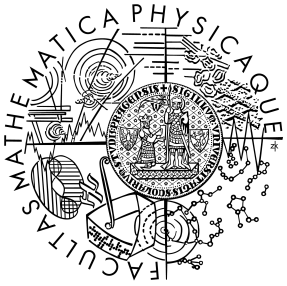
Department of Probability and Mathematical Statistics



FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

NMSA407 Linear Regression

Course Notes

2016–17

Arnošt Komárek

Last modified on December 8, 2016.

These course notes contain an overview of notation, definitions, theorems and comments covered by the course "NMSA407 Linear Regression", which is a part of the curriculum of the Master's programs "Probability, Mathematical Statistics and Econometrics" and "Financial and Insurance Mathematics".

This document undergoes continuing development.

This version is dated December 8, 2016.

Arnošt Komárek

komarek@karlin.mff.cuni.cz

On Řečička, in Karlín from May 2015, partially based on lecture overheads used in fall 2013 and 2014.

Contents

Preface

- Basic literature: Khuri (2010); Zvára (2008).
- Supplementary literature: Seber and Lee (2003); Draper and Smith (1998); Sun (2003); Weisberg (2005); Anděl (2007); Cipra (2008); Zvára (1989).
- Principal computational environment: R software (R Core Team, 2016).

Notation and general conventions

General conventions

- Vectors are understood as *column* vectors (matrices with one column).
- Statements concerning equalities between two random quantities are understood as equalities *almost surely* even if "almost surely" is not explicitely stated.
- Measurability is understood with respect to the Borel σ-algebra on the Euclidean space.

General notation

- $Y \sim \big(\mu,\, \sigma^2\big)$ means that the random variable Y follows a distribution satisfying

$$\mathbb{E}\big(Y\big) = \mu, \quad \mathsf{var}\big(Y\big) = \sigma^2.$$

- $\boldsymbol{Y} \sim \big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)$ means that the random vector $\boldsymbol{Y}$ follows a distribution satisfying

$$\mathbb{E}\big(\boldsymbol{Y}\big) = \boldsymbol{\mu}, \quad \mathsf{var}\big(\boldsymbol{Y}\big) = \boldsymbol{\Sigma}.$$

Notation related to the linear model

- Generic response random variable: covariate random vector (length p), regressor random vector (length k, elements indexed from 0):

$$Y, \quad \boldsymbol{Z} = \big(Z_1,\, \ldots,\, Z_p\big)^\top, \quad \boldsymbol{X} = \big(X_0,\, \ldots,\, X_{k-1}\big)^\top.$$

- Response vector (length n): $\boldsymbol{Y} = \big(Y_1,\, \ldots,\, Y_n\big)^\top$.
- Covariates (p covariates):
  - $\boldsymbol{Z}_i = \big(Z_{i,1},\, \ldots,\, Z_{i,p}\big)^\top$ $(i = 1,\, \ldots,\, n)$:
    vector of covariates for observation i;
  - $\boldsymbol{Z}^j = \big(Z_{1,j},\, \ldots,\, Z_{n,j}\big)^\top$ $(j = 0,\, \ldots,\, p)$:
    values of the jth covariate for n observations.

- Covariate matrix (dimension $n \times p$):

$$\mathbb{Z} = \begin{pmatrix} Z_{1,1} & \dots & Z_{1,p} \\ \vdots & \vdots & \vdots \\ Z_{n,1} & \dots & Z_{n,p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_1^\top \\ \vdots \\ \boldsymbol{Z}_n^\top \end{pmatrix} = \big(\boldsymbol{Z}^1,\, \dots,\, \boldsymbol{Z}^p\big).$$

- Regressors (k regressors indexed from 0):
  - $\boldsymbol{X}_i = \big(X_{i,0},\, \dots,\, X_{i,k-1}\big)^\top$ $(i = 1,\, \dots,\, n)$:
    vector of regressors for observation i;
  - $\boldsymbol{X}^j = \big(X_{1,j},\, \dots,\, X_{n,j}\big)^\top$ $(j = 0,\, \dots,\, k-1)$:
    values of the jth regressor for n observations.

- Model matrix (dimension $n \times k$):

$$\mathbb{X} = \begin{pmatrix} X_{1,0} & \dots & X_{1,k-1} \\ \vdots & \vdots & \vdots \\ X_{n,0} & \dots & X_{n,k-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \big(\boldsymbol{X}^0,\, \dots,\, \boldsymbol{X}^{k-1}\big).$$

- Rank of the model: $r = \mathsf{rank}(\mathbb{X})$ $(\leq k < n)$ (almost surely).
- Error terms: $\boldsymbol{\varepsilon} = \big(\varepsilon_1,\, \dots,\, \varepsilon_n\big)^\top = \big(Y_1 - \boldsymbol{X}_1^\top\boldsymbol{\beta},\, \dots,\, Y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\big)^\top = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$.
- Regression space: $\mathcal{M}\big(\mathbb{X}\big)$ (linear span of columns of $\mathbb{X}$)
  - vector dimension r (almost surely);
  - orthonormal basis $\mathbb{Q}_{n\times r} = \big(\boldsymbol{q}_1,\, \dots,\, \boldsymbol{q}_r\big)$.
- Residual space: $\mathcal{M}\big(\mathbb{X}\big)^\perp$
  - vector dimension $n - r$ (almost surely);
  - orthonormal basis $\mathbb{N}_{n\times r} = \big(\boldsymbol{n}_1,\, \dots,\, \boldsymbol{n}_{n-r}\big)$.
- Hat matrix: $\mathbb{H} = \mathbb{Q}\mathbb{Q}^\top = \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top$.
- Residual projection matrix: $\mathbb{M} = \mathbb{N}\mathbb{N}^\top = \mathbf{I}_n - \mathbb{H}$.
- Fitted values: $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1, \dots, \widehat{Y}_n\big)^\top = \mathbb{H}\boldsymbol{Y}$.
- Residuals: $\boldsymbol{U} = \big(U_1, \dots, U_n\big)^\top = \mathbb{M}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.
- Residual sum of squares: $\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2 = \big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2$.
- Residual degrees of freedom: $\nu_e = n - r$.
- Residual mean square: $\mathsf{MS}_e = \mathsf{SS}_e/(n-r)$.
- Sum of squares: $\mathsf{SS} : \mathbb{R}^k \longrightarrow \mathbb{R}, \quad \mathsf{SS}(\boldsymbol{\beta}) = \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big\|^2$, $\boldsymbol{\beta} \in \mathbb{R}^k$.

Chapter 1

Linear Model

1.1 Regression analysis

Start of Lecture #1 (05/10/2016)

Linear regression[1] is a basic method of so called *regression analysis*[2] which covers a variety of methods to *model* on how distribution of one variable depends on one or more other variables. A principal tool of linear regression is then so called *linear model*[3] which will be the main topic of this lecture.

1.1.1 Data

Basic methods of regression analysis assume that data can be represented by n *independent and identically distributed (i.i.d.)* random vectors $\left(Y_i,\ \boldsymbol{Z}_i^\top\right)^\top$, $i = 1, \ldots, n$, being distributed as a generic random vector $\left(Y,\ \boldsymbol{Z}^\top\right)^\top$. That is,

$$\begin{pmatrix} Y_i \\ \boldsymbol{Z}_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \begin{pmatrix} Y \\ \boldsymbol{Z} \end{pmatrix}, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{Z} = \left(Z_1, \ldots, Z_p\right)^\top$. This will also be a basic assumption used for majority of the lecture.

Terminology *(Response, covariates).*

- Y is called *response*[4] or *dependent variable*[5].
- The components of $\boldsymbol{Z}$ are called *covariates*[6], *explanatory variables*[7], *predictors*[8], or *independent variables*[9].
- The sample space[10] of the covariates will be denoted as $\mathcal{Z}$. That is, $\mathcal{Z} \subseteq \mathbb{R}^p$, and among the other things, $\mathsf{P}(\boldsymbol{Z} \in \mathcal{Z}) = 1$.

[1] *lineární regrese* [2] *regresní analýza* [3] *lineární model* [4] *odezva* [5] *závisle proměnná* [6] *Nepřekládá se. Výraz „kovariáty“ nepoužívat!* [7] *vysvětlující proměnné* [8] *prediktory* [9] *nezávisle proměnné* [10] *výběrový prostor*

Notation and terminology *(Response vector, covariate matrix).*

Further, let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{Z} = \begin{pmatrix} Z_{1,1} & \dots & Z_{1,p} \\ \vdots & \vdots & \vdots \\ Z_{n,1} & \dots & Z_{n,p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_1^\top \\ \vdots \\ \boldsymbol{Z}_n^\top \end{pmatrix} = \big(\boldsymbol{Z}^1,\, \dots,\, \boldsymbol{Z}^p\big).$$

- Vector $\boldsymbol{Y}$ is called the *response vector*[11].
- The $n \times p$ matrix $\mathbb{Z}$ is called the *covariate matrix*[12].
- The vector $\boldsymbol{Z}_i = \big(Z_{i,1},\, \dots,\, Z_{i,p}\big)^\top$ $(i = 1,\, \dots,\, n)$ represents the covariate values for the ith observation.
- The vector $\boldsymbol{Z}^j = \big(Z_{1,j},\, \dots,\, Z_{n,j}\big)^\top$ $(j = 1,\, \dots,\, p)$ represent the values of the jth covariate for the n observations in a sample.

Notation. Letter Y (or y) will always denote a response related quantity. Letters Z (or z) and later also X (or x) will always denote a quantity related to the covariates.

This lecture:

- Response Y is *continuous*.
- Interest in modelling dependence of only the *expected value* (the mean) of Y on the covariates.
- Covariates can be of any type (numeric, categorical).

1.1.2 Probabilistic model for the data

Any statistical analysis is based on specifying a stochastic mechanism which is assumed to generate the data. In our situation, with i.i.d. data $\big(Y_i,\, \boldsymbol{Z}_i^\top\big)^\top$, $i = 1, \dots, n$, the data generating mechanism corresponds to a joint distribution of a generic random vector $\big(Y,\, \boldsymbol{Z}^\top\big)^\top$ which can be given by a joint density

$$f_{Y,\boldsymbol{Z}}(y,\, \boldsymbol{z}), \qquad y \in \mathbb{R},\ \boldsymbol{z} \in \mathcal{Z}$$

(with respect to some σ-finite product measure $\lambda_Y \times \lambda_{\boldsymbol{Z}}$). For the purpose of this lecture, λ_Y will always be a Lebesgue measure on $\big(\mathbb{R},\, \mathcal{B}\big)$.

It is known from basic lectures on probability that any joint density can be decomposed into a product of a conditional and a marginal density as

$$f_{Y,\boldsymbol{Z}}(y,\, \boldsymbol{z}) = f_{Y|\boldsymbol{Z}}\big(y \,\big|\, \boldsymbol{z}\big)\, f_{\boldsymbol{Z}}(\boldsymbol{z}), \qquad y \in \mathbb{R},\ \boldsymbol{z} \in \mathcal{Z}.$$

With the regression analysis, and with the linear regression in particular, the interest lies in revealing certain features of the conditional distribution $Y \,\big|\, \boldsymbol{Z}$ (given by the density $f_{Y|\boldsymbol{Z}}$) while considering the marginal distribution of the covariates $\boldsymbol{Z}$ (given by the density $f_{\boldsymbol{Z}}$) as nuisance. It will be shown during the lecture that a valid statistical inference is possible for suitable characteristics of the conditional distribution of the response given the covariates while leaving the covariates distribution $f_{\boldsymbol{Z}}$ practically unspecified. Moreover, to infer on certain characteristics of the conditional distribution $Y \,\big|\, \boldsymbol{Z}$, e.g., on the conditional mean $\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z}\big)$, even the density $f_{Y|\boldsymbol{Z}}$ might be left practically unspecified for many tasks.

[11] *vektor odezvy* [12] *matice vysvětlujících proměnných*

1.1.3 Regressors

In the reminder of the lecture, we will mainly attempt to model the conditional mean $\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z}\big)$. When doing so, transformations of the original covariates are usually considered. The response (conditional) expectation is then assumed to be a function of the transformed covariates.

In the following, let $\boldsymbol{t} : \mathcal{Z} \longrightarrow \mathcal{X} \subseteq \mathbb{R}^k$ be a measurable function, $\boldsymbol{t} = \big(t_0, \ldots, t_{k-1}\big)^\top$ (for reasons which become clear in a while, we start indexing of the elements of this transformation by zero). Further, let

$$\begin{aligned} \boldsymbol{X} &= \big(X_0, \ldots, X_{k-1}\big)^\top &&= \big(t_0(\boldsymbol{Z}), \ldots, t_{k-1}(\boldsymbol{Z})\big)^\top &&= \boldsymbol{t}(\boldsymbol{Z}), \\ \boldsymbol{X}_i &= \big(X_{i,0}, \ldots, X_{i,k-1}\big)^\top &&= \big(t_0(\boldsymbol{Z}_i), \ldots, t_{k-1}(\boldsymbol{Z}_i)\big)^\top &&= \boldsymbol{t}(\boldsymbol{Z}_i), \qquad i = 1, \ldots, n. \end{aligned}$$

Subsequently, we will assume that

$$\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z}\big) = m\big(\boldsymbol{t}(\boldsymbol{Z})\big) = m(\boldsymbol{X})$$

for some measurable function $m : \mathcal{X} \longrightarrow \mathbb{R}$.

Terminology *(Regressors, regression function).*

- The vectors $\boldsymbol{X}$, $\boldsymbol{X}_i$, $i = 1, \ldots, n$, are called the *regressor vectors*[13] for a particular unit in a sample.
- Function m which relates the response expectation to the regressors is called the *regression function*[14].
- The vector $\boldsymbol{X}^j := \big(X_{1,j}, \ldots, X_{n,j}\big)^\top$ $(j = 0, \ldots, k-1)$ is called the *jth regressor vector.*[15]

All *theoretical* considerations in this lecture will assume that the transformation $\boldsymbol{t}$ which relates the regressor vector $\boldsymbol{X}$ to the covariate vector $\boldsymbol{Z}$ is given and known. If the original data $\big(Y_i, \boldsymbol{Z}_i^\top\big)^\top$, $i = 1, \ldots, n$ are i.i.d. having the distribution of the generic response-covariate vector $\big(Y, \boldsymbol{Z}^\top\big)^\top$, the (transformed) data $\big(Y_i, \boldsymbol{X}_i^\top\big)^\top$, $i = 1, \ldots, n$ are again i.i.d., now having the distribution of the generic response-regressor vector $\big(Y, \boldsymbol{X}^\top\big)^\top$ which is obtained from the distribution of $\big(Y, \boldsymbol{Z}^\top\big)^\top$ by a transformation theorem. The joint density of $\big(Y, \boldsymbol{X}^\top\big)^\top$ can again be decomposed into a product of the conditional and the marginal density as

$$f_{Y,\boldsymbol{X}}(y, \boldsymbol{x}) = f_{Y|\boldsymbol{X}}\big(y \,\big|\, \boldsymbol{x}\big)\, f_{\boldsymbol{X}}(\boldsymbol{x}), \qquad y \in \mathbb{R},\ \boldsymbol{x} \in \mathcal{X}. \tag{1.1}$$

Furthermore, it will overall be assumed that for almost all $\boldsymbol{z} \in \mathcal{Z}$

$$\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z} = \boldsymbol{z}\big) = \mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{t}(\boldsymbol{z})\big). \tag{1.2}$$

Consequently, to model the conditional expectation $\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z}\big)$, it is sufficient to model the conditional expectation $\mathbb{E}\big(Y \,\big|\, \boldsymbol{X}\big)$ using the data $\big(Y_i, \boldsymbol{X}_i^\top\big)^\top$, $i = 1, \ldots, n$ and then to use (1.2) to get $\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z}\big)$. In the reminder of the lecture, if it is not necessary to mention the transformation $\boldsymbol{t}$ which relates the original covariates to the regressors, we will say that the data are directly composed of the response and the regressors.

[13] *vektory regresorů* [14] *regresní funkce* [15] *vektor $j$tého regresoru*

1.2 Linear model: Basics

1.2.1 Linear model with i.i.d. data

Definition 1.1 Linear model with i.i.d. data.
The data $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y,\, \boldsymbol{X}^\top\right)^\top$, $i = 1, \ldots, n$, *satisfy a* linear model *if*

$$\mathbb{E}\left(Y \,\middle|\, \boldsymbol{X}\right) = \boldsymbol{X}^\top \boldsymbol{\beta}, \qquad \mathsf{var}\left(Y \,\middle|\, \boldsymbol{X}\right) = \sigma^2,$$

where $\boldsymbol{\beta} = \left(\beta_0,\, \ldots,\, \beta_{k-1}\right)^\top \in \mathbb{R}^k$ *and* $0 < \sigma^2 < \infty$ *are unknown parameters.*

Terminology (Regression coefficients, residual variance and standard deviation).

- $\boldsymbol{\beta} = \left(\beta_0,\, \ldots,\, \beta_{k-1}\right)^\top$ is called the vector of *regression coefficients*[16] or *regression parameters.*[17]
- σ^2 is called *the residual variance.*[18]
- $\sigma = \sqrt{\sigma^2}$ is called *the residual standard deviation.*[19]

The linear model as specified by Definition 1.1 deals with specifying only the first two moments of the conditional distribution $Y \,\middle|\, \boldsymbol{X}$. For the rest, both the density $f_{Y|\boldsymbol{X}}$ and the density $f_{\boldsymbol{X}}$ from (1.1) can be arbitrary. The regression function of the linear model is

$$m(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta} = \beta_0\, x_0 + \cdots + \beta_{k-1}\, x_{k-1}, \qquad \boldsymbol{x} = \left(x_0,\, \ldots,\, x_{k-1}\right)^\top \in \mathcal{X}.$$

The term "linear" points to the fact that the regression function is *linear* with respect to the regression coefficients vector $\boldsymbol{\beta}$. Note that the regressors $\boldsymbol{X}$ might be (and often are) linked to the original covariates $\boldsymbol{Z}$ (the transformation $\boldsymbol{t}$) in an arbitrary, i.e., also in a non-linear way.

Notation and terminology (Linear model with intercept).

