\, j = 1,\ldots,r-1 \qquad \text{and} \qquad \frac{\frac{d}{d\lambda}\, p_r(\lambda)}{p_r(\lambda)} = \frac{\displaystyle\sum_{i=r}^{\infty} \frac{\left(\frac{i-1}{\lambda} - 1\right)\lambda^{i-1}}{(i-1)!}}{\displaystyle\sum_{i=r}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}}\,.$$

  - This and (81) imply that the ML estimate $\widehat{\lambda}$ fulfills the following equation:

$$\sum_{j=1}^{r-1} Z_j(x_1,\ldots,x_n)\, \left(\frac{j-1}{\lambda} - 1\right) + Z_r(x_1,\ldots,x_n)\, \frac{\displaystyle\sum_{i=r}^{\infty} \frac{\left(\frac{i-1}{\lambda} - 1\right)\lambda^{i-1}}{(i-1)!}}{\displaystyle\sum_{i=r}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}} = 0\,. \tag{82}$$

- For each n there is one $r_0 = r_0(n) \in \mathbb{N}$, such that $Z_r(x_1, \ldots, x_n) = 0$ for each $r > r_0$. This and (82) imply that for $r \to \infty$

$$\widehat{\lambda}_n = \widehat{\lambda}(x_1, \ldots, x_n) \to \overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \,.$$

  - For a sufficiently large number r of classes $\{0\}$, $\{1\}$, …, $\{r-2\}$, $\{r-1, r, r+1, \ldots\}$ the sample mean $\overline{x}_n$, which is a ML estimate for λ in the non–aggregated Poisson–model, is a good approximation for the ML estimation $\widehat{\lambda}_n$ for λ in aggregated Poisson–models.
  - The null hypothesis $H_0 : P \in \{\,\mathrm{Poi}(\lambda),\, \lambda > 0\}$ is therefore rejected if

$$\widehat{T}_n(x_1, \ldots, x_n) = \sum_{j=1}^{r} \frac{\big(Z_j(x_1, \ldots, x_n) - n\widehat{p}_j(x_1, \ldots, x_n)\big)^2}{n\widehat{p}_j(x_1, \ldots, x_n)} > \chi^2_{r-2,1-\alpha} \,,$$

    where $\widehat{p}_j(x_1, \ldots, x_n) = p_j(\overline{x}_n)$ with the function $p_j : (0, \infty) \to [0, 1]$ given in (80) and the estimation $\overline{x}_n$ for λ.

5.4.2 Pearson–Fisher Test for Normal Distribution

- Now let $\Theta = \mathbb{R} \times (0, \infty)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ and let $\{P_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\} = \{\,\mathrm{N}(\mu, \sigma^2),\, \mu \in \mathbb{R}, \sigma^2 > 0\}$ be the family of (one–dimensional) normal distributions.
  - The probabilities $p_j(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(a_j < X_1 \le b_j)$ are then given by

$$p_j(\boldsymbol{\theta}) = \int_{a_j}^{b_j} f(x; \boldsymbol{\theta})\, dx \,, \qquad \text{where } f(x; \boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big). \tag{83}$$

  - According to (66), each maximum–likelihood estimate

$$\widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n) = (\widehat{\mu}(x_1, \ldots, x_n), \widehat{\sigma}^2(x_1, \ldots, x_n))^\top$$

    for $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ which is obtained from the aggregated data satisfies the system of equations

$$\sum_{j=1}^{r} Z_j(x_1, \ldots, x_n) \frac{\int_{a_j}^{b_j} \frac{\partial}{\partial \theta_i} f(x; \boldsymbol{\theta})\, dx}{\int_{a_j}^{b_j} f(x; \boldsymbol{\theta})\, dx} = 0 \qquad \text{for } i = 1, 2. \tag{84}$$

  - Here it follows from (83) that

$$\frac{\partial}{\partial \mu} f(x; \boldsymbol{\theta}) = \frac{x-\mu}{\sigma^2}\, f(x; \boldsymbol{\theta}) \qquad \text{or} \qquad \frac{\partial}{\partial \sigma^2} f(x; \boldsymbol{\theta}) = f(x; \boldsymbol{\theta}) \Big(\frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\Big) .$$

- This and (84) imply that the ML estimate $\widehat{\boldsymbol{\theta}}(x_1, \ldots, x_n)$ satisfies the following system of equations:

$$\sum_{j=1}^{r} Z_j(x_1, \ldots, x_n) \frac{\int_{a_j}^{b_j} (x-\mu) f(x; \boldsymbol{\theta})\, dx}{\int_{a_j}^{b_j} f(x; \boldsymbol{\theta})\, dx} = 0 \,, \qquad \sum_{j=1}^{r} Z_j(x_1, \ldots, x_n) \frac{\int_{a_j}^{b_j} (x-\mu)^2 f(x; \boldsymbol{\theta})\, dx}{\int_{a_j}^{b_j} f(x; \boldsymbol{\theta})\, dx} - n\sigma^2 = 0 \,,$$

  where the first equality of this system of equations is equivalent to

$$\sum_{j=1}^{r} Z_j(x_1, \ldots, x_n) \frac{\int_{a_j}^{b_j} x f(x; \boldsymbol{\theta})\, dx}{\int_{a_j}^{b_j} f(x; \boldsymbol{\theta})\, dx} - \mu \underbrace{\sum_{j=1}^{r} Z_j(x_1, \ldots, x_n)}_{=n} = 0 \,.$$

- Thus, the ML estimate $\widehat{\boldsymbol{\theta}}(x_1,\dots,x_n) = (\widehat{\mu}(x_1,\dots,x_n), \widehat{\sigma}^2(x_1,\dots,x_n))^\top$ fulfills the system of equations

$$\mu = \frac{1}{n}\sum_{j=1}^{r} Z_j(x_1,\dots,x_n)\frac{\int_{a_j}^{b_j} x f(x;\mu,\sigma^2)\,dx}{\int_{a_j}^{b_j} f(x;\mu,\sigma^2)\,dx}\,, \qquad \sigma^2 = \frac{1}{n}\sum_{j=1}^{r} Z_j(x_1,\dots,x_n)\frac{\int_{a_j}^{b_j} (x-\mu)^2 f(x;\mu,\sigma^2)\,dx}{\int_{a_j}^{b_j} f(x;\mu,\sigma^2)\,dx}\,,$$

  - which for a sufficiently large number r of classes $(a_1,b_1],\dots,(a_r,b_r]$ can be solved approximately in the following way:

$$\widehat{\mu} \approx \frac{1}{n}\sum_{j=1}^{r} c_j Z_j(x_1,\dots,x_n)\,, \qquad \widehat{\sigma}^2 \approx \frac{1}{n}\sum_{j=1}^{r} (c_j-\widehat{\mu})^2 Z_j(x_1,\dots,x_n)\,, \tag{85}$$

  - where $c_1 = b_1$ is the right endpoint of the first class, $c_r = b_{r-1}$ is the left endpoint of the r–th class and $c_j = (b_{j-1}+b_j)/2$ is the center of the j–th class for $j = 2,\dots,r-1$.

- The null hypothesis $H_0 : P \in \{\,\mathrm{N}(\mu,\sigma^2),\ \mu\in\mathbb{R}, \sigma^2>0\}$ is rejected if

$$\widehat{T}_n(x_1,\dots,x_n) = \sum_{j=1}^{r} \frac{\big(Z_j(x_1,\dots,x_n) - n\widehat{p}_j(x_1,\dots,x_n)\big)^2}{n\widehat{p}_j(x_1,\dots,x_n)} > \chi^2_{r-3,1-\alpha}\,,$$

  where $\widehat{p}_j(x_1,\dots,x_n) = p_j(\widehat{\mu},\widehat{\sigma}^2)$ with the function $p_j : \mathbb{R}\times(0,\infty)\to[0,1]$ given in (83) and $(\widehat{\mu},\widehat{\sigma}^2)$ is the estimation for (μ,σ^2), given in (85).

Remark

- The approximate solution (85) of the system of equations (84) shall now be used if the number r of classes is large enough.
  - This requires a sufficiently large sample size n.
  - In other words: If the sample size n is small, then the χ^2-goodness–of–fit test is not suited to verify the hypothesis for normality.
- Alternative tests for normal distribution are the following *goodness–of–fit tests of Shapiro–Wilk–type*, which lead to acceptable results even for a small sample size n.