Often, the regressor X_0 is constantly equal to one ($t_0(\boldsymbol{z}) = 1$ for any $\boldsymbol{z} \in \mathcal{Z}$). That is, the regressor vector $\boldsymbol{X}$ is $\boldsymbol{X} = \left(1,\, X_1,\, \ldots,\, X_{k-1}\right)^\top$ and the regression function becomes

$$m(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta} = \beta_0 + \beta_1\, x_1 + \cdots + \beta_{k-1}\, x_{k-1}, \qquad \boldsymbol{x} = \left(1,\, x_1,\, \ldots,\, x_{k-1}\right)^\top \in \mathcal{X}.$$

The related linear model is then called the *linear model with intercept*[20]. The regression coefficient β_0 is called the *intercept term*[21] of the model.

1.2.2 Interpretation of regression coefficients

The regression parameters express influence of the regressors on the response expectation. Let for a chosen $j \in \{0,\, 1,\, \ldots, k-1\}$

$$\boldsymbol{x} = \left(x_0,\, \ldots,\, x_j\, \ldots,\, x_{k-1}\right)^\top \in \mathcal{X}, \quad \text{and} \quad \boldsymbol{x}^{j(+1)} := \left(x_0,\, \ldots,\, x_j + 1\, \ldots,\, x_{k-1}\right)^\top \in \mathcal{X}.$$

[16] *regresní koeficienty* [17] *regresní parametry* [18] *reziduální rozptyl* [19] *reziduální směrodatná odchylka* [20] *lineární model s absolutním členem* [21] *absolutní člen*

We then have

$$\begin{aligned}
&\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}^{j(+1)}\big) \;-\; \mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big) \\
&\qquad = \mathbb{E}\big(Y \,\big|\, X_0 = x_0,\, \ldots,\, X_j = x_j + 1,\, \ldots, X_{k-1} = x_{k-1}\big) \\
&\qquad\qquad\qquad - \;\mathbb{E}\big(Y \,\big|\, X_0 = x_0,\, \ldots,\, X_j = x_j,\, \ldots, X_{k-1} = x_{k-1}\big) \\
&\qquad = \beta_0\, x_0 + \cdots + \beta_j\, (x_j + 1) + \cdots + \beta_{k-1}\, x_{k-1} \\
&\qquad\qquad\qquad - \;\big(\beta_0\, x_0 + \cdots + \beta_j\, x_j + \cdots + \beta_{k-1}\, x_{k-1}\big) \\
&\qquad = \beta_j.
\end{aligned}$$

That is, the regression coefficient β_j expresses a change of the response expectation corresponding to a *unity* change of the jth regressor while keeping the remaining regressors unchanged. Further, let for a fixed $\delta \in \mathbb{R}$

$$\boldsymbol{x}^{j(+\delta)} := \big(x_0,\, \ldots,\, x_j + \delta\, \ldots,\, x_{k-1}\big)^\top \in \mathcal{X},$$

we then have

$$\begin{aligned}
&\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}^{j(+\delta)}\big) \;-\; \mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big) \\
&\qquad = \mathbb{E}\big(Y \,\big|\, X_0 = x_0,\, \ldots,\, X_j = x_j + \delta,\, \ldots, X_{k-1} = x_{k-1}\big) \\
&\qquad\qquad\qquad - \;\mathbb{E}\big(Y \,\big|\, X_0 = x_0,\, \ldots,\, X_j = x_j,\, \ldots, X_{k-1} = x_{k-1}\big) \\
&\qquad = \beta_j\, \delta.
\end{aligned}$$

That is, if for a particular dataset a linear model is assumed, we assume, among the other things the following:

(i) The change of the response expectation corresponding to a constant change δ of the jth regressor does not depend on the value x_j of that regressor which is changed by δ.

(ii) The change of the response expectation corresponding to a constant change δ of the jth regressor does not depend on the values of the remaining regressors.

Terminology *(Effect of the regressor).*

The regression coefficient β_j is also called the *effect* of the jth regressor.

Linear model with intercept

In a model with intercept where X_0 is almost surely equal to one, it does not make sense to consider a change of this regressor by any fixed value. The intercept β_0 has then the following interpretation. If

$$\big(x_0,\, x_1,\, \ldots,\, x_{k-1}\big)^\top = \big(1,\, 0,\, \ldots,\, 0\big)^\top \in \mathcal{X},$$

that is, if the non-intercept regressors may all attain zero values, we have

$$\beta_0 = \mathbb{E}\big(Y \,\big|\, X_1 = 0,\, \ldots,\, X_{k-1} = 0\big).$$

1.2.3 Linear model with general data

Notation and terminology *(Model matrix).*

Let

$$\mathbb{X} = \begin{pmatrix} X_{1,0} & \dots & X_{1,k-1} \\ \vdots & \vdots & \vdots \\ X_{n,0} & \dots & X_{n,k-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \left(\boldsymbol{X}^0, \, \dots, \, \boldsymbol{X}^{k-1}\right).$$

The $n \times k$ matrix $\mathbb{X}$ is called the *model matrix*[22] or the *regression matrix*[23].

In the linear model with intercept, the model matrix becomes

$$\mathbb{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,k-1} \\ \vdots & \vdots & \vdots & \\ 1 & X_{n,1} & \dots & X_{n,k-1} \end{pmatrix} = \left(\mathbf{1}_n, \, \boldsymbol{X}^1, \, \dots, \, \boldsymbol{X}^{k-1}\right).$$

Its first column, the vector $\mathbf{1}_n$, is called the *intercept* column of the model matrix.

The response random vector $\boldsymbol{Y} = \left(Y_1, \, \dots, \, Y_n\right)^\top$, as well as the model matrix $\mathbb{X}$ are random quantities (in case of the model with intercept, the elements of the first column of the model matrix can be viewed as random variables with a Dirac distribution concentrated at the value of one). The joint distribution of the "long" random vector $\left(Y_1, \, \dots, \, Y_n, \, \boldsymbol{X}_1^\top, \, \dots, \, \boldsymbol{X}_n^\top\right)^\top \equiv \left(\boldsymbol{Y}, \, \mathbb{X}\right)$ has in general a density $f_{\boldsymbol{Y}, \mathbb{X}}$ (with respect to some σ-finite product measure $\lambda_{\boldsymbol{Y}} \times \lambda_{\mathbb{X}}$) which can again be decomposed into a product of a conditional and marginal density as

$$f_{\boldsymbol{Y}, \mathbb{X}}(\boldsymbol{y}, \, \mathbf{x}) = f_{\boldsymbol{Y} | \mathbb{X}}\left(\boldsymbol{y} \,\middle|\, \mathbf{x}\right) f_{\mathbb{X}}(\mathbf{x}). \tag{1.3}$$

In case of i.i.d. data, this can be further written as

$$f_{\boldsymbol{Y}, \mathbb{X}}(\boldsymbol{y}, \, \mathbf{x}) = \underbrace{\left\{ \prod_{i=1}^{n} f_{Y | \boldsymbol{X}}\left(y_i \,\middle|\, \boldsymbol{x}_i\right) \right\}}_{f_{\boldsymbol{Y} | \mathbb{X}}\left(\boldsymbol{y} \,\middle|\, \mathbf{x}\right)} \underbrace{\left\{ \prod_{i=1}^{n} f_{\boldsymbol{X}}(\boldsymbol{x}_i) \right\}}_{f_{\mathbb{X}}(\mathbf{x})}. \tag{1.4}$$

The linear model, if assumed for the i.i.d. data, implies statements concerning the (vector) expectation and the covariance matrix of the conditional distribution of the response random vector $\boldsymbol{Y}$ given the model matrix $\mathbb{X}$, i.e., concerning the properties of the first part of the product (1.3).

Lemma 1.1 Conditional mean and covariance matrix of the response vector.

Let the data $\left(Y_i, \, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y, \, \boldsymbol{X}^\top\right)^\top$, $i = 1, \dots, n$ *satisfy a linear model. Then*

$$\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{X}\boldsymbol{\beta}, \qquad \mathsf{var}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \sigma^2 \, \mathbf{I}_n. \tag{1.5}$$

---

[22] *matice modelu* [23] *regresní matice*

Proof. Trivial consequence of the definition of the linear model with the i.i.d. data. ❑

Property (1.5) is implied from assuming $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y,\, \boldsymbol{X}^\top\right)^\top$, $i = 1, \ldots, n$, where $\mathbb{E}\left(Y \,\middle|\, \boldsymbol{X}\right) = \boldsymbol{X}^\top\boldsymbol{\beta}$, $\mathsf{var}\left(Y \,\middle|\, \boldsymbol{X}\right) = \sigma^2$. To derive many results shown later in this lecture, it is sufficient to assume that the full data ($\equiv \left(\boldsymbol{Y}, \mathbb{X}\right)$) satisfy just the weaker condition (1.5) without requesting that the random vectors $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top$, $i = 1, \ldots, n$, which represent the individual observations, are independent or identically distributed. To allow to distinguish when it is necessary to assume the i.i.d. situation and when it is sufficient to assume just the weaker condition (1.5), we shall introduce the following definition.

Definition 1.2 Linear model with general data.
The data $\left(\boldsymbol{Y}, \mathbb{X}\right)$, *satisfy a* linear model *if*

$$\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{X}\boldsymbol{\beta}, \qquad \mathsf{var}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \sigma^2\, \mathbf{I}_n,$$

where $\boldsymbol{\beta} = \left(\beta_0,\, \ldots,\, \beta_{k-1}\right)^\top \in \mathbb{R}^k$ *and* $0 < \sigma^2 < \infty$ *are unknown parameters.*

Notation.

(i) The linear model with i.i.d. data, that is, the assumption $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y,\, \boldsymbol{X}^\top\right)^\top$, $i = 1, \ldots, n$, $\mathbb{E}\left(Y \,\middle|\, \boldsymbol{X}\right) = \boldsymbol{X}^\top\boldsymbol{\beta}$, $\mathsf{var}\left(Y \,\middle|\, \boldsymbol{X}\right) = \sigma^2$ will be briefly stated as

$$\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y,\, \boldsymbol{X}^\top\right)^\top,\ i = 1, \ldots, n, \qquad Y \,\middle|\, \boldsymbol{X} \sim \left(\boldsymbol{X}^\top\boldsymbol{\beta},\, \sigma^2\right).$$

(ii) The linear model with general data, that is, the assumption $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{X}\boldsymbol{\beta}$, $\mathsf{var}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \sigma^2\, \mathbf{I}_n$ will be indicated by

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

Note. If $\boldsymbol{Y} \,\middle|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\right)$ is assumed, we require that in (1.3)

- neither $f_{\boldsymbol{Y}\,|\,\mathbb{X}}$ is of a product type;
- nor $f_{\mathbb{X}}$ is of a product type

as indicated in (1.4).

1.2.4 Rank of the model

The k-dimensional regressor vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ (the $n \times k$ model matrix $\mathbb{X}$) are in general jointly generated by some $(n \cdot k)$-dimensional joint distribution with a density $f_{\mathbb{X}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = f_{\mathbb{X}}(\mathbf{x})$ (with respect to some σ-finite measure $\lambda_{\mathbb{X}}$). In the whole lecture, we will assume $n > k$. Next to it, we will additionally assume in the whole lecture that for a fixed $r \leq k$,

$$\mathsf{P}\left(\mathsf{rank}(\mathbb{X}) = r\right) = 1. \tag{1.6}$$

That is, we will assume that the (column) rank of the model matrix is fixed rather than being random. It should gradually become clear throughout the lecture that this assumption is not really restrictive for most of the practical applications of a linear model.

Convention. In the reminder of the lecture, we will only write $\mathsf{rank}(\mathbb{X}) = r$ which will mean that $\mathsf{P}\big(\mathsf{rank}(\mathbb{X}) = r\big) = 1$ if randomness of the covariates should be taken into account.

Definition 1.3 Full-rank linear model.

A full-rank linear model[24] *is such a linear model where* $r = k$.

Note. In a full-rank linear model, columns of the model matrix $\mathbb{X}$ are *linearly independent* vectors in $\mathbb{R}^n$ (almost surely).

1.2.5 Error terms

Notation and terminology (Error terms).

The random variables

$$\varepsilon_i := Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}, \qquad i = 1, \ldots, n,$$

will be called the *error terms (random errors, disturbances)*[25] of the model. The random vector

$$\boldsymbol{\varepsilon} = \big(\varepsilon_1, \ldots, \varepsilon_n\big)^\top = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$$

will be called the *error term vector*.

Lemma 1.2 Moments of the error terms.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$. *Then*

$$\begin{aligned} \mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) &= \mathbf{0}_n, & \qquad \mathbb{E}\big(\boldsymbol{\varepsilon}\big) &= \mathbf{0}_n, \\ \mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) &= \sigma^2\, \mathbf{I}_n, & \qquad \mathsf{var}\big(\boldsymbol{\varepsilon}\big) &= \sigma^2\, \mathbf{I}_n. \end{aligned}$$

Proof. $\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) - \mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta} - \mathbb{X}\boldsymbol{\beta} = \mathbf{0}_n.$

$\mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) = \mathsf{var}\big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} \,\big|\, \mathbb{X}\big) = \mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \sigma^2\, \mathbf{I}_n.$

$\mathbb{E}\big(\boldsymbol{\varepsilon}\big) = \mathbb{E}\Big\{\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big)\Big\} = \mathbb{E}\big(\mathbf{0}_n\big) = \mathbf{0}_n.$

$\mathsf{var}\big(\boldsymbol{\varepsilon}\big) = \mathbb{E}\Big\{\mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big)\Big\} + \mathsf{var}\Big\{\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big)\Big\} = \mathbb{E}\big(\sigma^2\, \mathbf{I}_n\big) + \mathsf{var}\big(\mathbf{0}_n\big) = \sigma^2\, \mathbf{I}_n.$

❑

---

[24] *lineární model o plné hodnosti* [25] *chybové členy, náhodné chyby*

Note. If $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y,\, \boldsymbol{X}^\top\right)^\top$, $i = 1, \ldots, n$, then indeed

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \varepsilon,\ i = 1,\, \ldots,\, n, \quad \varepsilon \sim \left(0,\, \sigma^2\right).$$

1.2.6 Distributional assumptions

To derive some of the results, it is necessary not only to assume a certain form of the conditional expectations of the response given the regressors but to specify more closely the whole conditional distribution of the response given the regressors. For example, with i.i.d. data $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top \overset{\text{i.i.d.}}{\sim} \left(Y,\, \boldsymbol{X}^\top\right)^\top$, $i = 1, \ldots, n$, many results can be derived (see Chapter 3) if it is assumed

$$Y \,\big|\, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{X}^\top \boldsymbol{\beta},\, \sigma^2\right).$$

1.2.7 Fixed or random covariates

In certain application areas (e.g., designed experiments), the covariates (and regressors) can all (or some of them) be fixed rather than random variables. This means that the covariate values are determined/set by the analyst rather than being observed on (randomly selected) subjects. For majority of the theory presented throughout this course, it does not really matter whether the covariates are considered as random or as fixed quantities. The proofs (majority that appear in this lecture) very often work with conditional statements given the covariate/regressor values and hence proceed in exactly the same way in both situations. Nevertheless, especially when dealing with asymptotic properties of the estimators used in the context of a linear model (see Chapter 13), care must be taken on whether the covariates are considered as random or as fixed.

1.2.8 Limitations of a linear model

> "*Essentially, all models are wrong, but some are useful. The practical question is how wrong do they have to be to not be useful.*"
>
> George E. P. Box (1919 – 2013)

Linear model is indeed only one possibility (out of infinitely many) on how to model dependence of the response on the covariates. The linear model as defined by Definition 1.1 is (possibly seriously) wrong if, for example,

- The expected value $\mathbb{E}\left(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\right)$, $\boldsymbol{x} \in \mathcal{X}$, cannot be expressed as a linear function of $\boldsymbol{x}$.
  $\Rightarrow$ *Incorrect regression function.*
- The conditional variance $\mathsf{var}\left(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\right)$, $\boldsymbol{x} \in \mathcal{X}$, is not constant. It may depend on $\boldsymbol{x}$ as well, it may depend on other factors.
  $\Rightarrow$ *Heteroscedasticity.*
- Response random variables are not conditionally uncorrelated/independent (the error terms are not uncorrelated/independent). This is often the case if response is measured repeatedly (e.g., over time) on n subjects included in the study.

Additionally, the linear model deals with modelling of only the first two (conditional) moments of the response. In many application areas, other characteristics of the conditional distribution $Y \,\big|\, \boldsymbol{X}$ are of (primary) interest.

End of Lecture #1 (05/10/2016)

Chapter 2

Least Squares Estimation

Start of Lecture #2 (05/10/2016)

In this chapter, we shall consider a set of n random vectors $\left(Y_i,\ \boldsymbol{X}_i^\top\right)^\top$, $\boldsymbol{X}_i = \left(X_{i,0}, \ldots, X_{i,k-1}\right)^\top$, $i = 1, \ldots, n$, which are not necessarily i.i.d. but satisfy a linear model. That is,

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\right), \qquad \mathsf{rank}\left(\mathbb{X}_{n\times k}\right) = r \leq k < n, \tag{2.1}$$

where $\boldsymbol{Y} = \left(Y_1, \ldots, Y_n\right)^\top$, $\mathbb{X}$ is a matrix with vectors $\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top$ in its rows and $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{k-1}\right) \in \mathbb{R}^k$ and $\sigma^2 > 0$ are unknown parameters. In this chapter, we introduce a *method of least squares*[1] to estimate the unknown parameters of the linear model (2.1). All results in this chapter will be derived from the assumption (2.1), i.e., without assuming i.i.d. data or even normally distributed response.

[1] *metoda nejmenších čtverců*

2.1 Regression and residual space, projections

2.1.1 Regression and residual space

Notation *(Linear span of columns of the model matrix and its orthogonal complement).*

For given dataset and a linear model, the model matrix $\mathbb{X}$ is a real $n \times k$ matrix. Let $\boldsymbol{x}^0, \ldots, \boldsymbol{x}^{k-1} \in \mathbb{R}^n$ denote its columns, i.e.,

$$\mathbb{X} = \big(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^{k-1}\big).$$

- The *linear span*[2] of columns of $\mathbb{X}$, i.e., a vector space generated by vectors $\boldsymbol{x}^0, \ldots, \boldsymbol{x}^{k-1}$ will be denoted as $\mathcal{M}\big(\mathbb{X}\big)$, that is,

$$\mathcal{M}\big(\mathbb{X}\big) = \big\{\boldsymbol{v} : \boldsymbol{v} = \sum_{j=0}^{k-1} \beta_j \boldsymbol{x}^j,\ \boldsymbol{\beta} = \big(\beta_0, \ldots, \beta_{k-1}\big)^\top \in \mathbb{R}^k\big\}.$$

- The orthogonal complement to $\mathcal{M}\big(\mathbb{X}\big)$ will be denoted as $\mathcal{M}\big(\mathbb{X}\big)^\perp$, that is,

$$\mathcal{M}\big(\mathbb{X}\big)^\perp = \big\{\boldsymbol{u} : \boldsymbol{u} \in \mathbb{R}^n,\ \boldsymbol{v}^\top \boldsymbol{u} = 0 \text{ for all } \boldsymbol{v} \in \mathcal{M}\big(\mathbb{X}\big)\big\}.$$

Note. We know from linear algebra lectures that the linear span of column of $\mathbb{X}$, $\mathcal{M}\big(\mathbb{X}\big)$, is a vector subspace of dimension r of the n-dimensional Euclidean space $\mathbb{R}^n$. Similarly, $\mathcal{M}\big(\mathbb{X}\big)^\perp$ is a vector subspace of dimension $n - r$ of the n-dimensional Euclidean space $\mathbb{R}^n$. We have

$$\mathcal{M}\big(\mathbb{X}\big) \cup \mathcal{M}\big(\mathbb{X}\big)^\perp = \mathbb{R}^n, \qquad \mathcal{M}\big(\mathbb{X}\big) \cap \mathcal{M}\big(\mathbb{X}\big)^\perp = \{\mathbf{0}_n\},$$
$$\text{for any } \boldsymbol{v} \in \mathcal{M}\big(\mathbb{X}\big),\ \boldsymbol{u} \in \mathcal{M}\big(\mathbb{X}\big)^\perp \quad \boldsymbol{v}^\top \boldsymbol{u} = 0.$$

Definition 2.1 Regression and residual space of a linear model.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}) = r$. *The* regression space[3] *of the model is a vector space* $\mathcal{M}\big(\mathbb{X}\big)$. *The* residual space[4] *of the model is the orthogonal complement of the regression space, i.e., a vector space* $\mathcal{M}\big(\mathbb{X}\big)^\perp$.

Notation *(Orthonormal vector bases of the regression and residual space).*

Let $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r$ be (any) orthonormal vector basis of the regression space $\mathcal{M}\big(\mathbb{X}\big)$ and let $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_{n-r}$ be (any) orthonormal vector basis of the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$. That is, $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r, \boldsymbol{n}_1, \ldots, \boldsymbol{n}_{n-r}$ is an orthonormal vector basis of the n-dimensinal Euclidean space $\mathbb{R}^n$. We will denote

- $\mathbb{Q}_{n\times r} = \big(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r\big)$.
- $\mathbb{N}_{n\times(n-r)} = \big(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_{n-r}\big)$.

[2] *lineární obal* [3] *regresní prostor* [4] *reziduální prostor*

- $\mathbb{P}_{n\times n} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r, \boldsymbol{n}_1, \ldots, \boldsymbol{n}_{n-r}) = (\mathbb{Q}, \mathbb{N})$.

Notes. It follows from the linear algebra lectures.

- Properties of the columns of the $\mathbb{Q}$ matrix:
  - $\boldsymbol{q}_j^\top \boldsymbol{q}_j = 1, \quad j = 1, \ldots, r;$
  - $\boldsymbol{q}_j^\top \boldsymbol{q}_l = 0, \quad j, l = 1, \ldots, r,\ j \neq l.$
- Properties of the columns of the $\mathbb{N}$ matrix:
  - $\boldsymbol{n}_j^\top \boldsymbol{n}_j = 1, \quad j = 1, \ldots, n-r;$
  - $\boldsymbol{n}_j^\top \boldsymbol{n}_l = 0, \quad j, l = 1, \ldots, n-r,\ j \neq l.$
- Mutual properties of the columns of the $\mathbb{Q}$ and $\mathbb{N}$ matrix:
  - $\boldsymbol{q}_j^\top \boldsymbol{n}_l = \boldsymbol{n}_l^\top \boldsymbol{q}_j = 0, \quad j = 1, \ldots, r,\ l = 1, \ldots, n-r.$
- Above properties written in a matrix form:

$$\mathbb{Q}^\top \mathbb{Q} = \mathbf{I}_r, \quad \mathbb{N}^\top \mathbb{N} = \mathbf{I}_{n-r},$$
$$\mathbb{Q}^\top \mathbb{N} = \mathbf{0}_{r\times(n-r)}, \quad \mathbb{N}^\top \mathbb{Q} = \mathbf{0}_{(n-r)\times r},$$
$$\mathbb{P}^\top \mathbb{P} = \mathbf{I}_n. \tag{2.2}$$

- It follows from (2.2) that $\mathbb{P}^\top$ is inverse to $\mathbb{P}$ and hence

$$\mathbf{I}_n = \mathbb{P}\mathbb{P}^\top = \begin{pmatrix}\mathbb{Q}, & \mathbb{N}\end{pmatrix} \begin{pmatrix}\mathbb{Q}^\top \\ \mathbb{N}^\top\end{pmatrix} = \mathbb{Q}\mathbb{Q}^\top + \mathbb{N}\mathbb{N}^\top.$$

- It is also useful to remind

$$\mathcal{M}(\mathbb{X}) = \mathcal{M}(\mathbb{Q}), \qquad \mathcal{M}(\mathbb{X})^\perp = \mathcal{M}(\mathbb{N}), \qquad \mathbb{R}^n = \mathcal{M}(\mathbb{P}).$$

Notation. In the following, let

$$\mathbb{H} = \mathbb{Q}\mathbb{Q}^\top, \qquad \mathbb{M} = \mathbb{N}\mathbb{N}^\top.$$

Note. Matrices $\mathbb{H}$ and $\mathbb{M}$ are symmetric and idempotent:

$$\mathbb{H}^\top = \left(\mathbb{Q}\mathbb{Q}^\top\right)^\top = \mathbb{Q}\mathbb{Q}^\top = \mathbb{H}, \qquad \mathbb{H}\mathbb{H} = \mathbb{Q}\mathbb{Q}^\top\mathbb{Q}\mathbb{Q}^\top = \mathbb{Q}\mathbf{I}_r\mathbb{Q}^\top = \mathbb{Q}\mathbb{Q}^\top = \mathbb{H},$$
$$\mathbb{M}^\top = \left(\mathbb{N}\mathbb{N}^\top\right)^\top = \mathbb{N}\mathbb{N}^\top = \mathbb{M}, \qquad \mathbb{M}\mathbb{M} = \mathbb{N}\mathbb{N}^\top\mathbb{N}\mathbb{N}^\top = \mathbb{N}\mathbf{I}_{n-r}\mathbb{N}^\top = \mathbb{N}\mathbb{N}^\top = \mathbb{M}.$$

2.1.2 Projections

Let $\boldsymbol{y} \in \mathbb{R}^n$. We can then write (while using identity in Expression 2.2)

$$\boldsymbol{y} = \mathbf{I}_n\boldsymbol{y} = \left(\mathbb{Q}\mathbb{Q}^\top + \mathbb{N}\mathbb{N}^\top\right)\boldsymbol{y} = (\mathbb{H} + \mathbb{M})\boldsymbol{y} = \mathbb{H}\boldsymbol{y} + \mathbb{M}\boldsymbol{y}.$$

We have

- $\widehat{\boldsymbol{y}} := \mathbb{H}\boldsymbol{y} = \mathbb{Q}\left(\mathbb{Q}^\top \boldsymbol{y}\right) \in \mathcal{M}(\mathbb{X})$.
- $\boldsymbol{u} := \mathbb{M}\boldsymbol{y} = \mathbb{N}\left(\mathbb{N}^\top \boldsymbol{y}\right) \in \mathcal{M}(\mathbb{X})^\perp$.
- $\boldsymbol{y}^\top \boldsymbol{u} = \boldsymbol{y}^\top \mathbb{Q}\left(\mathbb{Q}^\top \mathbb{N}\right) \mathbb{N}^\top \boldsymbol{y} = \boldsymbol{y}^\top \mathbb{Q}\, \mathbf{0}_{r\times(n-r)}\, \mathbb{N}^\top \boldsymbol{y} = 0$.

That is, we have decomposition of any $\boldsymbol{y} \in \mathbb{R}^n$ into

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{u}, \qquad \widehat{\boldsymbol{y}} \in \mathcal{M}(\mathbb{X}),\ \boldsymbol{u} \in \mathcal{M}(\mathbb{X})^\perp, \quad \widehat{\boldsymbol{y}} \perp \boldsymbol{u}.$$

In other words, $\widehat{\boldsymbol{y}}$ and $\boldsymbol{u}$ are *projections* of $\boldsymbol{y}$ into $\mathcal{M}(\mathbb{X})$ and $\mathcal{M}(\mathbb{X})^\perp$, respectively, and $\mathbb{H}$ and $\mathbb{M}$ are corresponding *projection matrices*.

Notes. It follows from the linear algebra lectures.

- Decomposition $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{u}$ is unique.
- Projection matrices $\mathbb{H}$, $\mathbb{M}$ are unique. That is $\mathbb{H} = \mathbb{Q}\mathbb{Q}^\top$ does not depend on a choice of the orthonormal vector basis of $\mathcal{M}(\mathbb{X})$ included in the $\mathbb{Q}$ matrix and $\mathbb{M} = \mathbb{N}\mathbb{N}^\top$ does not depend on a choice of the orthonormal vector basis of $\mathcal{M}(\mathbb{X})^\perp$ included in the $\mathbb{N}$ matrix.
- Vector $\widehat{\boldsymbol{y}} = \left(\widehat{y}_1, \ldots, \widehat{y}_n\right)^\top$ is the closest point (in the Euclidean metric) in the regression space $\mathcal{M}(\mathbb{X})$ to a given vector $\boldsymbol{y} = \left(y_1, \ldots, y_n\right)^\top$, that is,

$$\forall\, \widetilde{\boldsymbol{y}} = \left(\widetilde{y}_1, \ldots, \widetilde{y}_n\right)^\top \in \mathcal{M}(\mathbb{X}) \qquad \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \le \sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2 = \|\boldsymbol{y} - \widetilde{\boldsymbol{y}}\|^2.$$

Definition 2.2 Hat matrix, residual projection matrix.

Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\right)$, where $\mathbb{Q}$ and $\mathbb{N}$ are the orthonormal bases of the regression and the residual space, respectively.

1. *The* hat matrix[5] *of the model is the matrix $\mathbb{Q}\,\mathbb{Q}^\top$ which is denoted as $\mathbb{H}$.*
2. *The* residual projection matrix[6] *of the model is the matrix $\mathbb{N}\,\mathbb{N}^\top$ which is denoted as $\mathbb{M}$.*

Lemma 2.1 Expressions of the projection matrices using the model matrix.

The hat matrix $\mathbb{H}$ and the residual projection matrix $\mathbb{M}$ can be expressed as

$$\begin{aligned} \mathbb{H} &= \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top, \\ \mathbb{M} &= \mathbf{I}_n - \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top. \end{aligned}$$

Proof.

- Five matrices rule (Theorem A.2):

$$\begin{aligned} \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top\mathbb{X} &= \mathbb{X}, \\ \left\{\mathbf{I}_n - \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top\right\}\mathbb{X} &= \mathbf{0}_{n\times k}. \end{aligned}$$

[5] *regresní projekční matice, lze však užívat též výrazu „hat matice“* [6] *reziduální projekční matice*

- Let $\widetilde{\mathbb{H}} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top$,
  $\widetilde{\mathbb{M}} = \mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top = \mathbf{I}_n - \widetilde{\mathbb{H}}$.

- We have $\widetilde{\mathbb{M}}\mathbb{X} = \mathbf{0}_{n\times k}$, both $\widetilde{\mathbb{H}}$ and $\widetilde{\mathbb{M}}$ are symmetric.

- We now have:
$$\boldsymbol{y} = \mathbf{I}_n\,\boldsymbol{y} = \big(\widetilde{\mathbb{H}} + \mathbf{I}_n - \widetilde{\mathbb{H}}\big)\boldsymbol{y} = \widetilde{\mathbb{H}}\boldsymbol{y} + \widetilde{\mathbb{M}}\boldsymbol{y}.$$

- Clearly, $\widetilde{\mathbb{H}}\boldsymbol{y} = \mathbb{X}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\Big\}\boldsymbol{y} \in \mathcal{M}\big(\mathbb{X}\big)$.

- For any $\boldsymbol{z} = \mathbb{X}\boldsymbol{b} \in \mathcal{M}\big(\mathbb{X}\big)$: $\boldsymbol{z}^\top\widetilde{\mathbb{M}}\boldsymbol{y} = \boldsymbol{b}^\top\mathbb{X}^\top\widetilde{\mathbb{M}}\boldsymbol{y} = \boldsymbol{y}^\top\underbrace{\widetilde{\mathbb{M}}\mathbb{X}}_{\mathbf{0}_n}\boldsymbol{b} = 0$. Hence $\widetilde{\mathbb{M}}\boldsymbol{y} \in \mathcal{M}\big(\mathbb{X}\big)^\perp$.

$\Rightarrow$ Uniqueness of projections and projection matrices

$$\mathbb{H} = \widetilde{\mathbb{H}} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top,$$
$$\mathbb{M} = \widetilde{\mathbb{M}} = \mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top.$$

❑

Notes.

- Expression $\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top$ does not depend on a choice of the pseudoinverse matrix $\big(\mathbb{X}^\top\mathbb{X}\big)^-$.
- If $r = \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$ then

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

2.2 Fitted values, residuals, Gauss–Markov theorem

Before starting to deal with estimation of the principal parameters of the linear model, which are the regression coefficients $\boldsymbol{\beta}$, we will deal, for a given model matrix $\mathbb{X}$ based on the observed data, with estimation of the full (conditional) mean $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}$ of the response vector $\boldsymbol{Y}$ and its (conditional) covariance matrix $\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbf{I}_n$ for which it is sufficient to estimate the residual variance σ^2.

Notation. We denote

$$\boldsymbol{\mu} := \mathbb{X}\boldsymbol{\beta} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big).$$

By saying that we are now interested in estimation of the full (conditional) expectation $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)$ we mean that we want to estimate the parameter vector $\boldsymbol{\mu}$ on its own without necessity to know its decomposition into $\mathbb{X}\boldsymbol{\beta}$.

Definition 2.3 Fitted values, residuals, residual sum of squares.
Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$.

1. *The* fitted values[7] *or the vector of fitted values of the model is a vector* $\mathbb{H}\boldsymbol{Y}$ *which will be denoted as* $\widehat{\boldsymbol{Y}}$. *That is,*
$$\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1, \ldots, \widehat{Y}_n\big)^\top = \mathbb{H}\boldsymbol{Y}.$$

2. *The* residuals[8] *or the vector of residuals of the model is a vector* $\mathbb{M}\boldsymbol{Y}$ *which will be denoted as* $\boldsymbol{U}$. *That is,*
$$\boldsymbol{U} = \big(U_1, \ldots, U_n\big)^\top = \mathbb{M}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}.$$

3. *The* residual sum of squares[9] *of the model is a quantity* $\big\|\boldsymbol{U}\big\|^2$ *which will be denoted as* SS_e. *That is,*

$$\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2 = \boldsymbol{U}^\top\boldsymbol{U} = \sum_{i=1}^n U_i^2$$

$$= \sum_{i=1}^n \big(Y_i - \widehat{Y}_i\big)^2 = \big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big)^\top\big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big) = \big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2.$$

Notes.

- The fitted values $\widehat{\boldsymbol{Y}}$ and the residuals $\boldsymbol{U}$ are projections of the response vector $\boldsymbol{Y}$ into the regression space $\mathcal{M}\big(\mathbb{X}\big)$ and the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$, respectively.
- Using different quantities and expressions introduced in Section 2.1, we can write

$$\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \mathbb{Q}\,\mathbb{Q}^\top\boldsymbol{Y} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\boldsymbol{Y},$$

$$\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = \mathbb{N}\,\mathbb{N}^\top\boldsymbol{Y} = \Big\{\mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\Big\}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}.$$

[7] *vyrovnané hodnoty* [8] *rezidua* [9] *reziduální součet čtverců*

- It follows from the projection properties that the vector $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1, \dots, \widehat{Y}_n\big)^\top$ is the nearest point of the regression space $\mathcal{M}\big(\mathbb{X}\big)$ to the response vector $\boldsymbol{Y} = \big(Y_1, \dots, Y_n\big)^\top$, that is,

$$\forall\, \widetilde{\boldsymbol{Y}} = \big(\widetilde{Y}_1, \dots, \widetilde{Y}_n\big)^\top \in \mathcal{M}(\mathbb{X})$$

$$\big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \le \sum_{i=1}^{n} (Y_i - \widetilde{Y}_i)^2 = \big\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}\big\|^2. \quad (2.3)$$

- The Gauss–Markov theorem introduced below shows that $\widehat{\boldsymbol{Y}}$ is a suitable estimator of $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$. Owing to (2.3), it is also called the *least squares estimator (LSE)*.[10] The method of estimation is then called the *method of least squares*[11] or the method of *ordinary least squares (OLS)*.

Theorem 2.2 Gauss–Markov.

Assume a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I}_n\big)$. *Then the vector of fitted values* $\widehat{\boldsymbol{Y}}$ *is, conditionally given* $\mathbb{X}$, *the* best linear unbiased estimator (BLUE)[12] *of a vector parameter* $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)$. *Further,*

$$\mathsf{var}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X}\big) = \sigma^2\, \mathbb{H} = \sigma^2\, \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top.$$

Proof.

Linearity means that $\widehat{\boldsymbol{Y}}$ is a *linear* function of the response vector $\boldsymbol{Y}$ which is clear from the expression $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$.