5.4.3 Goodness–of–Fit Tests of Shapiro–Wilk–Type

- In this section we discuss two goodness–of–fit tests of Shapiro–Wilk–type which can also be used to verify the hypothesis $H_0 : P \in \{\,\mathrm{N}(\mu,\sigma^2),\ \mu\in\mathbb{R}, \sigma^2>0\}$.
- Here the *order statistics* $X_{(1)},\dots,X_{(n)}$ of the (independent and identically distributed) sample variables $X_1,\dots,X_n$ are considered, which have already been introduced in Section I–1.4.
  - *Recall*: The order statistics are defined by using the sample function $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, where

$$(x_1,\dots,x_n) \to (x_{(1)},\dots,x_{(n)}) = \varphi(x_1,\dots,x_n) \qquad \text{with } x_{(i)} = \min\big\{x_j : \#\{k : x_k \le x_j\} \ge i\big\} \tag{86}$$

    for each $i\in\{1,\dots,n\}$.
  - The mapping $\varphi : \mathbb{R}^n\to\mathbb{R}^n$, given in (86), is a *permutation* of the components of the vector $(x_1,\dots,x_n)$, such that

$$x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}\,.$$

– For each $\omega \in \Omega$ let now

$$(X_{(1)}(\omega), \ldots, X_{(n)}(\omega)) = \varphi\big(X_1(\omega), \ldots, X_n(\omega)\big)$$

be the (measurable) permutation, given in (86), of $(X_1(\omega), \ldots, X_n(\omega))$, such that

$$X_{(1)}(\omega) \leq \ldots \leq X_{(n)}(\omega)\,. \tag{87}$$

– The random variables $X_{(1)}, \ldots, X_{(n)} : \Omega \to \mathbb{R}$ are called the *order statistics* of $(X_1, \ldots, X_n)$.

- If $X_i \sim \mathrm{N}(\mu, \sigma^2)$ for certain $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, then it is easy to see that the following representation formula holds for the expectation $b_i = \mathbb{E}\, X_{(i)}$ of the order statistics $X_{(i)}$:

$$b_i = \mu + \sigma a_i \qquad \forall\, i = 1, \ldots, n\,, \tag{88}$$

– where $a_i = \mathbb{E}\, Y_{(i)}$ is the expectation of the i–th order statistic $Y_{(i)}$ for $\mathrm{N}(0,1)$–distributed sample variables $Y_1, \ldots, Y_n$.

– The benefit of the representation formula (88) is that the expectations $a_1, \ldots, a_n$ are available in the form of tables or can be determined using Monte Carlo simulation.

- Since the vectors $(b_1, \ldots, b_n)$ and $(X_{(1)}, \ldots, X_{(n)})$ should differ only little under H_0, the following *empirical correlation coefficient* is considered to verify the null hypothesis
$H_0 : P \in \{\, \mathrm{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$:

$$\widetilde{T}(X_1, \ldots, X_n) = \frac{\sum\limits_{i=1}^{n} \big(b_i - \overline{b}\big)\big(X_{(i)} - \overline{X}\big)}{\sqrt{\sum\limits_{i=1}^{n} (b_i - \overline{b})^2}\sqrt{\sum\limits_{i=1}^{n} \big(X_{(i)} - \overline{X}\big)^2}}\,, \tag{89}$$

where $\overline{b} = \sum_{i=1}^{n} b_i/n$ and $\overline{X} = \sum_{i=1}^{n} X_i/n$.

1. *Shapiro–Francia test*

- Since correlation coefficients are invariant under linear transformations, we are able to replace b_i in (89) by a_i for each $i \in \{1, \ldots, n\}$, where $\overline{a} = \sum_{i=1}^{n} a_i/n = 0$ holds.
- Furthermore, it holds that

$$\sum_{i=1}^{n} \big(X_{(i)} - \overline{X}\big)^2 = \sum_{i=1}^{n} \big(X_i - \overline{X}\big)^2 \qquad \text{and} \qquad \sum_{i=1}^{n} a_i \overline{X} = 0\,,$$

i.e., the definition of $\widetilde{T}(X_1, \ldots, X_n)$ in (89) is equivalent to

$$\widetilde{T}(X_1, \ldots, X_n) = \frac{\sum\limits_{i=1}^{n} a_i X_{(i)}}{\sqrt{\sum\limits_{i=1}^{n} a_i^2}\sqrt{\sum\limits_{i=1}^{n} \big(X_i - \overline{X}\big)^2}}\,. \tag{90}$$

- Since it always holds that $|\widetilde{T}(X_1, \ldots, X_n)| \leq 1$, the null hypothesis H_0 is rejected if $\widetilde{T}(X_1, \ldots, X_n) < q_{n,\alpha}$, where $q_{n,\alpha}$ denotes the α–quantile of the distribution of $\widetilde{T}(X_1, \ldots, X_n)$.
- This is the so–called *Shapiro–Francia test* for normal distribution, where the quantiles $q_{n,\alpha}$ of the distribution of $\widetilde{T}(X_1, \ldots, X_n)$ can either be taken from a table or be determined using Monte Carlo simulation.

2. *Shapiro–Wilk test*

- In (90) it is possible to consider the linear transformation

$$(a'_1, \ldots, a'_n)^\top = \mathbf{K}^{-1}(a_1, \ldots, a_n)^\top$$

instead of $a_1, \ldots, a_n$, where the covariance matrix $\mathbf{K} = (k_{ij})$ is given by

$$k_{ij} = \mathbb{E}\left((Y_{(i)} - a_i)(Y_{(j)} - a_j)\right) \qquad \text{with } Y_i \sim \mathrm{N}(0,1).$$

- The test constructed in this way is called *Shapiro–Wilk test*.

6 Nonparametric Localization Tests

6.1 Two Simple Examples of One–Sample Problems

6.1.1 Binomial Test

- The χ^2 goodness–of–fit test considered in Section 5.2 can be replaced by the following *binomial test* if $r = 2$, i.e., if only two classes are considered (for example when dealing with binary data).
  - Then we partition the domain of the (independent and identically distributed) sampling variables $X_1, \ldots, X_n$ into two subsets $(a_1, b_1]$ and $(a_2, b_2]$, such that
$$(a_1, b_1] \cap (a_2, b_2] = \emptyset \qquad \text{and} \qquad \mathbb{P}\big(X_1 \in (a_1, b_1] \cup (a_2, b_2]\big) = 1\,,$$
    and consider the "class size"
$$T(X_1, \ldots, X_n) = \#\{i : 1 \le i \le n,\, a_1 < X_i \le b_1\}\,.$$
  - One can easily see that $T = T(X_1, \ldots, X_n)$ is binomial distributed, i.e.,
$$T \sim \mathrm{Bin}(n, p)\,, \qquad \text{where } p = \mathbb{P}(a_1 < X_1 \le b_1). \tag{1}$$
- To begin with, we consider the problem of testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, where $p_0 \in (0, 1)$ is an arbitrary fixed number.
  - Due to (1), H_0 is rejected if $T \le t_{\alpha_1}$ or $T \ge t_{1-\alpha_2}$,
  - where the "critical values" t_{α_1} and $t_{1-\alpha_2}$ for arbitrary $\alpha_1, \alpha_2 \in (0, 1)$ with $\alpha_1 + \alpha_2 = \alpha$ are given by
$$\begin{aligned} t_{\alpha_1} &= \max\{t \in \mathbb{R} : \mathbb{P}_{p_0}(T \le t) \le \alpha_1\} \\ &= \max\Big\{k \in \{0, 1, \ldots, n\} : \sum_{i=0}^{k} \binom{n}{i} p_0^i (1 - p_0)^{n-i} \le \alpha_1\Big\} \end{aligned}$$
    and
$$\begin{aligned} t_{1-\alpha_2} &= \min\{t \in \mathbb{R} : \mathbb{P}_{p_0}(T \ge t) \le \alpha_2\} \\ &= \min\Big\{k \in \{0, 1, \ldots, n\} : \sum_{i=k}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i} \le \alpha_2\Big\}\,. \end{aligned}$$
  - For $p_0 = 0.5$ one usually chooses $\alpha_1 = \alpha_2 = \alpha/2$. If p_0 is close to 0 or 1, it is advisable to choose α_1 smaller or greater than α_2, respectively.
  - The quantiles t_{α_1} and $t_{1-\alpha_2}$ of the binomial distribution $\mathrm{Bin}(n, p_0)$ can either be taken from tables or be determined using Monte Carlo simulation.
- The (one–sided) problem of testing $H_0 : p \le p_0$ versus $H_1 : p > p_0$ can be treated in a similar way. Here H_0 is rejected if $T \ge t_{1-\alpha}$.
- In an analogous way, one obtains a decision rule for the (one–sided) problem of testing $H_0 : p \ge p_0$ versus $H_1 : p < p_0$, where H_0 is rejected if $T \le t_\alpha$.