Unbiasedness. Let us calculate $\mathbb{E}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X}\big)$.

$$\mathbb{E}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\mathbb{H}\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{H}\,\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{H}\,\mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu}.$$

The pre-last equality holds due to the fact that $\mathbb{H}\mathbb{X}$ is a projection of each column of $\mathbb{X}$ into $\mathcal{M}\big(\mathbb{X}\big)$ which is generated by those columns. That is $\mathbb{H}\mathbb{X} = \mathbb{X}$.

Optimality. Let $\widetilde{\boldsymbol{Y}} = \boldsymbol{a} + \mathbb{B}\boldsymbol{Y}$ be some other linear unbiased estimator of $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$.

- That is,

$$\begin{aligned} &\forall\, \boldsymbol{\beta} \in \mathbb{R}^k \quad \mathbb{E}\big(\widetilde{\boldsymbol{Y}} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}, \\ &\forall\, \boldsymbol{\beta} \in \mathbb{R}^k \quad \boldsymbol{a} + \mathbb{B}\,\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}, \\ &\forall\, \boldsymbol{\beta} \in \mathbb{R}^k \quad \boldsymbol{a} + \mathbb{B}\mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}. \end{aligned}$$

  It follows from here, by using above equality with $\boldsymbol{\beta} = \mathbf{0}_k$, that $\boldsymbol{a} = \mathbf{0}_n$.
- That is, from unbiasedness, we have that $\forall\, \boldsymbol{\beta} \in \mathbb{R}^k\ \mathbb{B}\,\mathbb{X}\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}$. Take now $\boldsymbol{\beta} = \big(0, \dots, 1, \dots, 0\big)^\top$ while changing a position of one. From here, it follows that $\mathbb{B}\mathbb{X} = \mathbb{X}$.
- We now have:

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{a} + \mathbb{B}\boldsymbol{Y} \text{ unbiased estimator of } \boldsymbol{\mu} \quad \Longrightarrow \quad \boldsymbol{a} = \mathbf{0}_k \ \&\ \mathbb{B}\mathbb{X} = \mathbb{X}.$$

  Trivially (but we will not need it here), also the opposite implication holds (if $\widetilde{\boldsymbol{Y}} = \mathbb{B}\boldsymbol{Y}$ with $\mathbb{B}\mathbb{X} = \mathbb{X}$ then $\widetilde{\boldsymbol{Y}}$ is the unbiased estimator of $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$). In other words,

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{a} + \mathbb{B}\boldsymbol{Y} \text{ is unbiased estimator of } \boldsymbol{\mu} \quad \Longleftrightarrow \quad \boldsymbol{a} = \mathbf{0}_n \ \&\ \mathbb{B}\mathbb{X} = \mathbb{X}.$$

[10] *odhad metodou nejmenších čtverců* [11] *metoda nejmenších čtverců (MNČ)* [12] *nejlepší lineární nestranný odhad*

- Let us now explore what can be concluded from the equality $\mathbb{B}\mathbb{X} = \mathbb{X}$.

$$\begin{aligned} \mathbb{B}\mathbb{X} &= \mathbb{X}, && \Big| \quad \cdot \left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top \\ \mathbb{B}\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top &= \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top, \\ \mathbb{B}\mathbb{H} &= \mathbb{H}, && (2.4) \\ \mathbb{H}^\top\mathbb{B}^\top &= \mathbb{H}^\top, \\ \mathbb{H}\mathbb{B}^\top &= \mathbb{H}. && (2.5) \end{aligned}$$

- Let us calculate $\mathsf{var}\left(\widehat{\boldsymbol{Y}} \,\middle|\, \mathbb{X}\right)$:

$$\begin{aligned} \mathsf{var}\left(\widehat{\boldsymbol{Y}} \,\middle|\, \mathbb{X}\right) &= \mathsf{var}\left(\mathbb{H}\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{H}\,\mathsf{var}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right)\mathbb{H}^\top = \mathbb{H}\left(\sigma^2\mathbf{I}_n\right)\mathbb{H}^\top \\ &= \sigma^2\,\mathbb{H}\mathbb{H}\mathbb{H}^\top = \sigma^2\,\mathbb{H} = \sigma^2\,\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top. \end{aligned}$$

- Analogously, we calculate $\mathsf{var}\left(\widetilde{\boldsymbol{Y}} \,\middle|\, \mathbb{X}\right)$ for $\widetilde{\boldsymbol{Y}} = \mathbb{B}\boldsymbol{Y}$, where $\mathbb{B}\mathbb{X} = \mathbb{X}$:

$$\begin{aligned} \mathsf{var}\left(\widetilde{\boldsymbol{Y}} \,\middle|\, \mathbb{X}\right) &= \mathsf{var}\left(\mathbb{B}\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{B}\,\mathsf{var}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right)\mathbb{B}^\top = \mathbb{B}\left(\sigma^2\mathbf{I}_n\right)\mathbb{B}^\top \\ &= \sigma^2\,\mathbb{B}\mathbb{B}^\top = \sigma^2\left(\mathbb{H}+\mathbb{B}-\mathbb{H}\right)\left(\mathbb{H}+\mathbb{B}-\mathbb{H}\right)^\top \\ &= \sigma^2\left\{\underbrace{\mathbb{H}\mathbb{H}\mathbb{H}^\top}_{\mathbb{H}} + \underbrace{\mathbb{H}(\mathbb{B}-\mathbb{H})^\top}_{\mathbf{0}_n} + \underbrace{(\mathbb{B}-\mathbb{H})\mathbb{H}^\top}_{\mathbf{0}_n} + (\mathbb{B}-\mathbb{H})\left(\mathbb{B}-\mathbb{H}\right)^\top\right\} \\ &= \sigma^2\,\mathbb{H} + \sigma^2\left(\mathbb{B}-\mathbb{H}\right)\left(\mathbb{B}-\mathbb{H}\right)^\top, \end{aligned}$$

  where $\mathbb{H}(\mathbb{B}-\mathbb{H})^\top = (\mathbb{B}-\mathbb{H})\mathbb{H}^\top = \mathbf{0}_n$ follow from (2.4) and (2.5) and from the fact that $\mathbb{H}$ is symmetric and idempotent.
- Hence finally,

$$\mathsf{var}\left(\widetilde{\boldsymbol{Y}} \,\middle|\, \mathbb{X}\right) - \mathsf{var}\left(\widehat{\boldsymbol{Y}} \,\middle|\, \mathbb{X}\right) = \sigma^2\left(\mathbb{B}-\mathbb{H}\right)\left(\mathbb{B}-\mathbb{H}\right)^\top,$$

  which is a positive semidefinite matrix. That is, the estimator $\widehat{\boldsymbol{Y}}$ is not worse than the estimator $\widetilde{\boldsymbol{Y}}$.

❑

Note. It follows from the Gauss–Markov theorem that

$$\widehat{\boldsymbol{Y}} \,\middle|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbb{H}\right).$$

Historical remarks

- The method of least squares was used in astronomy and geodesy already at the beginning of the 19th century.
- 1805: First documented publication of least squares.

  Adrien-Marie Legendre. Appendix "*Sur le méthode des moindres quarrés*" ("*On the method of least squares*") in the book *Nouvelles Méthodes Pour la Détermination des Orbites des Comètes* (*New Methods for the Determination of the Orbits of the Comets*).

- 1809: Another (supposedly independent) publication of least squares.
  Carl Friedrich Gauss. In Volume 2 of the book *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium* (*The Theory of the Motion of Heavenly Bodies Moving Around the Sun in Conic Sections*).
  - C. F. Gauss claimed he had been using the method of least squares since 1795 (which is probably true).
- The Gauss–Markov theorem was first proved by C. F. Gauss in 1821 – 1823.
- In 1912, A. A. Markov provided another version of the proof.
- In 1934, J. Neyman described the Markov's proof as being "elegant" and stated that Markov's contribution (written in Russian) had been overlooked in the West.
  $\Rightarrow$ The name Gauss–Markov theorem.

Theorem 2.3 Basic properties of the residuals and the residual sum of squares.
Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \le k < n$. *The following then holds:*

(i) $\boldsymbol{U} = \mathbb{M}\boldsymbol{\varepsilon}$, *where* $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$.

(ii) $\mathsf{SS}_e = \boldsymbol{Y}^\top\mathbb{M}\boldsymbol{Y} = \boldsymbol{\varepsilon}^\top\mathbb{M}\boldsymbol{\varepsilon}$.

(iii) $\mathbb{E}\big(\boldsymbol{U} \,\big|\, \mathbb{X}\big) = \mathbf{0}_n,\ \ \mathsf{var}\big(\boldsymbol{U} \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbb{M}$.

(iv) $\mathbb{E}\big(\mathsf{SS}_e \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\mathsf{SS}_e\big) = (n-r)\,\sigma^2$.

Proof.

(i) $$\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = \mathbb{M}\,(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \underbrace{\mathbb{M}\mathbb{X}}_{\mathbf{0}_n}\boldsymbol{\beta} + \mathbb{M}\boldsymbol{\varepsilon} = \mathbb{M}\boldsymbol{\varepsilon}.$$

(ii) $$\begin{aligned}\mathsf{SS}_e &= \boldsymbol{U}^\top\boldsymbol{U} = (\mathbb{M}\boldsymbol{Y})^\top\,\mathbb{M}\boldsymbol{Y}\\ &= \boldsymbol{Y}^\top\underbrace{\mathbb{M}^\top\mathbb{M}}_{\mathbb{M}}\boldsymbol{Y} = \boldsymbol{Y}^\top\mathbb{M}\boldsymbol{Y}\\ &\overset{(i)}{=} \boldsymbol{\varepsilon}^\top\mathbb{M}^\top\mathbb{M}\,\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbb{M}\,\boldsymbol{\varepsilon}.\end{aligned}$$

(iii) $$\begin{aligned}\mathbb{E}\big(\boldsymbol{U} \,\big|\, \mathbb{X}\big) &= \mathbb{E}\big(\mathbb{M}\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \underbrace{\mathbb{M}\mathbb{X}}_{\mathbf{0}_n}\boldsymbol{\beta} = \mathbf{0}_n.\\ \mathsf{var}\big(\boldsymbol{U} \,\big|\, \mathbb{X}\big) &= \mathsf{var}\big(\mathbb{M}\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{M}\,\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)\,\mathbb{M}^\top = \mathbb{M}\,\big(\sigma^2\mathbf{I}_n\big)\,\mathbb{M}^\top = \sigma^2\,\mathbb{M}\mathbb{M}^\top = \sigma^2\,\mathbb{M}.\end{aligned}$$

(iv) $$\begin{aligned}\mathbb{E}\big(\mathsf{SS}_e \,\big|\, \mathbb{X}\big) &= \mathbb{E}\big(\boldsymbol{\varepsilon}^\top\mathbb{M}\,\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) = \mathbb{E}\Big\{\mathsf{tr}\big(\boldsymbol{\varepsilon}^\top\mathbb{M}\,\boldsymbol{\varepsilon}\big) \,\Big|\, \mathbb{X}\Big\} = \mathbb{E}\Big\{\mathsf{tr}\big(\mathbb{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big) \,\Big|\, \mathbb{X}\Big\}\\ &= \mathsf{tr}\Big\{\mathbb{E}\big(\mathbb{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big) \,\Big|\, \mathbb{X}\Big\} = \mathsf{tr}\Big\{\mathbb{M}\,\underbrace{\mathbb{E}\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \,\big|\, \mathbb{X}\big)}_{\mathsf{var}(\boldsymbol{\varepsilon}|\mathbb{X})}\Big\} = \mathsf{tr}\big(\mathbb{M}\,\sigma^2\,\mathbf{I}_n\big) = \mathsf{tr}\big(\sigma^2\,\mathbb{M}\big)\\ &= \sigma^2\,\mathsf{tr}(\mathbb{M}) = \sigma^2\,\mathsf{tr}\big(\mathbb{N}\mathbb{N}^\top\big) = \sigma^2\,\mathsf{tr}\big(\mathbb{N}^\top\mathbb{N}\big) = \sigma^2\,\mathsf{tr}\big(\mathbf{I}_{n-r}\big) = \sigma^2\,(n-r),\end{aligned}$$

where (to remind) $\mathbb{N}$ denotes an $n \times (n-r)$ matrix whose columns form an orthonormal vector basis of a residual space $\mathcal{M}(\mathbb{X})^{\perp}$.

Finally, $\mathbb{E}(\mathsf{SS}_e) = \mathbb{E}\Big\{\mathbb{E}(\mathsf{SS}_e \,|\, \mathbb{X})\Big\} = \mathbb{E}\big\{(n-r)\,\sigma^2\big\} = (n-r)\,\sigma^2$.

❑

Notes.

- Point (i) of Theorem 2.3 says that the residuals can be obtained not only by projecting the response vector $\boldsymbol{Y}$ into $\mathcal{M}(\mathbb{X})^{\perp}$ but also by projecting the vector of the error terms of the linear model into $\mathcal{M}(\mathbb{X})^{\perp}$.
- Point (iii) of Theorem 2.3 can also be briefly written as

$$\boldsymbol{U} \,|\, \mathbb{X} \sim \big(\mathbf{0}_n,\, \sigma^2\,\mathbb{M}\big).$$

Definition 2.4 Residual mean square and residual degrees of freedom.
Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}) = r$.

1. *The* residual mean square[13] *of the model is a quantity* $\mathsf{SS}_e/(n-r)$ *and will be denoted as* MS_e. *That is,*

$$\mathsf{MS}_e = \frac{\mathsf{SS}_e}{n-r}.$$

2. *The* residual degrees of freedom[14] *of the model is the dimension of the residual space and will be denotes as* ν_e. *That is,*

$$\nu_e = n - r.$$

Theorem 2.4 Unbiased estimator of the residual variance.
The residual mean square MS_e *is an* unbiased *estimator (both conditionally given* $\mathbb{X}$ *and also with respect to the joint distribution of* $\boldsymbol{Y}$ *and* $\mathbb{X}$*) of the residual variance* σ^2 *in a linear model* $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \le k < n$.

Proof. Direct consequence of Theorem 2.3, point (iv).

❑

End of Lecture #2 (05/10/2016)

[13] *reziduální střední čtverec* [14] *reziduální stupně volnosti*

2.3 Normal equations

Start of Lecture #3 (06/10/2016)

The vector of fitted values $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$ is a projection of the response vector into $\mathcal{M}(\mathbb{X})$. Hence, it must be possible to write $\widehat{\boldsymbol{Y}}$ as a linear combination of the columns of the model matrix $\mathbb{X}$. That is, there exists $\boldsymbol{b} \in \mathbb{R}^k$ such that

$$\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}. \tag{2.6}$$

Notes.

- In a full-rank model ($\mathsf{rank}(\mathbb{X}_{n\times k}) = k$), linearly independent columns of $\mathbb{X}$ form a vector basis of $\mathcal{M}(\mathbb{X})$. Hence $\boldsymbol{b} \in \mathbb{R}^k$ such that $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$ is unique.
- If $\mathsf{rank}(\mathbb{X}_{n\times k}) = r < k$, a vector $\boldsymbol{b} \in \mathbb{R}^k$ such that $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$ is *not* unique.

We already know from the Gauss-Markov theorem (Theorem 2.2) that $\mathbb{E}(\widehat{\boldsymbol{Y}} \,|\, \mathbb{X}) = \mathbb{X}\boldsymbol{\beta}$. Hence if we manage to express $\widehat{\boldsymbol{Y}}$ as $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$ and $\boldsymbol{b}$ will be unique, we have a natural candidate for an estimator of the regression coefficients $\boldsymbol{\beta}$. Nevertheless, before we proceed to estimation of $\boldsymbol{\beta}$, we derive conditions that $\boldsymbol{b} \in \mathbb{R}^k$ must satisfy to fulfill also (2.6).

Definition 2.5 Sum of squares.
Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. *The function* $\mathsf{SS} : \mathbb{R}^k \longrightarrow \mathbb{R}$ *given as follows*

$$\mathsf{SS}(\boldsymbol{\beta}) = \left\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right\|^2 = \left(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right), \qquad \boldsymbol{\beta} \in \mathbb{R}^k$$

will be called the sum of squares[15] *of the model.*

Theorem 2.5 Least squares and normal equations.
Assume a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. *The vector of fitted values* $\widehat{\boldsymbol{Y}}$ *equals to* $\mathbb{X}\boldsymbol{b}$, $\boldsymbol{b} \in \mathbb{R}^k$ *if and only if* $\boldsymbol{b}$ *solves a linear system*

$$\mathbb{X}^\top\mathbb{X}\boldsymbol{b} = \mathbb{X}^\top\boldsymbol{Y}. \tag{2.7}$$

Proof.

$\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$, is a projection of $\boldsymbol{Y}$ into $\mathcal{M}(\mathbb{X})$

$\Leftrightarrow$ $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$ is the closest point to $\boldsymbol{Y}$ in $\mathcal{M}(\mathbb{X})$

$\Leftrightarrow$ $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$, where $\boldsymbol{b}$ minimizes $\mathsf{SS}(\boldsymbol{\beta}) = \left\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right\|^2$ over $\boldsymbol{\beta} \in \mathbb{R}^k$.

Let us find conditions under which the term $\mathsf{SS}(\boldsymbol{\beta})$ attains its minimal value over $\boldsymbol{\beta} \in \mathbb{R}^k$. To this end, a vector of the first derivatives (a gradient) and a matrix of the second derivatives (a Hessian) of $\mathsf{SS}(\boldsymbol{\beta})$ are needed.

$$\begin{aligned}
\mathsf{SS}(\boldsymbol{\beta}) &= \left\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right\|^2 = \left(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right) = \boldsymbol{Y}^\top\boldsymbol{Y} - 2\boldsymbol{Y}^\top\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta}. \\
\frac{\partial \mathsf{SS}}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) &= -2\,\mathbb{X}^\top\boldsymbol{Y} + 2\,\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta}.
\end{aligned}$$

[15] *součet čtverců*

$$\frac{\partial^2 \mathsf{SS}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}(\boldsymbol{\beta}) \;=\; 2\,\mathbb{X}^\top \mathbb{X}. \tag{2.8}$$

For any $\boldsymbol{\beta} \in \mathbb{R}^k$, the Hessian (2.8) is a positive semidefinite matrix and hence $\boldsymbol{b}$ minimizes $\mathsf{SS}(\boldsymbol{\beta})$ over $\boldsymbol{\beta} \in \mathbb{R}^k$ if and only if

$$\frac{\partial \mathsf{SS}}{\partial \boldsymbol{\beta}}(\boldsymbol{b}) = \mathbf{0}_k,$$

that is, if and only if

$$\mathbb{X}^\top \mathbb{X} \boldsymbol{b} = \mathbb{X}^\top \boldsymbol{Y}.$$

❑

Definition 2.6 Normal equations.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$. *The* system of normal equations[16] *or concisely* normal equations[17] *of the model is the linear system*

$$\mathbb{X}^\top \mathbb{X} \boldsymbol{b} = \mathbb{X}^\top \boldsymbol{Y},$$

or equivalently, the linear system

$$\mathbb{X}^\top (\boldsymbol{Y} - \mathbb{X}\boldsymbol{b}) = \mathbf{0}_k.$$

Note. In general, the linear system (2.7) of normal equations would not have to have a solution (a minimum of the sum of squares would not have to exist). Nevertheless, in our case, existence of the solution (and hence existence of a minimum of the sum of squares) follows from the fact that it corresponds to the projection $\widehat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ into the regression space $\mathcal{M}(\mathbb{X})$ and existence of the projection $\widehat{\boldsymbol{Y}}$ is guaranteed by the projection properties known from the linear algebra lectures. On the other hand, we can also show quite easily that there exists a solution to the normal equations (and hence there exists a minimum of the sum of squares) by using the following lemma.

Lemma 2.6 Vector spaces generated by the rows of the model matrix.

Let $\mathbb{X}_{n\times k}$ *be a real matrix. Then*

$$\mathcal{M}\big(\mathbb{X}^\top \mathbb{X}\big) = \mathcal{M}\big(\mathbb{X}^\top\big).$$

Proof. First note that $\mathcal{M}\big(\mathbb{X}^\top \mathbb{X}\big) = \mathcal{M}\big(\mathbb{X}^\top\big)$ is equivalent to $\mathcal{M}\big(\mathbb{X}^\top \mathbb{X}\big)^\perp = \mathcal{M}\big(\mathbb{X}^\top\big)^\perp$. We will show this by showing that for any $\boldsymbol{a} \in \mathbb{R}^k$ $\boldsymbol{a} \in \mathcal{M}\big(\mathbb{X}^\top\big)^\perp$ if and only if $\boldsymbol{a} \in \mathcal{M}\big(\mathbb{X}^\top \mathbb{X}\big)^\perp$.

(i) $\boldsymbol{a} \in \mathcal{M}\big(\mathbb{X}^\top\big)^\perp \;\Rightarrow\; \boldsymbol{a}^\top \mathbb{X}^\top = \mathbf{0}_n^\top \;\Rightarrow\; \boldsymbol{a}^\top \mathbb{X}^\top \mathbb{X} = \mathbf{0}_k^\top$
$\Leftrightarrow\; \boldsymbol{a} \in \mathcal{M}\big(\mathbb{X}^\top \mathbb{X}\big)^\perp$

[16] *systém normálních rovnic* [17] *normální rovnice*

(ii) $\boldsymbol{a} \in \mathcal{M}\left(\mathbb{X}^\top\mathbb{X}\right)^\perp \;\Rightarrow\; \boldsymbol{a}^\top\mathbb{X}^\top\mathbb{X} = \mathbf{0}_k^\top \;\Rightarrow\; \boldsymbol{a}^\top\mathbb{X}^\top\mathbb{X}\,\boldsymbol{a} = 0$
$\Rightarrow\; \|\mathbb{X}\boldsymbol{a}\| = 0 \;\Leftrightarrow\; \mathbb{X}\boldsymbol{a} = \mathbf{0}_n \;\Leftrightarrow\; \boldsymbol{a}^\top\mathbb{X}^\top = \mathbf{0}_n^\top$
$\Leftrightarrow\; \boldsymbol{a} \in \mathcal{M}\left(\mathbb{X}^\top\right)^\perp$

❑

Note. A vector space $\mathcal{M}\left(\mathbb{X}^\top\right)$ is a vector space generated by the columns of the matrix $\mathbb{X}^\top$, that is, it is a vector space generated by the *rows* of the matrix $\mathbb{X}$.

Notes.

- Existence of a solution to normal equations (2.7) follows from the fact that its right-hand side $\mathbb{X}^\top\boldsymbol{Y} \in \mathcal{M}\left(\mathbb{X}^\top\right)$ and $\mathcal{M}\left(\mathbb{X}^\top\right)$ is (by Lemma 2.6) the same space as a vector space generated by the columns of the matrix of the linear system ($\mathbb{X}^\top\mathbb{X}$).
- By Theorem A.1, all solutions to normal equations, i.e., a set of points that minimize the sum of squares $\mathsf{SS}(\boldsymbol{\beta})$ are given as $\boldsymbol{b} = \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{X}^\top\boldsymbol{Y}$, where $\left(\mathbb{X}^\top\mathbb{X}\right)^-$ is any pseudoinverse to $\mathbb{X}^\top\mathbb{X}$ (if $\mathsf{rank}\left(\mathbb{X}_{n\times k}\right) = r < k$, this pseudoinverse is not unique).
- We also have that for any $\boldsymbol{b} = \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{X}^\top\boldsymbol{Y}$:

$$\mathsf{SS}_e = \mathsf{SS}(\boldsymbol{b}).$$

Notation.

- In the following, symbol $\boldsymbol{b}$ will be exclusively used to denote any solution to normal equations, that is,

$$\boldsymbol{b} = \left(b_0, \ldots, b_{k-1}\right)^\top = \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{X}^\top\boldsymbol{Y}.$$

- For a full-rank linear model, ($\mathsf{rank}\left(\mathbb{X}_{n\times k}\right) = k$), the following holds:
  - The only pseudoinverse $\left(\mathbb{X}^\top\mathbb{X}\right)^-$ is $\left(\mathbb{X}^\top\mathbb{X}\right)^- = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$.
  - The only solution of normal equations is $\boldsymbol{b} = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ which is also a unique minimizer of the sum of squares $\mathsf{SS}(\boldsymbol{\beta})$.

  In this case, we will denote the unique solution to normal equations as $\widehat{\boldsymbol{\beta}}$. That is,

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \ldots, \widehat{\beta}_{k-1}\right)^\top = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y}.$$

2.4 Estimable parameters

We have seen in the previous section that the sum of squares $\mathsf{SS}(\boldsymbol{\beta})$ does not necessarily attain a unique minimum. This happens if the model matrix $\mathbb{X}_{n\times k}$ has linearly dependent columns (its rank $r < k$) and hence there exist (infinitely) many possibilities on how to express the vector of the fitted values $\widehat{\boldsymbol{Y}} \in \mathcal{M}(\mathbb{X})$ as a linear combination of the columns of the model matrix $\mathbb{X}$. In other words, there exist (infinitely) many vectors $\boldsymbol{b} \in \mathbb{R}^k$ such that $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$. This could also be interpreted as that there are (infinitely) many estimators of the regression parameters $\boldsymbol{\beta}$ leading to the (unique) unbiased estimator of the response mean $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{X}) = \mathbb{X}\boldsymbol{\beta}$. It then does not make much sense to talk about estimation of the regression parameters $\boldsymbol{\beta}$. To avoid such situations, we now define a notion of an *estimable parameter*[18] of a linear model.

Definition 2.7 Estimable parameter.
Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$. *Let* $\mathbf{l} \in \mathbb{R}^k$. *We say that a parameter*

$$\theta = \mathbf{l}^\top\boldsymbol{\beta}$$

is an estimable parameter *of the model if for all* $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{X})$ *the expression* $\mathbf{l}^\top\boldsymbol{\beta}$ *does not depend on a choice of a solution to the linear system* $\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu}$.

Notes.

- Definition of an estimable parameter is equivalent to the requirement

$$\forall\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^k \qquad \mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 \;\;\Rightarrow\;\; \mathbf{l}^\top\boldsymbol{\beta}_1 = \mathbf{l}^\top\boldsymbol{\beta}_2.$$

  That is, the estimable parameter is such a linear combination of the regression coefficients $\boldsymbol{\beta}$ which does not depend on a choice of the $\boldsymbol{\beta}$ leading to the same vector in the regression space $\mathcal{M}(\mathbb{X})$ (leading to the same vector of the response expectation $\boldsymbol{\mu}$).

- In a full-rank model ($\mathsf{rank}(\mathbb{X}_{n\times k}) = k$), columns of the model matrix $\mathbb{X}$ form a vector basis of the regression space $\mathcal{M}(\mathbb{X})$. It then follows from the properties of a vector basis that for any $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{X})$ there exist a unique $\boldsymbol{\beta}$ such that $\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu}$. Trivially, for any $\mathbf{l} \in \mathbb{R}^k$, the expression $\mathbf{l}^\top\boldsymbol{\beta}$ then does not depend on a choice of a solution to the linear system $\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ since there is only one such solution. In other words, in a full-rank model, any linear function of the regression coefficients $\boldsymbol{\beta}$ is estimable.

Definition 2.8 Estimable vector parameter.
Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$. *Let* $\mathbf{l} \in \mathbb{R}^k$. *Let* $\mathbf{l}_1, \ldots, \mathbf{l}_m \in \mathbb{R}^k$. *Let* $\mathbb{L}$ *be an* $m \times k$ *matrix having vectors* $\mathbf{l}_1^\top, \ldots, \mathbf{l}_m^\top$ *in its rows. We say that a vector parameter*

$$\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$$

is an estimable vector parameter *of the model if all parameters* $\theta_j = \mathbf{l}_j^\top\boldsymbol{\beta}$, $j = 1, \ldots, m$, *are estimable.*

---

[18] *odhadnutelný parametr*

Notes.

- Definition of an estimable parameter is equivalent to the requirement

$$\forall \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^k \qquad \mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 \;\; \Rightarrow \;\; \mathbb{L}\boldsymbol{\beta}_1 = \mathbb{L}\boldsymbol{\beta}_2.$$

- Trivially, a vector parameter $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}$ is always estimable. We also already know its BLUE which is the vector of fitted values $\widehat{\boldsymbol{Y}}$.
- In a full-rank model ($\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$), the regression coefficients vector $\boldsymbol{\beta}$ is an estimable vector parameter.

Example 2.1 (Overparameterized two-sample problem).

Consider a two-sample problem:

$$\begin{aligned} &\textit{Sample 1:} && Y_1, \ldots, Y_{n_1} && \overset{\textit{i.i.d.}}{\sim} Y^{(1)}, && Y^{(1)} \sim \big(\mu_1, \sigma^2\big), \\ &\textit{Sample 2:} && Y_{n_1+1}, \ldots, Y_{n_1+n_2} && \overset{\textit{i.i.d.}}{\sim} Y^{(2)}, && Y^{(2)} \sim \big(\mu_2, \sigma^2\big), \end{aligned}$$

and $Y_1, \ldots, Y_{n_1}, Y_{n_1+1}, \ldots, Y_{n_1+n_2}$ are assumed to be independent. This situation can be described by a linear model $\boldsymbol{Y} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $n = n_1 + n_2$, where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ \vdots \\ Y_{n_1+n_2} \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta} = \begin{pmatrix} \beta_0 + \beta_1 \\ \vdots \\ \beta_0 + \beta_1 \\ \hline \beta_0 + \beta_2 \\ \vdots \\ \beta_0 + \beta_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \hline \mu_2 \\ \vdots \\ \mu_2 \end{pmatrix}.$$

(i) Parameters $\mu_1 = \beta_0 + \beta_1$ and $\mu_2 = \beta_0 + \beta_2$ are (trivially) estimable.

(ii) None of the elements of the vector $\boldsymbol{\beta}$ is estimable. For example, take $\boldsymbol{\beta}_1 = \big(0,\, 1,\, 0\big)^\top$ and $\boldsymbol{\beta}_2 = \big(1,\, 0,\, -1\big)^\top$. We have $\mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 = \big(1, \ldots, 1,\, 0,\, \ldots,\, 0\big)^\top$ but none of the elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ is equal. This corresponds to the fact that two means μ_1 and μ_2 can be expressed in infinitely many ways using three numbers $\beta_0,\, \beta_1,\, \beta_2$ as $\mu_1 = \beta_0 + \beta_1$ and $\mu_2 = \beta_0 + \beta_2$.

(iii) A non-trivial estimable parameter is, e.g.,

$$\theta = \mu_2 - \mu_1 = \beta_2 - \beta_1 = \mathbf{l}^\top\boldsymbol{\beta}, \qquad \mathbf{l} = \big(0,\, -1,\, 1\big)^\top.$$

We have for $\boldsymbol{\beta}_1 = \big(\beta_{1,0},\, \beta_{1,1},\, \beta_{1,2}\big)^\top \in \mathbb{R}^3$ and $\boldsymbol{\beta}_2 = \big(\beta_{2,0},\, \beta_{2,1},\, \beta_{2,2}\big)^\top \in \mathbb{R}^3$:

$$\begin{aligned} \mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 \;\Leftrightarrow\; & \beta_{1,0} + \beta_{1,1} = \beta_{2,0} + \beta_{2,1}, \\ & \beta_{1,0} + \beta_{1,2} = \beta_{2,0} + \beta_{2,2} \\ \Rightarrow\; & \beta_{1,2} - \beta_{1,1} = \beta_{2,2} - \beta_{2,1} \;\Leftrightarrow\; \mathbf{l}^\top\boldsymbol{\beta}_1 = \mathbf{l}^\top\boldsymbol{\beta}_2. \end{aligned}$$

Definition 2.9 Contrast.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$. *An estimable parameter* $\theta = \boldsymbol{c}^\top\boldsymbol{\beta}$, *given by a real vector* $\boldsymbol{c} = \big(c_0, \ldots, c_{k-1}\big)^\top$ *which satisfies*

$$\boldsymbol{c}^\top \mathbf{1}_k = 0, \quad \textit{i.e.,} \sum_{j=0}^{k-1} c_j = 0,$$

is called contrast[19].

Definition 2.10 Orthogonal contrasts.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$. *Contrasts* $\theta = \boldsymbol{c}^\top\boldsymbol{\beta}$ *and* $\eta = \boldsymbol{d}^\top\boldsymbol{\beta}$ *given by orthogonal vectors* $\boldsymbol{c} = \big(c_0, \ldots, c_{k-1}\big)^\top$ *and* $\boldsymbol{d} = \big(d_0, \ldots, d_{k-1}\big)^\top$, *i.e., given by vectors* $\boldsymbol{c}$ *and* $\boldsymbol{d}$ *that satisfy* $\boldsymbol{c}^\top\boldsymbol{d} = 0$, *are called (mutually)* orthogonal *contrasts.*

Theorem 2.7 Estimable parameter, necessary and sufficient condition.

Assume a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$.

(i) Let $\mathbf{l} \in \mathbb{R}^k$. *Parameter* $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ *is an estimable parameter* if and only if

$$\mathbf{l} \;\in\; \mathcal{M}\big(\mathbb{X}^\top\big).$$

(ii) A vector $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ *is an estimable vector parameter* if and only if

$$\mathcal{M}\big(\mathbb{L}^\top\big) \;\subset\; \mathcal{M}\big(\mathbb{X}^\top\big).$$

Proof.

(i) $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ is estimable

$\Leftrightarrow \quad \forall\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^k \quad \mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 \;\Rightarrow\; \mathbf{l}^\top\boldsymbol{\beta}_1 = \mathbf{l}^\top\boldsymbol{\beta}_2$

$\Leftrightarrow \quad \forall\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^k \quad \mathbb{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = \mathbf{0}_n \;\Rightarrow\; \mathbf{l}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = 0$

$\Leftrightarrow \quad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^k \quad \mathbb{X}\boldsymbol{\gamma} = \mathbf{0}_n \;\Rightarrow\; \mathbf{l}^\top\boldsymbol{\gamma} = 0$

$\Leftrightarrow \quad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^k \quad \boldsymbol{\gamma}$ orthogonal to all *rows* of $\mathbb{X} \;\Rightarrow\; \mathbf{l}^\top\boldsymbol{\gamma} = 0$

$\Leftrightarrow \quad \forall\, \boldsymbol{\gamma} \in \mathbb{R}^k \quad \boldsymbol{\gamma} \in \mathcal{M}\big(\mathbb{X}^\top\big)^\perp \;\Rightarrow\; \mathbf{l}^\top\boldsymbol{\gamma} = 0$

$\Leftrightarrow \quad \mathbf{l} \in \mathcal{M}\big(\mathbb{X}^\top\big).$

(ii) Direct consequence of point (i).

[19] *kontrast*

❑

Note. In a full-rank model ($\mathsf{rank}(\mathbb{X}_{n\times k}) = k < n$), $\mathcal{M}(\mathbb{X}^\top) = \mathbb{R}^k$. That is, any linear function of $\boldsymbol{\beta}$ is indeed estimable (statement that we already concluded from the definition of an estimable parameter).

Theorem 2.8 Gauss–Markov for estimable parameters.

Let $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ be an estimable parameter of a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta},\ \sigma^2\mathbf{I}_n)$. Let $\boldsymbol{b}$ be any solution to the normal equations. The statistic

$$\widehat{\theta} \ = \ \mathbf{l}^\top\boldsymbol{b}$$

then satisfies:

(i) $\widehat{\theta}$ does not depend on a choice of the solution $\boldsymbol{b}$ of the normal equations, i.e., it does not depend on a choice of a pseudoinverse in $\boldsymbol{b} = (\mathbb{X}^\top\mathbb{X})^-\mathbb{X}^\top\boldsymbol{Y}$.

(ii) $\widehat{\theta}$ is, conditionally given $\mathbb{X}$, the best linear unbiased estimator (BLUE) *of the parameter θ.*

(iii) $\mathsf{var}(\widehat{\theta} \,|\, \mathbb{X}) = \sigma^2\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}$ *, that is,*

$$\widehat{\theta}\,|\,\mathbb{X} \ \sim \ \Big(\theta,\ \sigma^2\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}\Big),$$

where $\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}$ does not depend on a choice of the pseudoinverse $(\mathbb{X}^\top\mathbb{X})^-$.