Remark

- The binomial test described above is also called *sign test* because the generation of 2 classes can be perceived as binarization of the original data.
- In the two one–sided problems of testing, the critical values $t_{1-\alpha}$ and t_α are determined for $p = p_0$ although the null hypothesis is $H_0 : p \le p_0$ or $H_0 : p \ge p_0$, respectively.

  - Considering the values $t_{1-\alpha}$ and t_α anyway does here not contradict the fact that for each $p < p_0$ or $p > p_0$ the critical value would be smaller than $t_{1-\alpha}$ or greater than t_α and that H_0 would thus have to be rejected more often.
  - The choice of the critical values $t_{1-\alpha}$ and t_α can be explained by the fact that one does not consider a single p with $p < p_0$ or $p > p_0$, but that p can be arbitrarily close to p_0 and hence, in particular, also $p = p_0$ is allowed.

- If the sample size n is large and if p_0 is close to 0 or 1,
  - the direct computation of the quantiles $t_{1-\alpha}$ and t_α of the binomial distribution $\mathrm{Bin}(n, p_0)$ is difficult.
  - The law of rare events (cf. Section WR–3.2.2) implies that $t_{1-\alpha}$ and t_α can be approximated by quantiles of the Poisson distribution $\mathrm{Poi}(\lambda)$ in this case, where $\lambda = np_0$ or $\lambda = n(1-p_0)$, respectively.
- Moreover, for arbitrary fixed $p_0 \in (0,1)$ the critical values $t_{1-\alpha}$ and t_α can be approximated by suitably transformed quantiles of the normal distribution $\mathrm{N}(0,1)$ if the sample size n is „sufficiently large".
  - In this case, it follows from the central limit theorem of DeMoivre–Laplace (cf. Theorem WR–3.6) that the transformed test statistic
$$T' = \frac{T - np_0}{\sqrt{np_0(1-p_0)}}$$
is approximately $\mathrm{N}(0,1)$–distributed, i.e., that
$$\mathbb{P}(T \le t) = \mathbb{P}(T' \le t') \approx \Phi(t')\,, \qquad \text{where} \quad t' = \frac{t - np_0}{\sqrt{np_0(1-p_0)}}$$
and $\Phi : \mathbb{R} \to [0,1]$ is the distribution function of the $\mathrm{N}(0,1)$–distribution.
  - Therefore, one gets that $t_\alpha \approx np_0 + z_\alpha\sqrt{np_0(1-p_0)}$, where z_α is the α–quantile of the $\mathrm{N}(0,1)$–distribution.
- Possible criteria for "sufficiently large" which are mentioned in literature are, e.g., the conditions $n \ge 20$ and $10 \le np_0 \le n - 10$.
  - When investigating the (two–sided) problem of testing $H_0 : p = p_0$ versus $H_1 : p \ne p_0$, then H_0 is rejected if
$$T \le np_0 + z_{\alpha_1}\sqrt{np_0(1-p_0)} \qquad \text{or} \qquad T \ge np_0 + z_{1-\alpha_2}\sqrt{np_0(1-p_0)}\,.$$
  - Similar approximation formulas can be derived for the critical values of the one–sided tests mentioned above.

Example

- Let the distribution function $F : \mathbb{R} \to [0,1]$ of the sampling variable $X_1, \ldots, X_n$ be continuous and let γ_p be the p–quantile of F, i.e., let $F(\gamma_p) = p$ for $p \in (0,1)$.
- In order to verify the hypothesis $H_0 : \gamma_p = \gamma_p^0$, one can consider the "coarsened" random sample $(Y_1, \ldots, Y_n)$ with
$$Y_i = \begin{cases} 1\,, & \text{if } X_i \le \gamma_p, \\ 0\,, & \text{if } X_i > \gamma_p. \end{cases}$$
- Then $Y_i \sim \mathrm{Bin}(1,p)$ for each $i = 1, \ldots, n$ and the hypothesis $H_0 : \gamma_p = \gamma_p^0$ with respect to $(X_1, \ldots, X_n)$ is equivalent to the hypothesis $H_0 : p = p_0$ with respect to $(Y_1, \ldots, Y_n)$.
- Hence, the binomial test can in particular be used to verify the hypothesis $H_0 : \gamma_{0.5} = 0$.

6.1.2 Run Test for Randomness

- In this section it is *not* assumed that the sampling variables $X_1, \ldots, X_n$ are independent.
  - We merely assume that $X_1, \ldots, X_n$ can only take the values 0 or 1, where the value 0 shall occur n_1 times and the value 1 shall occur n_2 times; $n_2 = n - n_1$.
  - Thus, there are altogether $\binom{n}{n_1}$ possible realizations of the random sample $(X_1, \ldots, X_n)$.
  - We now want to verify the null hypothesis H_0 that each of these $\binom{n}{n_1}$ realizations occurs with the same probability.
  - In other words: We want to check whether the localization, i.e., the order in which the n_1 ones and n_2 zeros are arranged, is "purely at random".
- As a test statistic $T : \Omega \to \{0, 1, \ldots\}$ we consider the number $T(\omega)$ of *runs* in the (concrete) sample $\omega = (x_1, \ldots, x_n)$, i.e., the number of (sub–) sequences of consecutive equal symbols $\omega = (x_1, \ldots, x_n)$.

Example

- Let $n = 20$ with $n_1 = 12$ and $n_2 = 8$. For
$$\omega = (1,1,0,0,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,1) \tag{2}$$
one then gets that $T(\omega) = 7$.
- We now investigate the question whether the data given in (2) is compatible with the hypothesis H_0 that the order is "purely at random" or whether H_0 should be rejected.
- For this purpose, we specify the distribution of T by considering a suitably chosen (Laplace) probability space, cf. Section WR–2.4.1.

Theorem 6.1 *Assuming that H_0 is true, it holds for each $i = 1, 2, \ldots, \min\{n_1, n_2\}$ that*

$$\mathbb{P}(T = k) = \begin{cases} \dfrac{2\binom{n_1 - 1}{i - 1}\binom{n_2 - 1}{i - 1}}{\binom{n}{n_1}}, & \text{wenn } k = 2i, \\[2ex] \dfrac{\binom{n_1 - 1}{i}\binom{n_2 - 1}{i - 1} + \binom{n_1 - 1}{i - 1}\binom{n_2 - 1}{i}}{\binom{n}{n_1}}, & \text{wenn } k = 2i + 1. \end{cases} \tag{3}$$

Furthermore, it holds that

$$\mathbb{E}\, T = 1 + \frac{2n_1 n_2}{n} \qquad \text{and} \qquad \operatorname{Var} T = \frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n - 1)} \,. \tag{4}$$

Proof

- We only prove (3) for the case $k = 2i$ since the proof for the case $k = 2i + 1$ proceeds analogously.
  - Hence, let $k = 2i$. Then there are at a time i runs consisting of ones and zeros, respectively.
  - For the decomposition of the n_1 zeros into i subsets, there are $\binom{n_1 - 1}{i - 1}$ possibilities.
  - For each of these decompositions, there are $\binom{n_2 - 1}{i - 1}$ possibilities to divide the n_2 ones into i subsets.

- If we now additionally assume that the sample $\omega = (x_1, \ldots, x_n)$ can begin either with $x_1 = 0$ or with $x_1 = 1$, we obtain a total of $2\binom{n_1-1}{i-1}\binom{n_2-1}{i-1}$ decomposition possibilities.
- Therefore, (3) is proved for the case $k = 2i$.

• In order to determine the expectation $\mathbb{E}\,T$, we use the following consideration.