If additionally $\mathbf{l} \neq \mathbf{0}_k$ then $\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l} > 0$.

Let further $\theta_1 = \mathbf{l}_1^\top\boldsymbol{\beta}$ and $\theta_2 = \mathbf{l}_2^\top\boldsymbol{\beta}$ be estimable parameters. Let

$$\widehat{\theta}_1 \ = \ \mathbf{l}_1^\top\boldsymbol{b}, \qquad \widehat{\theta}_2 \ = \ \mathbf{l}_2^\top\boldsymbol{b}.$$

Then

$$\mathsf{cov}(\widehat{\theta}_1,\, \widehat{\theta}_2 \,|\, \mathbb{X}) \ = \ \sigma^2\mathbf{l}_1^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}_2,$$

where $\mathbf{l}_1^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}_2$ does not depend on a choice of the pseudoinverse $(\mathbb{X}^\top\mathbb{X})^-$.

Proof.

(i) Let $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ be two solutions to normal equations, that is,

$$\mathbb{X}^\top\boldsymbol{Y} = \mathbb{X}^\top\mathbb{X}\,\boldsymbol{b}_1 = \mathbb{X}^\top\mathbb{X}\,\boldsymbol{b}_2.$$

By Theorem 2.5 (Least squares and normal equations):

$$\Leftrightarrow \quad \widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}_1 \quad \& \quad \widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}_2,$$

that is, $\mathbb{X}\boldsymbol{b}_1 = \mathbb{X}\boldsymbol{b}_2$.

Estimability of θ:

$$\Rightarrow \quad \mathbf{l}^\top\boldsymbol{b}_1 = \mathbf{l}^\top\boldsymbol{b}_2.$$

(ii) Parameter $\theta = \mathbf{l}^\top \boldsymbol{\beta}$ is estimable. By Theorem 2.7:

$$\begin{aligned} &\Leftrightarrow \quad \mathbf{l} \in \mathcal{M}(\mathbb{X}^\top) \\ &\Leftrightarrow \quad \mathbf{l} = \mathbb{X}^\top \boldsymbol{a} \text{ for some } \boldsymbol{a} \in \mathbb{R}^n \\ &\Rightarrow \quad \widehat{\theta} = \boldsymbol{a}^\top \mathbb{X} \boldsymbol{b} = \boldsymbol{a}^\top \widehat{\boldsymbol{Y}}. \end{aligned}$$

That is, $\widehat{\theta}$ is a linear function of $\widehat{\boldsymbol{Y}}$ which is the BLUE of $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$. It then follows that $\widehat{\theta}$ is the BLUE of the parameter

$$\boldsymbol{a}^\top \boldsymbol{\mu} = \boldsymbol{a}^\top \mathbb{X} \boldsymbol{\beta} = \mathbf{l}^\top \boldsymbol{\beta} = \theta.$$

(iii) **Proof/calculations were available on the blackboard in K1.**

❑

Theorem 2.9 Gauss–Markov for estimable vector parameter.

Let $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ be an estimable vector parameter of a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta},\ \sigma^2 \mathbf{I}_n)$. Let $\boldsymbol{b}$ be any solution to normal equations. The statistic

$$\widehat{\boldsymbol{\theta}} \ = \ \mathbb{L}\boldsymbol{b}$$

then satisfies:

(i) $\widehat{\boldsymbol{\theta}}$ does not depend on a choice of the solution $\boldsymbol{b}$ of the normal equations.

(ii) $\widehat{\boldsymbol{\theta}}$ is, conditionally given $\mathbb{X}$, the best linear unbiased estimator (BLUE) *of the vector parameter $\boldsymbol{\theta}$.*

(iii) $\mathsf{var}(\widehat{\boldsymbol{\theta}} \,|\, \mathbb{X}) = \sigma^2 \mathbb{L}(\mathbb{X}^\top \mathbb{X})^- \mathbb{L}^\top$, that is,

$$\widehat{\boldsymbol{\theta}} \,|\, \mathbb{X} \ \sim \ \Big(\boldsymbol{\theta},\ \sigma^2 \mathbb{L}(\mathbb{X}^\top \mathbb{X})^- \mathbb{L}^\top\Big),$$

where $\mathbb{L}(\mathbb{X}^\top \mathbb{X})^- \mathbb{L}^\top$ does not depend on a choice of the pseudoinverse $(\mathbb{X}^\top \mathbb{X})^-$.

If additionally $m \leq r$ and the rows of the matrix $\mathbb{L}$ are linearly independent then $\mathbb{L}(\mathbb{X}^\top \mathbb{X})^- \mathbb{L}^\top$ is a positive definite (invertible) matrix.

Proof. Direct consequence of Theorem 2.8, except positive definiteness of $\mathbb{L}(\mathbb{X}^\top \mathbb{X})^- \mathbb{L}^\top$ in situations when $\mathbb{L}$ has linearly independent rows.

Positive definiteness of $\mathbb{L}(\mathbb{X}^\top \mathbb{X})^- \mathbb{L}^\top$ if $\mathbb{L}_{m\times k}$ has linearly independent rows:

Proof/calculations were available on the blackboard in K1.

❑

Consequence of Theorem 2.9.

Assume a full-rank linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k < n$. *The statistic*

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

then satisfies:

(i) $\widehat{\boldsymbol{\beta}}$ *is, conditionally given* $\mathbb{X}$, *the* best linear unbiased estimator (BLUE) *of the regression coefficients* $\boldsymbol{\beta}$.

(ii) $\mathsf{var}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{X}\big) = \sigma^2\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$, *that is,*

$$\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{X} \;\sim\; \Big(\boldsymbol{\beta},\, \sigma^2\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big).$$

Proof. Use $\mathbb{L} = \mathbf{I}_k$ in Theorem 2.9. ❑

End of Lecture #3 (06/10/2016)

2.5 Parameterizations of a linear model

Start of Lecture #4 (12/10/2016)

For given response $\boldsymbol{Y} = \big(Y_1, \ldots, Y_n\big)^\top$ and given set of covariates $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, many different sets of regressors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and related model matrices $\mathbb{X}$ can be proposed. In this section, we define a notion of equivalent linear models which basically says when two (or more) different sets of regressors, i.e., two (or more) different model matrices (derived from one set of covariates) provide models that do not differ with respect to fundamental model properties.

2.5.1 Equivalent linear models

Definition 2.11 Equivalent linear models.

Assume two linear models: M_1*:* $\boldsymbol{Y} \,\big|\, \mathbb{X}_1 \sim \big(\mathbb{X}_1\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\big)$*, where* $\mathbb{X}_1$ *is an* $n \times k$ *matrix with* $\mathsf{rank}\big(\mathbb{X}_1\big) = r$ *and* M_2*:* $\boldsymbol{Y} \,\big|\, \mathbb{X}_2 \sim \big(\mathbb{X}_2\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n\big)$*, where* $\mathbb{X}_2$ *is an* $n \times l$ *matrix with* $\mathsf{rank}\big(\mathbb{X}_2\big) = r$*. We say that models* M_1 *and* M_2 *are* equivalent *if their regression spaces are the same. That is, if*

$$\mathcal{M}\big(\mathbb{X}_1\big) = \mathcal{M}\big(\mathbb{X}_2\big).$$

Notes.

- The two equivalent models:
  - have the same hat matrix $\mathbb{H} = \mathbb{X}_1\big(\mathbb{X}_1^\top\mathbb{X}_1\big)^-\mathbb{X}_1^\top = \mathbb{X}_2\big(\mathbb{X}_2^\top\mathbb{X}_2\big)^-\mathbb{X}_2^\top$ and a vector of fitted values $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$;
  - have the same residual projection matrix $\mathbb{M} = \mathbf{I}_n - \mathbb{H}$ and a vector of residuals $\boldsymbol{U} = \mathbb{M}\boldsymbol{Y}$;
  - have the same value of the residual sum of squares $\mathsf{SS}_e = \boldsymbol{U}^\top\boldsymbol{U}$, residual degrees of freedom $\nu_e = n - r$ and the residual mean square $\mathsf{MS}_e = \mathsf{SS}_e/(n - r)$.
- The two equivalent models provide two different parameterizations of one situation. Nevertheless, practical interpretation of the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^k$ and $\boldsymbol{\gamma} \in \mathbb{R}^l$ in the two models might be different. In practice, both parameterizations might be useful and this is also the reason why it often makes sense to deal with both parameterizations.

2.5.2 Full-rank parameterization of a linear model

Any linear model can be parameterized such that the model matrix has linearly independent columns, i.e., is of a full-rank. To see this, consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\big)$, where $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \leq k < n$. If $\mathbb{Q}_{n\times r}$ is a matrix with the orthonormal vector basis of $\mathcal{M}\big(\mathbb{X}\big)$ in its columns (that is, $\mathsf{rank}(\mathbb{Q}) = r$), the linear model

$$\boldsymbol{Y} \,\big|\, \mathbb{Q} \sim \big(\mathbb{Q}\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n\big) \tag{2.9}$$

is equivalent to the original model with the model matrix $\mathbb{X}$. Nevertheless, parameterization of a model using the orthonormal basis and the $\mathbb{Q}$ matrix is only rarely used in practice since the interpretation of the regression coefficients $\boldsymbol{\gamma}$ in model (2.9) is usually quite awkward.

Parameterization of a linear model using the orthonormal basis matrix $\mathbb{Q}$ is indeed not the only full-rank parameterization of a given linear model. There always exist infinitely many full-rank

parameterizations and in reasonable practical analyses, it should always be possible to choose such a full-rank parameterization or even parameterizations that also provide practically interpretable regression coefficients.

Example 2.2 (Different parameterizations of a two-sample problem).
Let us again consider a two-sample problem (see also Example 2.1). That is,

$$\begin{aligned} &\textit{Sample 1:} && Y_1, \ldots, Y_{n_1} && \overset{\text{i.i.d.}}{\sim} Y^{(1)}, && Y^{(1)} \sim \big(\mu_1, \sigma^2\big), \\ &\textit{Sample 2:} && Y_{n_1+1}, \ldots, Y_{n_1+n_2} && \overset{\text{i.i.d.}}{\sim} Y^{(2)}, && Y^{(2)} \sim \big(\mu_2, \sigma^2\big), \end{aligned}$$

and $Y_1, \ldots, Y_{n_1}, Y_{n_1+1}, \ldots, Y_{n_1+n_2}$ are assumed to be independent. This situation can be described by differently parameterized linear models $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta}, \sigma^2\, \mathbf{I}_n\big)$, $n = n_1 + n_2$ where the model matrix $\mathbb{X}$ is always divided into two blocks as

$$\mathbb{X} = \left(\frac{\mathbb{X}_1}{\mathbb{X}_2}\right),$$

where $\mathbb{X}_1$ is an $n_1 \times k$ matrix having n_1 identical rows $\boldsymbol{x}_1^\top$ and $\mathbb{X}_2$ is an $n_2 \times k$ matrix having n_2 identical rows $\boldsymbol{x}_2^\top$. The response mean vector $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)$ is then

$$\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta} = \left(\frac{\mathbb{X}_1\boldsymbol{\beta}}{\mathbb{X}_2\boldsymbol{\beta}}\right) = \left(\frac{\begin{matrix}\boldsymbol{x}_1^\top\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_1^\top\boldsymbol{\beta}\end{matrix}}{\begin{matrix}\boldsymbol{x}_2^\top\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_2^\top\boldsymbol{\beta}\end{matrix}}\right) = \left(\frac{\begin{matrix}\mu_1 \\ \vdots \\ \mu_1\end{matrix}}{\begin{matrix}\mu_2 \\ \vdots \\ \mu_2\end{matrix}}\right).$$

That is, parameterization of the model is given by choices of vectors $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, $\boldsymbol{x}_1 \neq \mathbf{0}_k$, $\boldsymbol{x}_2 \neq \mathbf{0}_k$ leading to expressions of the means of the two samples as

$$\mu_1 = \boldsymbol{x}_1^\top\boldsymbol{\beta}, \qquad \mu_2 = \boldsymbol{x}_2^\top\boldsymbol{\beta}.$$

The rank of the model is always $r = 2$.

Overparameterized model $\boldsymbol{x}_1 = \big(1,\, 1,\, 0\big)^\top$, $\boldsymbol{x}_2 = \big(1,\, 0,\, 1\big)^\top$:

$$\mathbb{X} = \left(\frac{\begin{matrix}1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0\end{matrix}}{\begin{matrix}1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1\end{matrix}}\right), \quad \boldsymbol{\beta} = \begin{pmatrix}\beta_0 \\ \beta_1 \\ \beta_2\end{pmatrix}, \qquad \begin{aligned}\mu_1 &= \beta_0 + \beta_1, \\ \mu_2 &= \beta_0 + \beta_2.\end{aligned}$$

Orthonormal basis $\boldsymbol{x}_1 = \big(1/\sqrt{n_1},\, 0\big)^\top$, $\boldsymbol{x}_2 = \big(0,\, 1/\sqrt{n_2}\big)^\top$:

$$\mathbb{X} = \mathbb{Q} = \left(\begin{array}{cc} \frac{1}{\sqrt{n_1}} & 0 \\ \vdots & \vdots \\ \frac{1}{\sqrt{n_1}} & 0 \\ \hline 0 & \frac{1}{\sqrt{n_2}} \\ \vdots & \vdots \\ 0 & \frac{1}{\sqrt{n_2}} \end{array}\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \begin{aligned} \mu_1 &= \frac{1}{\sqrt{n_1}}\,\beta_1, \qquad \beta_1 = \sqrt{n_1}\,\mu_1, \\ \mu_2 &= \frac{1}{\sqrt{n_2}}\,\beta_2, \qquad \beta_2 = \sqrt{n_2}\,\mu_2. \end{aligned}$$

Group means $\boldsymbol{x}_1 = \big(1,\, 0\big)^\top$, $\boldsymbol{x}_2 = \big(0,\, 1\big)^\top$:

$$\mathbb{X} = \left(\begin{array}{cc} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{array}\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \begin{aligned} \mu_1 &= \beta_1, \\ \mu_2 &= \beta_2. \end{aligned}$$

This could also be viewed as the overparameterized model constrained by a condition $\beta_0 = 0$.

Group differences $\boldsymbol{x}_1 = \big(1,\, 1\big)^\top$, $\boldsymbol{x}_2 = \big(1,\, 0\big)^\top$:

$$\mathbb{X} = \left(\begin{array}{cc} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ \hline 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{array}\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \begin{aligned} \mu_1 &= \beta_0 + \beta_1, \\ \mu_2 &= \beta_0, \qquad\qquad \beta_1 = \mu_1 - \mu_2. \end{aligned}$$

This could also be viewed as the overparameterized model constrained by a condition $\beta_2 = 0$.

Deviations from the mean of the means $\boldsymbol{x}_1 = \big(1,\, 1\big)^\top$, $\boldsymbol{x}_2 = \big(1,\, -1\big)^\top$:

$$\mathbb{X} = \left(\begin{array}{cc} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ \hline 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{array}\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \begin{aligned} \mu_1 &= \beta_0 + \beta_1, \qquad \beta_0 = \frac{\mu_1 + \mu_2}{2}, \\ \mu_2 &= \beta_0 - \beta_1, \qquad \beta_1 = \mu_1 - \frac{\mu_1 + \mu_2}{2} \\ &\qquad\qquad\quad = \frac{\mu_1 + \mu_2}{2} - \mu_2. \end{aligned}$$

This could also be viewed as the overparameterized model constrained by a condition $\beta_1 + \beta_2 = 0$.

Except the overparameterized model, all above parameterizations are based on a model matrix having full-rank $r = 2$.

2.6 Matrix algebra and a method of least squares

We have seen in Section 2.5 that any linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ can be reparameterized such that the model matrix $\mathbb{X}$ has linearly independent columns, that is, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$. Remind now expressions of some quantities that must be calculated when dealing with the least squares estimation of parameters of the full-rank linear model:

$$\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top, \qquad \mathbb{M} = \mathbf{I}_n - \mathbb{H} = \mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top,$$

$$\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}, \quad \mathsf{var}\big(\widehat{\boldsymbol{Y}} \,|\, \mathbb{X}\big) = \sigma^2\mathbb{H} = \sigma^2\,\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top,$$

$$\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}, \qquad \mathsf{var}\big(\boldsymbol{U} \,|\, \mathbb{X}\big) = \sigma^2\mathbb{M} = \sigma^2\Big\{\mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\Big\},$$

$$\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}, \qquad \mathsf{var}\big(\widehat{\boldsymbol{\beta}} \,|\, \mathbb{X}\big) = \sigma^2\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}.$$

The only non-trivial calculation involved in above expressions is calculation of the inverse $\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$. Nevertheless, all above expressions (and many others needed in a context of the least squares estimation) can be calculated without explicit evaluation of the matrix $\mathbb{X}^\top\mathbb{X}$. Some of above expressions can even be evaluated without knowing explicitely the form of the $\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$ matrix. To this end, methods of matrix algebra can be used (and are used by all reasonable software routines dealing with the least squares estimation). Two methods, known from the course *Fundamentals of Numerical Mathematics (NMNM201)*, that have direct usage in the context of least squares are:

- QR decomposition;
- Singular value decomposition (SVD)

applied to the model matrix $\mathbb{X}$. Both of them can be used, among the other things, to find the orthonormal vector basis of the regression space $\mathcal{M}\big(\mathbb{X}\big)$ and to calculate expressions mentioned above.

2.6.1 QR decomposition

QR decomposition of the model matrix is used, for example, by the R software (R Core Team, 2016) to estimate a linear model by the method of least squares. If $\mathbb{X}_{n\times k}$ is a real matrix with $\mathsf{rank}\big(\mathbb{X}\big) = k < n$ then we know from the course *Fundamentals of Numerical Mathematics (NMNM201)* that it can be decomposed as

$$\mathbb{X} = \mathbb{Q}\boldsymbol{R},$$

where

$$\mathbb{Q}_{n\times k} = \big(\boldsymbol{q}_1,\, \ldots,\, \boldsymbol{q}_k\big), \qquad \boldsymbol{q}_j \in \mathbb{R}^k,\ j = 1, \ldots, k,$$

$\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$ is an orthonormal basis of $\mathcal{M}\big(\mathbb{X}\big)$ and $\boldsymbol{R}_{k\times k}$ is *upper triangular* matrix. That is,

$$\mathbb{Q}^\top\mathbb{Q} = \mathbf{I}_k, \qquad \mathbb{Q}\mathbb{Q}^\top = \mathbb{H}.$$

We then have

$$\mathbb{X}^\top\mathbb{X} = \boldsymbol{R}^\top\underbrace{\mathbb{Q}^\top\mathbb{Q}}_{\mathbf{I}_k}\boldsymbol{R} = \boldsymbol{R}^\top\boldsymbol{R}. \tag{2.10}$$

That is, $\boldsymbol{R}^\top\boldsymbol{R}$ is a Cholesky (square root) decomposition of the symmetric matrix $\mathbb{X}^\top\mathbb{X}$. Note that this is a special case of an LU decomposition for symmetric matrices. Decomposition (2.10) can now be used to get easily (i) matrix $\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$, (ii) a value of its determinant or a value of determinant of $\mathbb{X}^\top\mathbb{X}$, (iii) solution to normal equations.

(i) Matrix $\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$.

$$\left(\mathbb{X}^\top\mathbb{X}\right)^{-1} = \left(\boldsymbol{R}^\top\boldsymbol{R}\right)^{-1} = \boldsymbol{R}^{-1}\left(\boldsymbol{R}^\top\right)^{-1} = \boldsymbol{R}^{-1}\left(\boldsymbol{R}^{-1}\right)^\top.$$

That is, to invert the matrix $\mathbb{X}^\top\mathbb{X}$, we only have to invert the upper triangular matrix $\boldsymbol{R}$.

(ii) Determinant of $\mathbb{X}^\top\mathbb{X}$ and $\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$.
Let $r_1, \ldots, r_k$ denote diagonal elements of the matrix $\boldsymbol{R}$. We then have

$$\det\left(\mathbb{X}^\top\mathbb{X}\right) = \det\left(\boldsymbol{R}^\top\boldsymbol{R}\right) = \left\{\det(\boldsymbol{R})\right\}^2 = \left(\prod_{j=1}^{k} r_j\right)^2,$$

$$\det\left\{\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right\} = \left\{\det\left(\mathbb{X}^\top\mathbb{X}\right)\right\}^{-1}.$$

(iii) Solution to normal equations $\widehat{\boldsymbol{\beta}} = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y}$.
We can obtain $\widehat{\boldsymbol{\beta}}$ by solving:

$$\begin{aligned}\mathbb{X}^\top\mathbb{X}\,\boldsymbol{b} &= \mathbb{X}^\top\boldsymbol{Y}\\ \boldsymbol{R}^\top\boldsymbol{R}\,\boldsymbol{b} &= \boldsymbol{R}^\top\mathbb{Q}^\top\boldsymbol{Y}\\ \boldsymbol{R}\,\boldsymbol{b} &= \mathbb{Q}^\top\boldsymbol{Y}. \qquad (2.11)\end{aligned}$$

That is, to get $\widehat{\boldsymbol{\beta}}$, it is only necessary to solve a linear system with the upper triangular system matrix which can easily be done by backward substitution.

Further, the right-hand-side $\boldsymbol{c} = \left(c_1, \ldots, c_k\right)^\top := \mathbb{Q}^\top\boldsymbol{Y}$ of the linear system (2.11) additionally serves to calculate the vector of fitted values. We have

$$\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \mathbb{Q}\mathbb{Q}^\top\boldsymbol{Y} = \mathbb{Q}\,\boldsymbol{c} = \sum_{j=1}^{k} c_j\boldsymbol{q}_j.$$

That is, the vector $\boldsymbol{c}$ provides coefficients of the linear combination of the orthonormal vector basis of the regression space $\mathcal{M}(\mathbb{X})$ that provide the fitted values $\widehat{\boldsymbol{Y}}$.

2.6.2 SVD decomposition

Use of the SVD decomposition for the least squares will not be explained in detail in this course. It is covered by the *Fundamentals of Numerical Mathematics (NMNM201)* course.

Chapter 3

Normal Linear Model

Until now, all proved theorems did not pose any distributional assumptions on the random vectors $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top$, $\boldsymbol{X}_i = \left(X_{i,0}, \ldots, X_{i,k-1}\right)^\top$, $i = 1, \ldots, n$, that represent the data. We only assumed a certain form of the (conditional) expectation and the (conditional) covariance matrix of $\boldsymbol{Y} = \left(Y_1, \ldots, Y_n\right)^\top$ given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ (given the model matrix $\mathbb{X}$). In this chapter, we will additionally assume that the response is conditionally normally distributed given the regressors which will lead us to the *normal* linear model.

3.1 Normal linear model

With i.i.d. data $\big(Y_i,\, \boldsymbol{X}_i^\top\big)^\top \overset{\text{i.i.d.}}{\sim} \big(Y,\, \boldsymbol{X}^\top\big)^\top$, $i = 1, \ldots, n$, we mentioned in Section 1.2.6 situation when it was additionally assumed that $Y \,\big|\, \boldsymbol{X} \sim \mathcal{N}\big(\boldsymbol{X}^\top\boldsymbol{\beta},\, \sigma^2\big)$. For the full data $(\boldsymbol{Y},\, \mathbb{X})$, this implies

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big). \tag{3.1}$$

Strictly speaking, the original data vectors $\big(Y_i,\, \boldsymbol{X}_i^\top\big)^\top$, $i = 1, \ldots, n$, do not have to be i.i.d. with respect to their joint distribution to satisfy (3.1). Remember that the joint density of the response vector and all the regressors can be decomposed as

$$f_{\boldsymbol{Y},\mathbb{X}}(\boldsymbol{y},\, \mathbf{x}) = f_{\boldsymbol{Y}|\mathbb{X}}\big(\boldsymbol{y} \,\big|\, \mathbf{x}\big)\, f_{\mathbb{X}}(\mathbf{x}), \qquad \boldsymbol{y} \in \mathbb{R}^n,\, \mathbf{x} \in \mathcal{X}^n.$$

Property (3.1) is related to the conditional density $f_{\boldsymbol{Y}|\mathbb{X}}$ which is then given as

$$f_{\boldsymbol{Y}\,|\,\mathbb{X}}\big(\boldsymbol{y} \,\big|\, \mathbf{x}\big) = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma}\, \varphi\Big(\frac{y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta}}{\sigma} \Big) \right\}, \quad \boldsymbol{y} \in \mathbb{R}^n,\, \mathbf{x} \in \mathcal{X}^n.$$

On the other hand, the property (3.1) says nothing concerning the joint distribution of the regressors represented by their joint density $f_{\mathbb{X}}$. Since most of the results shown in this chapter can be derived while assuming just (3.1) we will do so and open the space for applications of the developed theory even in situations when the regressors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are perhaps not i.i.d. but jointly generated by some distribution with a general density $f_{\mathbb{X}}$.

Definition 3.1 Normal linear model with general data.
The data $\big(\boldsymbol{Y},\, \mathbb{X}\big)$, *satisfy a* normal linear model[1] *if*

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

where $\boldsymbol{\beta} = \big(\beta_0,\, \ldots,\, \beta_{k-1}\big)^\top \in \mathbb{R}^k$ *and* $0 < \sigma^2 < \infty$ *are unknown parameters.*

Lemma 3.1 Error terms in a normal linear model.
Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$. *The error terms* $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} = \big(\varepsilon_1,\, \ldots,\, \varepsilon_n\big)^\top$ *then satisfy*

(i) $\boldsymbol{\varepsilon} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\, \mathbf{I}_n\big)$.

(ii) $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\, \mathbf{I}_n\big)$.

(iii) $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \varepsilon,\ i = 1, \ldots, n,\ \varepsilon \sim \mathcal{N}\big(0,\, \sigma^2\big)$.

Proof.

(i) follows from the fact that a multivariate normal distribution is preserved after linear transformations (only the mean and the covariance matrix changes accordingly).

[1] *normální lineární model*

(ii) follows from (i) and the fact that the conditional distribution $\varepsilon \mid \mathbb{X}$ does not depend on the condition and hence the (unconditional) distribution of ε must be the same.

(iii) follows from (ii) and basic properties of the multivariate normal distribution (indepedence is the same as uncorrelatedness, univariate margins are normal as well).

3.2 Properties of the least squares estimators under the normality

Theorem 3.2 Least squares estimators under the normality.

Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r$. *Let* $\mathbb{L}_{m\times k}$ *is a real matrix with non-zero rows* $\mathbf{l}_1^\top, \ldots, \mathbf{l}_m^\top$ *such that* $\boldsymbol{\theta} = \big(\theta_1, \ldots, \theta_m\big)^\top = \big(\mathbf{l}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \mathbb{L}\boldsymbol{\beta}$ *is an estimable parameter. Let* $\widehat{\boldsymbol{\theta}} = \big(\widehat{\theta}_1, \ldots, \widehat{\theta}_m\big)^\top = \big(\mathbf{l}_1^\top\boldsymbol{b}, \ldots, \mathbf{l}_m^\top\boldsymbol{b}\big)^\top = \mathbb{L}\boldsymbol{b}$ *be its least squares estimator. Further, let*

$$\mathbb{V} = \mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{L}^\top = \big(v_{j,t}\big)_{j,t=1,\ldots,m},$$

$$\mathbb{D} = \mathsf{diag}\left(\frac{1}{\sqrt{v_{1,1}}}, \ldots, \frac{1}{\sqrt{v_{m,m}}}\right),$$

$$T_j = \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\mathsf{MS}_e\, v_{j,j}}}, \qquad j = 1, \ldots, m,$$

$$\boldsymbol{T} = \big(T_1, \ldots, T_m\big)^\top = \frac{1}{\sqrt{\mathsf{MS}_e}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big).$$

The following then holds.

(i) $\widehat{\boldsymbol{Y}} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbb{H}\big)$.

(ii) $\boldsymbol{U} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\, \mathbb{M}\big)$.

(iii) $\widehat{\boldsymbol{\theta}} \,|\, \mathbb{X} \sim \mathcal{N}_m\big(\boldsymbol{\theta},\, \sigma^2\, \mathbb{V}\big)$.

(iv) Statistics $\widehat{\boldsymbol{Y}}$ *and* $\boldsymbol{U}$ *are conditionally, given* $\mathbb{X}$*, independent.*

(v) Statistics $\widehat{\boldsymbol{\theta}}$ *and* SS_e *are conditionally, given* $\mathbb{X}$*, independent.*

(vi) $\dfrac{\big\|\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big\|^2}{\sigma^2} \;\sim\; \chi^2_r.$

(vii) $\dfrac{\mathsf{SS}_e}{\sigma^2} \;\sim\; \chi^2_{n-r}.$

(viii) For each $j = 1, \ldots, m,$ $\qquad T_j \;\sim\; \mathsf{t}_{n-r}.$

(ix) $\boldsymbol{T} \,|\, \mathbb{X} \;\sim\; \mathsf{mvt}_{m,n-r}\big(\mathbb{D}\mathbb{V}\mathbb{D}\big).$

(x) If additionally $\mathsf{rank}\big(\mathbb{L}_{m\times k}\big) = m \le r$ *then the matrix* $\mathbb{V}$ *is invertible and*

$$\frac{1}{m}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)^\top\,\big(\mathsf{MS}_e\,\mathbb{V}\big)^{-1}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \;\sim\; \mathcal{F}_{m,\,n-r}.$$

End of Lecture #4 (12/10/2016)

Proof. **Proof/calculations were available on the blackboard in KI.** ❑

Start of Lecture #6 (19/10/2016)

In a full-rank linear model, we have $\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\,\mathbb{X}^\top\boldsymbol{Y}$ and under the normality assumption, Theorem 3.2 can be used to state additional properties of the LSE $\widehat{\boldsymbol{\beta}}$ of the regression coefficients $\boldsymbol{\beta}$.

Consequence of Theorem 3.2: Least squares estimator of the regression coefficients in a full-rank normal linear model.

Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$. *Further, let*

$$\mathbb{V} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} = \left(v_{j,t}\right)_{j,t=0,\ldots,k-1},$$

$$\mathbb{D} = \mathsf{diag}\left(\frac{1}{\sqrt{v_{0,0}}},\, \ldots,\, \frac{1}{\sqrt{v_{k-1,k-1}}}\right).$$

The following then holds.

(i) $\widehat{\boldsymbol{\beta}} \,|\, \mathbb{X} \sim \mathcal{N}_k(\boldsymbol{\beta},\, \sigma^2\, \mathbb{V})$.

(ii) Statistics $\widehat{\boldsymbol{\beta}}$ *and* SS_e *are conditionally, given* $\mathbb{X}$, *independent.*

(iii) For each $j = 0, \ldots, k-1$, $T_j := \dfrac{\widehat{\beta}_j - \beta_j}{\sqrt{\mathsf{MS}_e\, v_{j,j}}} \sim \mathsf{t}_{n-k}$.

(iv) $\boldsymbol{T} := \left(T_0, \ldots, T_{k-1}\right)^\top = \dfrac{1}{\sqrt{\mathsf{MS}_e}}\, \mathbb{D}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \mathsf{mvt}_{k,n-k}(\mathbb{D}\mathbb{V}\mathbb{D})$.

(v) $\dfrac{1}{k}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \mathsf{MS}_e^{-1}\, \mathbb{X}^\top\mathbb{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \mathcal{F}_{k,\,n-k}$.

Proof. Use $\mathbb{L} = \mathbf{I}_k$ in Theorem 3.2 and realize that the only pseudoinverse to the matrix $\mathbb{X}^\top\mathbb{X}$ in a full-rank model is the inverse $\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$.

❑

Theorem 3.2 and its consequence can now be used to perform principal statistical inference, i.e., calculation of confidence intervals and regions, testing statistical hypotheses, in a *normal* linear model.

3.2.1 Statistical inference in a full-rank normal linear model

Assume a full-rank normal linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$ and keep denoting $\mathbb{V} = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1} = \left(v_{j,t}\right)_{j,t=0,\ldots,k-1}$.

Inference on a chosen regression coefficient

First, take a chosen $j \in \{0, \ldots, k-1\}$. We then have the following.

- **Standard error of $\widehat{\beta}_j$ and confidence interval for β_j**
  We have $\mathsf{var}\left(\widehat{\beta}_j \,|\, \mathbb{X}\right) = \sigma^2\, v_{j,j}$ (Consequence of Theorem 2.9) which is unbiasedly estimated as $\mathsf{MS}_e\, v_{j,j}$ (Theorem 2.4). The square root of this quantity, i.e., estimated standard deviation of $\widehat{\beta}_j$ is then called as *standard error*[2] of the estimator $\widehat{\beta}_j$. That is,

$$\mathsf{S.E.}\left(\widehat{\beta}_j\right) = \sqrt{\mathsf{MS}_e\, v_{j,j}}. \tag{3.2}$$

[2] *směrodatná, příp. standardní chyba*

The standard error (3.2) is also the denominator of the t-statistic T_j from point (iii) of Consequence of Theorem 3.2. Hence the lower and the upper bounds of the Wald-type $(1-\alpha)\,100\%$ confidence interval for β_j based on the statistic T_j are

$$\widehat{\beta}_j \;\pm\; \mathsf{S.E.}\big(\widehat{\beta}_j\big)\,\mathsf{t}_{n-k}\Big(1-\frac{\alpha}{2}\Big).$$

Analogously, also one-sided confidence interval can be calculated.

- **Test on a value of β_j**

  Suppose that for a given $\beta_j^0 \in \mathbb{R}$, we aim in testing $\quad \mathrm{H}_0\colon\ \beta_j = \beta_j^0,$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathrm{H}_1\colon\ \beta_j \neq \beta_j^0.$

  The Wald-type test based on point (iii) of Consequence of Theorem 3.2 proceeds as follows:

  Test statistic: $\qquad T_{j,0} = \dfrac{\widehat{\beta}_j - \beta_j^0}{\mathsf{S.E.}\big(\widehat{\beta}_j\big)} = \dfrac{\widehat{\beta}_j - \beta_j^0}{\sqrt{\mathsf{MS}_e\, v_{j,j}}}.$

  Reject H_0 if $\qquad |T_{j,0}| \geq t_{n-k}\Big(1-\dfrac{\alpha}{2}\Big).$

  P-value when $T_{j,0} = t_{j,0}$: $\quad p = 2\,\mathsf{CDF}_{t,\,n-k}\big(-|t_{j,0}|\big).$

  Analogously, also one-sided tests can be conducted.

Simultaneous inference on a vector of regression coefficients

When the interest lies in the inference for the full vector of the regression coefficients $\boldsymbol{\beta}$, the following procedures can be used.

- **Simultaneous confidence region**[3] **for $\boldsymbol{\beta}$**

  It follows from point (v) of Consequence of Theorem 3.2 that the simultaneous $(1-\alpha)\,100\%$ confidence region for $\boldsymbol{\beta}$ is the set

  $$\Big\{\boldsymbol{\beta} \in \mathbb{R}^k :\ \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big)^\top \big(\mathsf{MS}_e^{-1}\mathbb{X}^\top\mathbb{X}\big)\,\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big) \;<\; k\,\mathcal{F}_{k,n-k}(1-\alpha)\Big\},$$

  which is an ellipsoid with $\quad$ center: $\quad \widehat{\boldsymbol{\beta}},$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ shape matrix: $\quad \mathsf{MS}_e\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1} = \widehat{\mathsf{var}}\big(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{X}\big),$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ diameter: $\quad \sqrt{k\,\mathcal{F}_{k,n-k}(1-\alpha)}.$

  Remember from the linear algebra and geometry lectures that the shape matrix determines the principal directions of the ellipsoid as those are given by the eigen vectors of this matrix. In this case, the principal directions of the *confidence ellipsoid* are given by the eigen vectors of the estimated covariance matrix $\widehat{\mathsf{var}}\big(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{X}\big)$.