- For each $j = 2, \ldots, n$ we consider the indicator variable $Y_j : \Omega \to \{0, 1\}$ with

$$Y_j = \begin{cases} 1\,, & \text{if a run starts at the } j\text{–th position,} \\ 0\,, & \text{else.} \end{cases}$$

- Then it holds that $\{\omega \in \Omega : Y_j(\omega) = 1\} = \{\omega \in \Omega : X_{j-1}(\omega) \neq X_j(\omega)\}$, i.e., there are $2\binom{n-2}{n_i-1}$ possibilities that a run starts at the j–th position.
- Therefore, one gets

$$\mathbb{E}\,Y_j \;=\; \mathbb{P}(Y_j = 1) \;=\; 2\,\frac{\binom{n-2}{n_i-1}}{\binom{n}{n_1}} \;=\; 2\,\frac{(n-2)!\,(n-n_1)!\,n_1!}{(n-n_1-1)!(n_1-1)!n!} \;=\; 2\,\frac{n_1(n-n_1)}{n(n-1)}\,.$$

- Together with the identity

$$T = 1 + \sum_{j=2}^{n} Y_j \tag{5}$$

this implies that

$$\mathbb{E}\,T \;=\; 1 + \sum_{j=2}^{n} \mathbb{E}\,Y_j \;=\; 1 + 2\,\frac{n_1(n-n_1)}{n}\,.$$

• The variance formula in (4) can be proved in a similar way because (5) implies that

$$\begin{aligned} \operatorname{Var} T \;&=\; \mathbb{E}\Big(\sum_{j=2}^{n} \mathbb{E}\,Y_j\Big)^2 - \Big(\sum_{j=2}^{n} \mathbb{E}\,Y_j\Big)^2 \\ &=\; \sum_{j=2}^{n} \mathbb{E}\,Y_j^2 \;+ \sum_{2 \le j_1, j_2 \le n,\, j_1 \neq j_2} \mathbb{E}\left(Y_{j_1} Y_{j_2}\right) \;-\; \Big(\sum_{j=2}^{n} \mathbb{E}\,Y_j\Big)^2 \\ &=\; \sum_{j=2}^{n} \mathbb{E}\,Y_j \;+ \sum_{2 \le j_1, j_2 \le n,\, j_1 \neq j_2} \mathbb{E}\left(Y_{j_1} Y_{j_2}\right) \;-\; \Big(\sum_{j=2}^{n} \mathbb{E}\,Y_j\Big)^2, \end{aligned}$$

hence one only needs to specify the moments $\mathbb{E}\left(Y_{j_1} Y_{j_2}\right)$. □

Remark

- A possible alternative to the null hypothesis H_0 that the localization of the zeros and ones is "purely at random" is their trend to form clumps or clusters.
- As rejection region of H_0 one then chooses the left–hand end of the distribution of T.
- In other words: H_0 is rejected if $T \le r_\alpha(n_1; n_2)$, where

$$r_\alpha(n_1; n_2) = \max\Big\{r \in \{1, 2, \ldots\} : \mathbb{P}(T \le r) \le \alpha\Big\}$$

is the α–quantile of the distribution of the test statistic T.

- The quantiles $r_\alpha(n_1; n_2)$ can be computed using the formulas for the probabilities $\mathbb{P}(T = k)$ given in Theorem 6.1. They can be taken from tables in literature.

Example (continued) For $\alpha = 0.1$ and $n_1 = 12$, $n_2 = 8$ one obtains that $r_{0.1}(12;8) = 7$. Moreover, it holds for the sample considered in (2) that

$$T(\omega) = 7 \ \Big(\leq r_{0.1}(12;8)\Big)\,,$$

i.e., H_0 is rejected.

If the (sub–) sample sizes n_1 and n_2 are large, the determination of the quantiles $r_\alpha(n_1;n_2)$ of $T = T_{n_1,n_2}$ involves a considerable computational cost. A way out is offered by the following central limit theorem, which we state without proof.

Theorem 6.2 *If n_1, $n_2 \to \infty$ such that $n_1/(n_1+n_2) \to p$ or equivalently $n_2/(n_1+n_2) \to 1-p$ for a $p \in (0,1)$, then it holds that*

$$\lim_{n_1,n_2\to\infty} \frac{1}{n_1+n_2}\,\mathbb{E}\,T_{n_1,n_2} = 2p(1-p) \qquad \textit{and} \qquad \lim_{n_1,n_2\to\infty} \frac{1}{n_1+n_2}\,\mathrm{Var}\,T_{n_1,n_2} = 4p^2(1-p)^2 \tag{6}$$

as well as

$$\lim_{n_1,n_2\to\infty} \mathbb{P}\Big(\frac{T_{n_1,n_2} - 2(n_1+n_2)p(1-p)}{2\sqrt{n_1+n_2}\,p(1-p)} \ \leq\ x\Big) = \Phi(x) \qquad \forall\, x \in \mathbb{R}\,, \tag{7}$$

where $\Phi : \mathbb{R} \to [0,1]$ is the distribution function of the $\mathrm{N}(0,1)$–distribution.

Remark Theorem 6.2 implies that for large n_1, n_2 the null hypothesis H_0 is rejected if

$$\frac{T_{n_1,n_2} - 2n_1n_2/(n_1+n_2)}{2n_1n_2/(n_1+n_2)^{3/2}} \ \leq\ z_\alpha\,, \tag{8}$$

where z_α is the α–quantile of the $\mathrm{N}(0,1)$–distribution.

6.2 Wilcoxon–Rank Test

6.2.1 Model Description; Median Test

- We now return to the case that the sampling variables $X_1, \ldots, X_n$ are independent and identically distributed with the distribution function $F : \mathbb{R} \to [0,1]$.
  - At the end of Section 6.1.1, in the context of the binomial or sign test, we have discussed a *median test* for verifying the hypothesis
    $$H_0 : \gamma_{0.5} = 0, \tag{9}$$
    where $\gamma_{0.5}$ is a median of F, i.e., $F(\gamma_{0.5}) = 0.5$.
  - In this section we consider another (more efficient) approach for testing the hypothesis given in (9).
- In doing so, we assume that the distribution function F of the sampling variables $X_1, \ldots, X_n$ belongs to the following (nonparametric) class of distribution functions.
  - Let $G : \mathbb{R} \to [0,1]$ be an arbitrary continuous distribution function with the following kind of symmetry with respect to the origin: For each $x \in \mathbb{R}$ it holds that $G(-x) = 1 - G(x)$.
  - This implies in particular that $G(0) = 1/2$, i.e., zero is a median of G.
  - Let the family Δ of distribution functions of the sampling variables $X_1, \ldots, X_n$ which is taken into account in the (two–sided) Wilcoxon test be given by $\Delta = \big\{F_\delta :\ F_\delta(x) = G(x-\delta)\ \forall\, x, \delta \in \mathbb{R}\big\}$.

– Since G is continuous, one then gets that

$$\mathbb{P}(X_i = x) = \mathbb{P}(X_i = X_j) = 0 \tag{10}$$

for each $x \in \mathbb{R}$ and for arbitrary $i, j = 1, \ldots, n$ with $i \neq j$.

- We discuss the (two–sided) problem of testing $H_0 : \delta = \delta_0$ vs. $H_1 : \delta \neq \delta_0$ for some $\delta_0 \in \mathbb{R}$.
  - Here we can (w.l.o.g.) set $\delta_0 = 0$; otherwise, the transformed sampling variables $X_1', \ldots, X_n'$ with $X_i' = X_i - \delta_0$ can be considered.
  - In a similar way, also the (one–sided) problem of testing $H_0 : \delta = 0$ vs. $H_1 : \delta > 0$ can be treated.
- For the verification of the null hypothesis $H_0 : \delta = 0$ we consider the *ranks* $R_1, \ldots, R_n$ of the random variables $|X_1|, \ldots, |X_n|$ with

$$R_i = \sum_{j=1}^{n} \mathbb{1}_{\{|X_j| \leq |X_i|\}} \qquad \forall\, i = 1, \ldots, n\,,$$

where the indicator variable $\mathbb{1}_{\{|X_j| \leq |X_i|\}} : \Omega \to \{0, 1\}$ is given by

$$\mathbb{1}_{\{|X_j| \leq |X_i|\}}(\omega) = \begin{cases} 1\,, & \text{if } |X_j(\omega)| \leq |X_i(\omega)|, \\ 0\,, & \text{else.} \end{cases}$$

- Then we consider the test statistics

$$T_n^+ = \sum_{i=1}^{n} R_i \mathbb{1}_{\{X_i > 0\}} \qquad \text{and} \qquad T_n^- = \sum_{i=1}^{n} R_i \mathbb{1}_{\{X_i < 0\}}\,. \tag{11}$$

Remark

- Due to (10) it holds with probability 1 that

$$T_n^- \;=\; \sum_{i=1}^{n} R_i - T_n^+ \;=\; \binom{n+1}{2} - T_n^+\,. \tag{12}$$

- One can show that, assuming that $H_0 : \delta = 0$ is true, it holds that $T_n^- \stackrel{\mathrm{d}}{=} T_n^+$; cf. (18).
- Thus, in the case that $H_0 : \delta = 0$ is true, the test statistics T_n^+ and T_n^- should take values that are approximately equal. Because of (12), this means that $T_n^+ \approx \binom{n+1}{2}/2$.
- Very small or very large values of T_n^+ hence indicate that the alternative hypothesis $H_1 : \delta \neq \delta_0$ might be true, i.e., $H_0 : \delta = 0$ is rejected if

$$T_n^+ \leq t_{\alpha/2} \qquad \text{or} \qquad T_n^+ \geq t_{1-\alpha/2}\,, \tag{13}$$

where the "critical values" $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are the $(\alpha/2)$–quantile and the $(1-\alpha/2)$–quantile of the distribution of T_n^+, respectively.

6.2.2 Distribution of the Test Statistic T_n^+ for Small Sample Sizes

- If the sample size n is not too large, then the quantiles $t_{\alpha/2}$ and $t_{1-\alpha/2}$ in (13) can be determined by combinatorial considerations.
  - Due to (10), the random vector $\mathbf{R} = (R_1, \ldots, R_n)$ of the ranks $R_1, \ldots, R_n$ of $|X_1|, \ldots, |X_n|$ is a (random) permutation of the numbers $1, \ldots, n$.

– Then the test statistic T_n^+ given in (11) can be represented as follows:

$$T_n^+ = \sum_{i=1}^{n} i\, Z_i\,, \qquad \text{where } Z_i = \mathbb{1}_{\{X_{R_i^{-1}}>0\}} \tag{14}$$

– and $\mathbf{R}^{-1} = (R_1^{-1}, \ldots, R_n^{-1})$ denotes the inverse permutation of $\mathbf{R}$, i.e., if $R_i = j$, then it holds that $R_j^{-1} = i$.

- Moreover, the following lemma is useful to determine the distribution of T_n^+.

Lemma 6.1 *Assuming that $H_0 : \delta = 0$ is true, it holds that:*

- *The random vectors $\big(\mathbb{1}_{\{X_1>0\}}, \ldots, \mathbb{1}_{\{X_n>0\}}\big)$ and $\mathbf{R} = (R_1, \ldots, R_n)$ are independent.*
- *The components $Z_1, \ldots, Z_n$ of $(Z_1, \ldots, Z_n)$ are independent and identically distributed with $Z_i \sim \mathrm{Bin}(1, 1/2)$.*

Proof

- We first show that the random variables $\mathbb{1}_{\{X_i>0\}}$ and $|X_i|$ are independent for each $i = 1, \ldots, n$.
  – For each $x \geq 0$ it holds that

$$\mathbb{P}\big(\mathbb{1}_{\{X_i>0\}} = 1,\, |X_i| \leq x\big) \;=\; \mathbb{P}\big(0 < X_i \leq x\big) \;=\; G(x) - \frac{1}{2}$$

  and

$$\mathbb{P}\big(\mathbb{1}_{\{X_i>0\}} = 1\big)\,\mathbb{P}\big(|X_i| \leq x\big) \;=\; \frac{1}{2}\,\big(G(x) - G(-x)\big) \;=\; \frac{1}{2}\,\big(G(x) - (1 - G(x))\big) \;=\; G(x) - \frac{1}{2}\,.$$

  – Moreover, it obviously holds for each $x < 0$ that

$$\mathbb{P}\big(\mathbb{1}_{\{X_i>0\}} = 1,\, |X_i| \leq x\big) \;=\; 0 \;=\; \mathbb{P}\big(\mathbb{1}_{\{X_i>0\}} = 1\big)\,\mathbb{P}\big(|X_i| \leq x\big)\,.$$

  – In the same way, it can be shown that

$$\mathbb{P}\big(\mathbb{1}_{\{X_i>0\}} = 0,\, |X_i| \leq x\big) \;=\; \mathbb{P}\big(\mathbb{1}_{\{X_i>0\}} = 0\big)\,\mathbb{P}\big(|X_i| \leq x\big)$$

  for each $x \in \mathbb{R}$.
- Since the independence of the sampling variables $X_1, \ldots, X_n$ implies the independence of the random vectors $\big(\mathbb{1}_{\{X_1>0\}}, |X_1|\big), \ldots, \big(\mathbb{1}_{\{X_n>0\}}, |X_n|\big)$,
  – it follows that $\big(\mathbb{1}_{\{X_1>0\}}, \ldots, \mathbb{1}_{\{X_n>0\}}\big)$ and $\big(|X_1|, \ldots, |X_n|\big)$ are independent random vectors.
  – Since $\mathbf{R} = (R_1, \ldots, R_n)$ is a Borel measurable function of $\big(|X_1|, \ldots, |X_n|\big)$, also the random vectors $\big(\mathbb{1}_{\{X_1>0\}}, \ldots, \mathbb{1}_{\{X_n>0\}}\big)$ and $\mathbf{R}$ are independent.
- Therefore, one gets for arbitrary $i \in \{1, \ldots, n\}$ and $z \in \{0, 1\}$ that

$$\begin{aligned}
\mathbb{P}(Z_i = z) \;&=\; \mathbb{P}\big(\mathbb{1}_{\{X_{R_i^{-1}}>0\}} = z\big) \\
&=\; \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_{R_i^{-1}}>0\}} = z \mid \mathbf{R} = \mathbf{r}\big)\,\mathbb{P}(\mathbf{R} = \mathbf{r}) \\
&=\; \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_{r_i^{-1}}>0\}} = z \mid \mathbf{R} = \mathbf{r}\big)\,\mathbb{P}(\mathbf{R} = \mathbf{r}) \\
&=\; \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_{r_i^{-1}}>0\}} = z\big)\,\mathbb{P}(\mathbf{R} = \mathbf{r}) \;=\; \frac{1}{2}\sum_{\mathbf{r}} \mathbb{P}(\mathbf{R} = \mathbf{r}) \;=\; \frac{1}{2}\,,
\end{aligned}$$

  where the summation extends over all permutations $\mathbf{r} = (r_1, \ldots, r_n)$ of the numbers $1, \ldots, n$.

- This implies that

$$\begin{aligned}
\mathbb{P}\big(Z_1 = z_1, \ldots, Z_n = z_n\big) &= \mathbb{P}\big(\mathbb{1}_{\{X_{R_1^{-1}}>0\}} = z_1, \ldots, \mathbb{1}_{\{X_{R_n^{-1}}>0\}} = z_n\big) \\
&= \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_{R_1^{-1}}>0\}} = z_1, \ldots, \mathbb{1}_{\{X_{R_n^{-1}}>0\}} = z_n \mid \mathbf{R} = \mathbf{r}\big)\, \mathbb{P}\big(\mathbf{R} = \mathbf{r}\big) \\
&= \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_{r_1^{-1}}>0\}} = z_1, \ldots, \mathbb{1}_{\{X_{r_n^{-1}}>0\}} = z_n \mid \mathbf{R} = \mathbf{r}\big)\, \mathbb{P}\big(\mathbf{R} = \mathbf{r}\big) \\
&= \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_1>0\}} = z_{r_1}, \ldots, \mathbb{1}_{\{X_n>0\}} = z_{r_n} \mid \mathbf{R} = \mathbf{r}\big)\, \mathbb{P}\big(\mathbf{R} = \mathbf{r}\big) \\
&= \sum_{\mathbf{r}} \mathbb{P}\big(\mathbb{1}_{\{X_1>0\}} = z_{r_1}, \ldots, \mathbb{1}_{\{X_n>0\}} = z_{r_n}\big)\, \mathbb{P}\big(\mathbf{R} = \mathbf{r}\big) \\
&= \frac{1}{2^n} \;=\; \mathbb{P}(Z_1 = z_1) \,\ldots\, \mathbb{P}(Z_n = z_n)
\end{aligned}$$

  for arbitrary $\mathbf{z} = (z_1, \ldots, z_n) \in \{0,1\}^n$.

Theorem 6.3 *Assuming that $H_0 : \delta = 0$ is true, the distribution of T_n^+ is given by*

$$\mathbb{P}(T_n^+ = k) \;=\; \frac{a_k}{2^n} \qquad \forall\, k = 0, 1, \ldots, n\,, \tag{15}$$

where

$$a_k = \#\Big\{\mathbf{z} = (z_1, \ldots, z_n) \in \{0,1\}^n : \sum_{i=1}^n i z_i = k\Big\}\,. \tag{16}$$

Moreover, it holds that

$$\mathbb{E}\, T_n^+ \;=\; \frac{n(n+1)}{4} \qquad \textit{and} \qquad \operatorname{Var} T_n^+ \;=\; \frac{n(n+1)(2n+1)}{24}\,. \tag{17}$$

Proof

- The representation formula (14) for T_n^+ and Lemma 6.1 imply that

$$\mathbb{P}(T_n^+ = k) \;=\; \mathbb{P}\Big(\sum_{i=1}^n i\, Z_i = k\Big) \;=\; \sum_{\mathbf{z}=(z_1,\ldots,z_n)\in\{0,1\}^n:\, \sum_{i=1}^n i z_i = k} \mathbb{P}\big(Z_1 = z_1, \ldots, Z_n = z_n\big) \;=\; \frac{a_k}{2^n}$$

  for each $k = 0, 1, \ldots, n$.
- Furthermore, one obtains that

$$\mathbb{E}\, T_n^+ \;=\; \mathbb{E}\,\Big(\sum_{i=1}^n i\, Z_i\Big) \;=\; \sum_{i=1}^n i\, \mathbb{E}\, Z_i \;=\; \frac{1}{2}\binom{n+1}{2} \;=\; \frac{n(n+1)}{4}$$

  and

$$\operatorname{Var} T_n^+ \;=\; \operatorname{Var}\Big(\sum_{i=1}^n i\, Z_i\Big) \;=\; \sum_{i=1}^n i^2 \operatorname{Var} Z_i \;=\; \frac{1}{4}\sum_{i=1}^n i^2 \;=\; \frac{n(n+1)(2n+1)}{24}\,. \qquad \square$$

Remark

- From (12) and (14) it moreover follows with Lemma 6.1 that

$$T_n^- \;=\; \binom{n+1}{2} - T_n^+ \;=\; \binom{n+1}{2} - \sum_{i=1}^n i\, Z_i \;=\; \sum_{i=1}^n i\,(1 - Z_i) \;\stackrel{\mathrm{d}}{=}\; \sum_{i=1}^n i\, Z_i \;=\; T_n^+\,,$$

  i.e., under $H_0 : \delta = 0$ it holds that

$$T_n^- \;\stackrel{\mathrm{d}}{=}\; T_n^+\,. \tag{18}$$

- Thus, (12) implies that for each $k = 0, 1, \ldots, n$

$$\mathbb{P}\big(T_n^+ = k\big) \;=\; \mathbb{P}\Big(T_n^- = \frac{n(n+1)}{2} - k\Big) \;=\; \mathbb{P}\Big(T_n^+ = \frac{n(n+1)}{2} - k\Big)\,,$$

  i.e., the distribution of T_n^+ is symmetric with respect to the expectation $\mathbb{E}\,T_n^+ = n(n+1)/4$.
- This means that also the quantiles $t_{\alpha,n} = \max\{t \in \mathbb{R} : \mathbb{P}(T_n^+ \le t) \le \alpha\}$ have this property of symmetry, i.e., for each $\alpha \in (0,1)$ it holds that

$$t_{\alpha,n} \;=\; \frac{n(n+1)}{2} \;- t_{1-\alpha,n}\,.$$

- The quantiles $t_{\alpha,n}$ can either be taken from tables or be determined using Monte Carlo simulation.

6.2.3 Asymptotic Distribution

- The direct determination of the quantiles $t_{\alpha/2}$ and $t_{1-\alpha/2}$ by using Theorem 6.3 is difficult if the sample size n is large.
  - Another way to approximatively determine the distribution of the test statistic T_n^+ is based on the representation formula (14).
  - In this context the fact is used that $T_n^+ = \sum_{i=1}^n i\,Z_i$ is a sum of independent random variables, which follows from Lemma 6.1.
  - In fact, the central limit theorem for sums of independent (but not necessary identically distributed) random variables implies that T_n^+ is normally distributed.
- For this purpose we consider the following stochastic model: For each $n \ge 1$ let $X_{n1}, \ldots, X_{nn} : \Omega \to \mathbb{R}$ be a sequence of independent random variables,
  - where we (w.l.o.g.) assume that for each $k \in \{1, \ldots, n\}$

$$\mathbb{E}\,X_{nk} = 0\,, \qquad 0 < \sigma_{nk}^2 = \mathrm{Var}\,X_{nk} < \infty \qquad \text{and} \qquad \sum_{k=1}^n \sigma_{nk}^2 = 1\,. \tag{19}$$

  - If the random variables $X_{n1}, \ldots, X_{nn}$ do not satisfy the conditions formulated in (19), then we consider the transformed random variables $X'_{n1}, \ldots, X'_{nn}$ with

$$X'_{nk} = \frac{X_{nk} - \mathbb{E}\,X_{nk}}{\sqrt{n \mathrm{Var}\,X_{nk}}}\,. \tag{20}$$

  - We denote the distribution function of X_{nk} by F_{nk}, where we do not exclude the case that F_{nk} for each $k \in \{1, \ldots, n\}$ can depend on the number n of considered random variables $X_{n1}, \ldots, X_{nn}$.

The following *central limit theorem of Lindeberg* (cf.Theorem WR–5.22) is the basis to show that T_n^+ is asymptotically normally distributed.

Lemma 6.2

- *For each $n \in \mathbb{N}$ let $X_{n1}, \ldots, X_{nn} : \Omega \to \mathbb{R}$ be a sequence of independent random variables, which satisfy the conditions* (19).

- *If furthermore for each* $\varepsilon > 0$

$$\lim_{n\to\infty} \sum_{k=1}^{n} \int_{\mathbb{R}\setminus(-\varepsilon,\varepsilon)} x^2 \, dF_{nk}(x) = 0\,, \tag{21}$$

then it holds for each $x \in \mathbb{R}$ *that*

$$\lim_{n\to\infty} P\big(X_{n1} + \ldots + X_{nn} \le x\big) = \Phi(x)\,, \tag{22}$$

where $\Phi : \mathbb{R} \to [0,1]$ *is the distribution function of the* $\mathrm{N}(0,1)$*–distribution.*

Theorem 6.4 *Under* $H_0 : \delta = 0$ *it holds that*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{T_n^+ - \mathbb{E}\,T_n^+}{\sqrt{\mathrm{Var}\,T_n^+}} \le x\Big) \;=\; \Phi(x) \qquad \forall\, x \in \mathbb{R}\,. \tag{23}$$

Proof

- Because of (14) it is sufficient to show that the random variables $X_{n1}, \ldots, X_{nn}$ with

$$X_{nk} = \frac{kZ_k - k\,\mathbb{E}\,Z_k}{\sqrt{\mathrm{Var}\,T_n^+}} \tag{24}$$

satisfy the conditions of Lemma 6.2.
  - It follows directly from equation (24) that (19) is fulfilled.
  - Therefore, it merely remains to show that the Lindeberg–condition (22) is satisfied.
- For the distribution function $F_{nk} : \mathbb{R} \to [0,1]$ of the random variable X_{nk}, introduced in (24), it follows from Lemma 6.1 that

$$F_{nk}(x) = \begin{cases} 0\,, & \text{if } x < \dfrac{-k}{2\sqrt{\mathrm{Var}\,T_n^+}}\,, \\ \dfrac{1}{2}\,, & \text{if } \dfrac{-k}{2\sqrt{\mathrm{Var}\,T_n^+}} \le x < \dfrac{k}{2\sqrt{\mathrm{Var}\,T_n^+}}\,, \\ 1\,, & \text{if } \dfrac{k}{2\sqrt{\mathrm{Var}\,T_n^+}} \le x\,. \end{cases}$$

  - This implies that $\int_{\mathbb{R}\setminus(-\varepsilon,\varepsilon)} x^2 \, dF_{nk}(x) = 0$ for each $k \in \{1, \ldots, n\}$ if n is chosen in such a way that

$$\frac{n^2}{4\mathrm{Var}\,T_n^+} \;=\; \frac{6n^2}{n(n+1)(2n+1)} \;<\; \varepsilon^2\,,$$

    where the last equality follows from the formula for $\mathrm{Var}\,T_n^+$ in Theorem 6.3.
  - Therefore, the validity of the Lindeberg–condition (22) is shown. □

Remark

- Because of Theorem 6.4 the following critical area is considered in the case of the (two–sided) test problem $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$.
- For sufficiently large n, $H_0 : \delta = 0$ is rejected if

$$\Big|\, \frac{T_n^+ - \mathbb{E}\,T_n^+}{\sqrt{\mathrm{Var}\,T_n^+}} \,\Big| \ge z_{1-\alpha/2}\,, \tag{25}$$

where $\mathbb{E}\,T_n^+$ or $\mathrm{Var}\,T_n^+$ are given in Theorem 6.3 and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$–quantile of the $\mathrm{N}(0,1)$–distribution.
- In the literature the condition $n \ge 20$ is suggested as a possible criterion for "sufficiently large".

6.3 Two–Sample Problems

- In this section we discuss nonparametric tests for the case that two independent random samples $(X_1, \ldots, X_{n_1})$ and $(Y_1, \ldots, Y_{n_2})$ are observed.
- In other words: We assume that the random variables $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ are completely independent with the (unknown) distribution functions F and G, i.e.,

$$F(x) = \mathbb{P}(X_i \leq x) \qquad \text{and} \qquad G(y) = \mathbb{P}(Y_j \leq y) \qquad \forall\, x, y \in \mathbb{R},\ i = 1, \ldots, n,\ j = 1, \ldots, m\,.$$

- Then, a (two–sided) test problem is for example given by

$$H_0 : F(x) = G(x) \quad \forall\, x \in \mathbb{R} \qquad \text{vs.} \qquad H_1 : F(x) \neq G(x) \quad \exists\, x \in \mathbb{R}\,. \tag{26}$$

- As a one–sided alternative to $H_0 : F(x) = G(x)\ \forall\, x \in \mathbb{R}$ the following hypotheses can be considered:

$$H_1 : F(x) \geq G(x) \quad \forall\, x \in \mathbb{R} \quad \text{and} \quad F(x) > G(x) \quad \exists\, x \in \mathbb{R} \tag{27}$$

or

$$H_1 : F(x) \leq G(x) \quad \forall\, x \in \mathbb{R} \quad \text{and} \quad F(x) < G \quad \exists\, x \in \mathbb{R}\,. \tag{28}$$

6.3.1 Run Test of Wald–Wolfowitz

- For the analysis of the test problem, given in (26), one can apply the run test for randomness, which has been discussed in Section 6.1.2.
  - For this purpose, we combine the sample variables $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ to one random sample

$$(X'_1, \ldots, X'_n) = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})\,, \qquad \text{where } n = n_1 + n_2,$$

    and consider the ordered sample $(X'_{(1)}, \ldots, X'_{(n)})$.
  - Here we assume that the distribution functions F and G are continuous, i.e., the mapping

$$(X'_1, \ldots, X'_n) \mapsto (X'_{(1)}, \ldots, X'_{(n)})$$

    is uniquely determined with probability 1.
- Under $H_0 : F(x) = G(x)\ \forall\, x \in \mathbb{R}$ it is to be expected that the X_i's and Y_j's in $(X'_{(1)}, \ldots, X'_{(n)})$ are "well mixed",
  - since the sample variables $X'_{(1)}, \ldots, X'_{(n)}$ then are independent and identically distributed.
  - If the trend for "clumping and clustering" is considered as an alternative, then H_0 is rejected if the number T of iterations in the (binary) sample $(Z_1, \ldots, Z_n)$ is "too small", where $Z_i = 0$ if $X'_{(i)} = X_j$ for some $j \in \{1, \ldots, n\}$ and $Z_i = 1$ if $X'_{(i)} = Y_j$ for some $j \in \{1, \ldots, n\}$.

Examples

- In a medical study the body heights of $n_1 = 8$ girls and $n_2 = 10$ boys were analyzed.
- The measurement results are:

x_i	117	121	122	124	125	126	128	132		
y_j	110	113	114	115	116	118	119	120	123	127

- If we order these measurements by size and assign a 0 to the heights of the boys and a 1 to the heights of the girls, then we obtain the sequence

$$\omega = (0,0,0,0,0,1,0,0,0,1,1,0,1,1,1,0,1,1) \qquad \text{with } T(\omega) = 8. \tag{29}$$

- Otherwise, Theorem 6.1 implies that for the α–quantile $r_\alpha(n_1; n_2)$ of the distribution of T it holds that $r_{0.05}(8; 10) = 6$ for $\alpha = 0.05$.
- In this case H_0 is therefore rejected because $T(\omega) = 8 > 6 = r_{0.05}(8; 10)$.

Remark

- The run test considered in this section is not able to identify alternatives of the type (27) or (28).
- The example given in (29) makes this clear: Since the number of iterations $T(\omega) = 8$ does not change, if we (in contrast to the previous approach) assign a 1 to the heights of the boys and a 0 to the heights of the girls.
- Also for two–sided alternatives the run test of Wald–Wolfowitz, also called "omnibus–test", should only be used if the form of the alternative is not specified further.
- For special alternatives, which for example only affect location or variability characteristics, other test methods are more efficient, cf Section 6.3.2.

6.3.2 Wilcoxon Rank–Sum Test for Location Alternatives

- We now discuss another nonparametric test for the case that two independent random samples $(X_1, \ldots, X_{n_1})$ and $(Y_1, \ldots, Y_{n_2})$ are observed.
- However, we will here consider more special alternatives as in (26) – (28).
  - We assume that the random variables $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are completely independent with the (unknown) continuous distribution functions F and G.
  - Similar as in Section 6.2 it is assumed that there is some $\delta \in \mathbb{R}$ such that

$$F(x) = G(x + \delta) \qquad \forall\, x \in \mathbb{R}\,.$$

  - A (two–sided) test problem, which is consistent with the above mentioned more general test problem (26), is then given by

$$H_0 : \delta = 0 \qquad \text{vs.} \qquad H_1 : \delta \neq 0\,. \tag{30}$$

  - The following hypotheses can be considered as one–sided alternatives to $H_0 : \delta = 0$:

$$H_1 : \delta > 0 \qquad \text{or} \qquad H_1 : \delta < 0\,. \tag{31}$$

- In the same way as in Section 6.3.1 we merge the sample variables $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ to one combined random sample $(X'_1, \ldots, X'_n) = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$, where $n = n_1 + n_2$.
  - Furthermore, we consider the (random) vector of the ranks $\mathbf{R}' = (R'_1, \ldots, R'_n)$ of the sample variables $X'_1, \ldots, X'_n$ in the combined sample, where

$$R'_i = \sum_{j=1}^{n} \mathbb{1}_{\{X'_j \leq X'_i\}} \qquad \forall\, i = 1, \ldots, n\,.$$

  - As in Section 6.3.1 it has to be expected under $H_0 : \delta = 0$ that the X_i's and Y_j's in the combined sample $(X'_{(1)}, \ldots, X'_{(n)})$ are "well mixed" because then the sample variables $X'_{(1)}, \ldots, X'_{(n)}$ are independent and identically distributed.

– Thus, for the two–sided test problem in (30), H_0 is rejected if the *rank–sum*

$$T_{n_1,n_2} = \sum_{i=1}^{n_1} R_i' \tag{32}$$

is "too small" or "too large".

- In order to perform the test, the distribution of the test statistic T_{n_1,n_2}, introduced in (32), has to be determined. For this purpose the following lemma is useful.

Lemma 6.3

- *Let $X : \Omega \to \{\ldots,-1,0,1,\ldots\}$ be a discrete random variable such that $\mathbb{E}\,|X| < \infty$ and that for some $\mu \in \mathbb{R}$ the following symmetry property is fulfilled:*

$$\mathbb{P}(X = \mu - k) = \mathbb{P}(X = \mu + k) \qquad \forall\, k \in \{\ldots,-1,0,1,\ldots\}\,. \tag{33}$$

- *Then it holds that $\mathbb{E}\,X = \mu$.*

Proof

- We can w.l.o.g. assume that $\mu = 0$ since otherwise the transformed random variable $X' = X - \mu$ can be considered.
- Then it follows from (33) with $\mu = 0$ that

$$\mathbb{E}\,X = \sum_{k=-\infty}^{\infty} k\,\mathbb{P}(X=k) \;=\; -\sum_{k=1}^{\infty} k\,\mathbb{P}(X=-k) \;+\; \sum_{k=1}^{\infty} k\,\mathbb{P}(X=k) \;\overset{(33)}{=}\; 0\,.$$

□

Theorem 6.5

- *Under $H_0 : \delta = 0$ the distribution of T_{n_1,n_2} is given by*

$$\mathbb{P}(T_{n_1,n_2} = k) \;=\; \frac{a_{k,n_1,n_2}}{\binom{n_1+n_2}{n_1}} \qquad \forall\, k = \frac{n_1(n_1+1)}{2},\ldots,n_1n_2 + \frac{n_1(n_1+1)}{2}\,, \tag{34}$$

where

$$a_{k,n_1,n_2} = \#\Big\{\mathbf{z} = (z_1,\ldots,z_{n_1+n_2}) \in \{0,1\}^{n_1+n_2} : \#\{i : z_i = 1\} = n_1,\ \sum_{i=1}^{n_1+n_2} i z_i = k\Big\}\,. \tag{35}$$

- *Furthermore, it holds that*

$$\mathbb{P}(T_{n_1,n_2} = k) \;=\; \mathbb{P}(T_{n_1,n_2} = 2\mu - k) \qquad \forall\, k \in \{\ldots,-1,0,1,\ldots\} \tag{36}$$

and therefore

$$\mathbb{E}\,T_{n_1,n_2} = \mu\,, \tag{37}$$

where $\mu = n_1(n_1+n_2+1)/2$.

Proof

- Under $H_0 : \delta = 0$ the sample variables $X_1', \ldots, X_{n_1+n_2}'$ are independent and identically distributed.

- Hence, each of the $\binom{n_1+n_2}{n_1}$ partitions of the n_1 variables $X_1, \ldots, X_{n_1}$ into the $n_1 + n_2$ existing rank spots has the same probability.
- Moreover, it holds for the minimum and maximum value $t_{\min}$ and $t_{\max}$, respectively, of T_{n_1,n_2} that

$$t_{\min} = \sum_{i=1}^{n_1} i \;=\; \frac{n_1(n_1+1)}{2} \qquad \text{and} \qquad t_{\max} = \sum_{i=n_2+1}^{n_2+n_1} i \;=\; n_1 n_2 \;+\; \frac{n_1(n_1+1)}{2}\,.$$

- From this the validity of (34) – (35) is obtained.

- In order to prove (36) we use the following symmetry property.
  - Each $\mathbf{z} = (z_1, \ldots, z_{n_1+n_2}) \in \{0,1\}^{n_1+n_2}$ with

$$\#\{i : z_i = 1\} = n_1 \qquad \text{and} \qquad \sum_{i=1}^{n_1+n_2} i z_i = k$$

  corresponds to some $\widetilde{\mathbf{z}} = (\widetilde{z}_1, \ldots, \widetilde{z}_{n_1+n_2}) \in \{0,1\}^{n_1+n_2}$ with

$$\#\{i : \widetilde{z}_{n_1+n_2+1-i} = 1\} = n_1 \qquad \text{and} \qquad \sum_{i=1}^{n_1+n_2} (n_1+n_2+1-i)\,\widetilde{z}_i = n_1(n_1+n_2+1) - k\,.$$

  - Since the sample variables $X'_{(1)}, \ldots, X'_{(n)}$ are independent and identically distributed, it thus follows for each $k \in \{\ldots, -1, 0, 1, \ldots\}$ that

$$\mathbb{P}(T_{n_1,n_2} = k) = \mathbb{P}(T_{n_1,n_2} = n_1(n_1+n_2+1) - k) = \mathbb{P}(T_{n_1,n_2} = 2\mu - k)\,, \tag{38}$$

  where $2\mu = n_1(n_1+n_2+1)$.

- In order to show (37) it is sufficient to substitute $k = \mu - i$ in (38).
  - Then (38) implies that

$$\mathbb{P}(T_{n_1,n_2} = \mu - i) = \mathbb{P}(T_{n_1,n_2} = \mu + i) \qquad \forall\, i \in \{\ldots, -1, 0, 1, \ldots\}\,.$$

  - From this and Lemma 6.3 the validity of (37) is obtained. □

Remark

- Now, (38) implies the following symmetry property for the quantiles t_{α,n_1,n_2} of T_{n_1,n_2}.
  - For each $\alpha \in (0,1)$ it holds that

$$t_{\alpha,n_1,n_2} \;=\; n_1(n_1+n_2+1) - t_{1-\alpha,n_1,n_2}\,.$$

  - The quantiles t_{α,n_1,n_2} can either be taken from tables or be determined using Monte Carlo simulation.
- The null hypothesis $H_0 : \delta = 0$ is rejected in favor of $H_1 : \delta \neq 0$ if

$$T_{n_1,n_2} \leq t_{\alpha/2,n_1,n_2} \qquad \text{or} \qquad T_{n_1,n_2} \geq n_1(n_1+n_2+1) - t_{\alpha/2,n_1,n_2}\,.$$

- Analogously, the null hypothesis $H_0 : \delta = 0$ is rejected in favor of $H_1 : \delta < 0$ or $H_1 : \delta > 0$ if

$$T_{n_1,n_2} \geq n_1(n_1+n_2+1) - t_{\alpha,n_1,n_2} \qquad \text{or} \qquad T_{n_1,n_2} \leq t_{\alpha,n_1,n_2}\,.$$

If the sample sizes n_1 and n_2 are large enough, it is difficult to determine the quantiles t_{α,n_1,n_2} directly via Theorem 6.5. However, then the distribution of the test statistic T_{n_1,n_2} can approximatively be determined from the following central limit theorem.

Theorem 6.6 *If $n_1, n_2 \to \infty$ such that $n_1/(n_1+n_2) \to p$ or $n_2/(n_1+n_2) \to 1-p$ for some $p \in (0,1)$, then it holds that*

$$\lim_{n_1,n_2\to\infty} \mathbb{P}\Big(\frac{T_{n_1,n_2} - \mathbb{E}\, T_{n_1,n_2}}{\sqrt{\mathrm{Var}\, T_{n_1,n_2}}} \le x \Big) = \Phi(x) \qquad \forall\, x \in \mathbb{R}\,, \tag{39}$$

where

$$\mathbb{E}\, T_{n_1,n_2} = \frac{n_1(n_1+n_2+1)}{2}\,, \qquad \mathrm{Var}\, T_{n_1,n_2} = \frac{n_1 n_2 (n_1+n_2+1)}{12}$$

and $\Phi : \mathbb{R} \to [0,1]$ is the distribution function of the $\mathrm{N}(0,1)$–distribution.

Remark

- Because of Theorem 6.6 the null hypothesis $H_0 : \delta = 0$ is rejected for large n_1, n_2 in favor of $H_1 : \delta \neq 0$ if
$$\Big| \frac{T_{n_1,n_2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \Big| \ge z_{1-\alpha/2}\,,$$
where z_α is the α–quantile of the $\mathrm{N}(0,1)$–distribution.
- Analogously, $H_0 : \delta = 0$ is rejected in favor of $H_1 : \delta > 0$ or $H_1 : \delta < 0$ if
$$\frac{T_{n_1,n_2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \ge z_{1-\alpha} \qquad \text{or} \qquad \frac{T_{n_1,n_2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \le -z_{1-\alpha}\,.$$