- **Test on a value of $\boldsymbol{\beta}$**

  Suppose that for a given $\boldsymbol{\beta}^0 \in \mathbb{R}^k$, we aim in testing $\quad \mathrm{H}_0\colon\ \boldsymbol{\beta} = \boldsymbol{\beta}^0,$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathrm{H}_1\colon\ \boldsymbol{\beta} \neq \boldsymbol{\beta}^0.$

  The Wald-type test based on point (v) of Consequence of Theorem 3.2 proceeds as follows:

  Test statistic: $\qquad Q_0 = \dfrac{1}{k}\,\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big)^\top\,\mathsf{MS}_e^{-1}\,\mathbb{X}^\top\mathbb{X}\,\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\big).$

  Reject H_0 if $\qquad Q_0 \geq \mathcal{F}_{k,n-k}(1-\alpha).$

  P-value when $Q_0 = q_0$: $\quad p = 1 - \mathsf{CDF}_{\mathcal{F},\,k,n-k}\big(q_0\big).$

[3] *simultánní konfidenční oblast*

3.2.2 Statistical inference in a general rank normal linear model

Let us now assume a geneal rank normal linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \leq k$.

Inference on an estimable parameter

Let $\theta = \mathbf{l}^\top \boldsymbol{\beta}$, $\mathbf{l} \neq \mathbf{0}_k$, be an estimable parameter and let $\widehat{\theta} = \mathbf{l}^\top \boldsymbol{b}$ be its least squares estimator.

- **Standard error of $\widehat{\theta}$ and confidence interval for θ**
  We have $\mathsf{var}\big(\widehat{\theta} \,|\, \mathbb{X}\big) = \sigma^2\, \mathbf{l}^\top \big(\mathbb{X}^\top \mathbb{X}\big)^- \mathbf{l}$ (Theorem 2.8) which is unbiasedly estimated as $\mathsf{MS}_e\, \mathbf{l}^\top \big(\mathbb{X}^\top \mathbb{X}\big)^- \mathbf{l}$ (Theorem 2.4). Hence the standard error of $\widehat{\theta}$ is

$$\mathsf{S.E.}\big(\widehat{\theta}\big) = \sqrt{\mathsf{MS}_e\, \mathbf{l}^\top \big(\mathbb{X}^\top \mathbb{X}\big)^- \mathbf{l}}. \tag{3.3}$$

  The standard error (3.3) is also the denominator of the appropriate t-statistic from point (viii) of Theorem 3.2. Hence the lower and the upper bounds of the Wald-type $(1-\alpha)\,100\%$ confidence interval for θ based on this t-statistic are

$$\widehat{\theta} \;\pm\; \mathsf{S.E.}\big(\widehat{\theta}\big)\, \mathsf{t}_{n-r}\Big(1 - \frac{\alpha}{2}\Big).$$

  Analogously, also one-sided confidence interval can be calculated.

- **Test on a value of θ**
  Suppose that for a given $\theta^0 \in \mathbb{R}$, we aim in testing $\quad \mathrm{H}_0\colon \;\; \theta = \theta^0,$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathrm{H}_1\colon \;\; \theta \neq \theta^0.$

  The Wald-type test based on point (viii) of Theorem 3.2 proceeds as follows:

  Test statistic: $\quad T_0 = \dfrac{\widehat{\theta} - \theta^0}{\mathsf{S.E.}\big(\widehat{\theta}\big)} = \dfrac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\, \mathbf{l}^\top \big(\mathbb{X}^\top \mathbb{X}\big)^- \mathbf{l}}}.$

  Reject H_0 if $\quad |T_0| \geq t_{n-r}\Big(1 - \dfrac{\alpha}{2}\Big).$

  P-value when $T_0 = t_0$: $\quad p = 2\, \mathsf{CDF}_{t,\,n-r}\big(-|t_0|\big).$

  Analogously, also one-sided tests can be conducted.

Simultaneous inference on an estimable vector parameter

Finally, let $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ be an estimable parameter, where $\mathbb{L}$ is an $m \times k$ matrix with $m \leq r$ linearly independent rows. Let $\widehat{\boldsymbol{\theta}} = \mathbb{L}\boldsymbol{b}$ be the least squares estimator of $\boldsymbol{\theta}$.

- **Simultaneous confidence region for $\boldsymbol{\theta}$**
  It follows from point (x) of Theorem 3.2 that the simultaneous $(1-\alpha)\,100\%$ confidence region for $\boldsymbol{\theta}$ is the set

$$\Big\{\boldsymbol{\theta} \in \mathbb{R}^m :\; \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)^\top \Big\{\mathsf{MS}_e\, \mathbb{L}\big(\mathbb{X}^\top \mathbb{X}\big)^- \mathbb{L}^\top\Big\}^{-1} \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big) \;<\; m\, \mathcal{F}_{m,n-r}(1-\alpha)\Big\},$$

  which is an ellipsoid with $\quad$ center: $\quad \widehat{\boldsymbol{\theta}},$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ shape matrix: $\quad \mathsf{MS}_e\, \mathbb{L}\big(\mathbb{X}^\top \mathbb{X}\big)^- \mathbb{L}^\top = \widehat{\mathsf{var}}\big(\widehat{\boldsymbol{\theta}} \,|\, \mathbb{X}\big),$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ diameter: $\quad \sqrt{m\, \mathcal{F}_{m,n-r}(1-\alpha)}.$

- **Test on a value of $\boldsymbol{\theta}$**

  Suppose that for a given $\boldsymbol{\theta}^0 \in \mathbb{R}^m$, we aim in testing $\quad \mathrm{H}_0\text{:} \quad \boldsymbol{\theta} = \boldsymbol{\theta}^0,$
  $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathrm{H}_1\text{:} \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}^0.$

  The Wald-type test based on point (x) of Theorem 3.2 proceeds as follows:

  Test statistic: $$Q_0 = \frac{1}{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^\top \left\{\mathsf{MS}_e\, \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right).$$

  Reject H_0 if $\qquad Q_0 \geq \mathcal{F}_{m,n-r}(1-\alpha).$

  P-value when $Q_0 = q_0$: $\quad p = 1 - \mathsf{CDF}_{\mathcal{F},\,m,n-r}\left(q_0\right).$

Note. Assume again a full-rank model ($r = k$) and take $\mathbb{L}$ as a submatrix of the identity matrix $\mathbf{I}_k$ by selecting some of its rows. The above procedures can then be used to infer simultaneously on a subvector of the regression coefficients $\boldsymbol{\beta}$.

Note. All tests, confidence intervals and confidence regions derived in this Section were derived under the assumption of a *normal* linear model. Nevertheless, we show in Chapter 13 that under certain conditions, all those methods of statistical inference remain *asymptotically* valid even if normality does not hold.

3.3 Confidence interval for the model based mean, prediction interval

We keep assuming that the data $\left(Y_i,\, \boldsymbol{X}_i^\top\right)^\top$, $i = 1, \ldots, n$, follow a normal linear model. That is,

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

from which it also follows

$$Y_i \,\big|\, \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma^2), \qquad i = 1, \ldots, n.$$

Furthermore, the error terms $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$ are i.i.d. distributed as $\varepsilon \sim \mathcal{N}(0,\, \sigma^2)$ (Lemma 3.1).

Remember that $\mathcal{X} \subseteq \mathbb{R}^k$ denotes a sample space of the regressor random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Let $\boldsymbol{x}_{new} \in \mathcal{X}$ and let

$$Y_{new} = \boldsymbol{x}_{new}^\top \boldsymbol{\beta} + \varepsilon_{new},$$

where $\varepsilon_{new} \sim \mathcal{N}(0,\, \sigma^2)$ is independent of $\boldsymbol{\varepsilon} = \left(\varepsilon_1,\, \ldots, \varepsilon_n\right)^\top$. A value of Y_{new} is thus a value of a "new" observation sampled from the conditional distribution

$$Y_{new} \,\big|\, \boldsymbol{X}_{new} = \boldsymbol{x}_{new} \sim \mathcal{N}(\boldsymbol{x}_{new}^\top \boldsymbol{\beta},\, \sigma^2)$$

independently of the "old" observations. We will now tackle two important problems:

(i) Interval estimation of $\mu_{new} := \mathbb{E}\big(Y_{new} \,\big|\, \boldsymbol{X}_{new} = \boldsymbol{x}_{new}\big) = \boldsymbol{x}_{new}^\top \boldsymbol{\beta}$.

(ii) Interval estimation of the value of the random variable Y_{new} itself, given the regressor vector $\boldsymbol{X}_{new} = \boldsymbol{x}_{new}$.

Solution to the outlined problems will be provided by the following theorem.

Theorem 3.3 Confidence interval for the model based mean, prediction interval. *Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r$. Let $\boldsymbol{x}_{new} \in \mathcal{X} \cap \mathcal{M}\big(\mathbb{X}^\top\big)$, $\boldsymbol{x}_{new} \neq \mathbf{0}_k$. Let $\varepsilon_{new} \sim \mathcal{N}(0,\, \sigma^2)$ is independent of $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$. Finally, let $Y_{new} = \boldsymbol{x}_{new}^\top \boldsymbol{\beta} + \varepsilon_{new}$. The following then holds:*

(i) $\mu_{new} = \boldsymbol{x}_{new}^\top \boldsymbol{\beta}$ is estimable,

$$\widehat{\mu}_{new} = \boldsymbol{x}_{new}^\top \boldsymbol{b}$$

is its best linear unbiased estimator (BLUE) with the standard error of

$$\mathsf{S.E.}\big(\widehat{\mu}_{new}\big) = \sqrt{\mathsf{MS}_e\, \boldsymbol{x}_{new}^\top \left(\mathbb{X}^\top \mathbb{X}\right)^- \boldsymbol{x}_{new}}$$

and the lower and the upper bound of the $(1-\alpha)\,100\%$ confidence interval for μ_{new} are

$$\widehat{\mu}_{new} \pm \mathsf{S.E.}\big(\widehat{\mu}_{new}\big)\, t_{n-r}\Big(1 - \frac{\alpha}{2}\Big). \tag{3.4}$$

(ii) A (random) interval with the bounds

$$\widehat{\mu}_{new} \pm \mathsf{S.E.P.}\big(\boldsymbol{x}_{new}\big)\, t_{n-r}\Big(1 - \frac{\alpha}{2}\Big), \tag{3.5}$$

where

$$\mathsf{S.E.P.}\big(\boldsymbol{x}_{new}\big) = \sqrt{\mathsf{MS}_e \left\{1 \,+\, \boldsymbol{x}_{new}^\top \big(\mathbb{X}^\top \mathbb{X}\big)^- \boldsymbol{x}_{new}\right\}}, \tag{3.6}$$

covers with the probability of $(1-\alpha)$ the value of Y_{new}.

Proof. **Proof/calculations were available on the blackboard in KI.** ❑

Terminology *(Confidence interval for the model based mean, prediction interval, standard error of prediction).*

- The interval with the bounds (3.4) is called the *confidence interval for the model based mean.*
- The interval with the bounds (3.5) is called the *prediction interval.*
- The quantity (3.6) is called the *standard error of prediction.*

Terminology *(Fitted regression function).*

Suppose that the corresponding linear model is of full-rank with the LSE $\widehat{\boldsymbol{\beta}}$ of the regression coefficients. The function

$$\widehat{m}(\boldsymbol{x}) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}, \qquad \boldsymbol{x} \in \mathcal{X},$$

which, by Theorem 3.3, provides BLUE's of the values of

$$\mu(\boldsymbol{x}) := \mathbb{E}\big(Y_{new} \,\big|\, \boldsymbol{X}_{new} = \boldsymbol{x}\big) = \boldsymbol{x}^\top \boldsymbol{\beta}$$

and also provides predictions for $Y_{new} = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon_{new}$, is called the *fitted regression function.*[4]

Terminology *(Confidence band around the regression function, prediction band).*

As was explained in Section 1.1.3, the regressors $\boldsymbol{X}_i \in \mathcal{X} \subseteq \mathbb{R}^k$ used in the linear model are often obtained by transforming some original covariates $\boldsymbol{Z}_i \in \mathcal{Z} \subseteq \mathbb{R}^p$. Common situation is that $\mathcal{Z} \subseteq \mathbb{R}$ is an interval and

$$\boldsymbol{X}_i = \big(X_{i,0}, \ldots, X_{i,k-1}\big)^\top = \big(t_0(Z_i), \ldots, t_{k-1}(Z_i)\big)^\top = \boldsymbol{t}(Z_i), \qquad i = 1, \ldots, n,$$

where $\boldsymbol{t} : \mathbb{R} \longrightarrow \mathbb{R}^k$ is a suitable transformation such that

$$\mathbb{E}\big(Y_i \,\big|\, Z_i\big) = \boldsymbol{t}^\top(Z_i)\boldsymbol{\beta} = \boldsymbol{X}_i^\top \boldsymbol{\beta}.$$

Suppose again that the corresponding linear model is of full-rank with the LSE $\widehat{\boldsymbol{\beta}}$ of the regression coefficients. Confidence intervals for the model based mean or prediction intervals can then be calculated for an (equidistant) sequence of values $z_{new,1}, \ldots, z_{new,N} \in \mathcal{Z}$ and then drawn over a scatterplot of observed data $\big(Y_1, Z_1\big)^\top, \ldots, \big(Y_n, Z_n\big)^\top$. In this way, two different bands with a fitted regression function

$$\widehat{m}(z) = \boldsymbol{t}^\top(z)\widehat{\boldsymbol{\beta}}, \qquad z \in \mathcal{Z},$$

going through the middle of both the bands, are obtained. In this context,

(i) The band based on the confidence intervals for the model based mean (Eq. 3.4) is called the *confidence band around the regression function;*[5]

(ii) The band based on the prediction intervals (Eq. 3.5) is called the *prediction band.*[6]

[4] *odhadnutá regresní funkce* [5] *pás spolehlivosti okolo regresní funkce* [6] *predikční pás*

3.4 Distribution of the linear hypotheses test statistics under the alternative

Beginning of skipped part

Section 3.2 provided classical tests of the linear hypotheses (hypotheses on the values of estimable parameters). To allow for power or sample size calculations, we additionally need distribution of the test statistics under the alternatives.

Theorem 3.4 Distribution of the linear hypothesis test statistics under the alternative.

Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *Let* $\mathbf{l} \neq \mathbf{0}_k$ *such that* $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ *is estimable. Let* $\widehat{\theta} = \mathbf{l}^\top\boldsymbol{b}$ *be its LSE. Let* $\theta^0,\, \theta^1 \in \mathbb{R}$, $\theta^0 \neq \theta^1$ *and let*

$$T_0 = \frac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\, \mathbf{l}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbf{l}}}.$$

Then under the hypothesis $\theta = \theta^1$,

$$T_0 \,|\, \mathbb{X} \sim \mathsf{t}_{n-r}(\lambda), \qquad \lambda = \frac{\theta^1 - \theta^0}{\sqrt{\sigma^2\, \mathbf{l}^\top\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbf{l}}}.$$

Proof. **Proof/calculations were available on the blackboard in K1.** ❑

Note. The statistic T_0 is the test statistic to test the null hypothesis H_0: $\theta = \theta^0$ using point (viii) of Theorem 3.2.

Theorem 3.5 Distribution of the linear hypotheses test statistics under the alternative.

Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *Let* $\mathbb{L}_{m\times k}$ *be a real matrix with* $m \leq r$ *linearly independent rows such that* $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ *is estimable. Let* $\widehat{\boldsymbol{\theta}} = \mathbb{L}\boldsymbol{b}$ *be its LSE. Let* $\boldsymbol{\theta}^0,\, \boldsymbol{\theta}^1 \in \mathbb{R}^m$, $\boldsymbol{\theta}^0 \neq \boldsymbol{\theta}^1$ *and let*

$$Q_0 = \frac{1}{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^\top \left\{\mathsf{MS}_e\, \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\right\}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right).$$

Then under the hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}^1$,

$$Q_0 \,|\, \mathbb{X} \sim \mathcal{F}_{m,n-r}(\lambda), \qquad \lambda = \left(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0\right)^\top \left\{\sigma^2\, \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\right\}^{-1}\left(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0\right).$$

Proof. **Proof/calculations were available on the blackboard in K1.** ❑

Note. The statistic Q_0 is the test statistic to test the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ using point (x) of Theorem 3.2.

Note. We derived only a conditional (given the regressors) distribution of the test statistics at hand. This corresponds to the fact that power and sample size calculations for linear models are mainly used in the area of *designed experiments*[7] where the regressor values, i.e., the model matrix $\mathbb{X}$ is assumed to be fixed and not random. A problem of the sample size calculation then involves not only calculation of needed sample size n but also determination of the form of the model matrix $\mathbb{X}$. More can be learned in the course *Experimental Design (NMST436)*.[8]

End of skipped part

[7] *navržené experimenty* [8] *Návrhy experimentů (NMST436)*

Chapter 4

Basic Regression Diagnostics

We will now start from considering the original response-covariate data. That is, we assume that data are represented by n random vectors $\left(Y_i,\, \boldsymbol{Z}_i^\top\right)^\top$, $\boldsymbol{Z}_i = \left(Z_{i,1},\, \ldots,\, Z_{i,p}\right)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$. We keep considering that the principal aim of the statistical analysis is to find a suitable model to express the (conditional) response expectation $\mathbb{E}\big(Y_i \,\big|\, \boldsymbol{Z}_i\big)$, $i = 1, \ldots, n$, in summary the response vector conditional expectation $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$, where $\mathbb{Z}$ is a matrix with vectors $\boldsymbol{Z}_1$, …, $\boldsymbol{Z}_n$ in its rows. Suppose that $\boldsymbol{t} : \mathcal{Z} \longrightarrow \mathcal{X} \subseteq \mathbb{R}^k$ is a transformation of the covariates leading to the model matrix of regressors

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{t}^\top(\boldsymbol{Z}_1) \\ \vdots \\ \boldsymbol{t}^\top(\boldsymbol{Z}_n) \end{pmatrix} =: \boldsymbol{t}(\mathbb{Z}), \qquad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \leq k.$$

4.1 (Normal) linear model assumptions

Basis for statistical inference shown by now was derived while assuming a linear model for the data, i.e., while assuming that $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \boldsymbol{t}^\top(\mathbb{Z})\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$ and $\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \sigma^2\,\mathbf{I}_n$. For the data $\big(Y_i,\, \boldsymbol{X}_i^\top\big)^\top$, $i = 1, \ldots, n$, where we directly work with the response-regressors pairs, this means the following assumptions ($i = 1, \ldots, n$):

(A1) $\mathbb{E}\big(Y_i \,\big|\, \boldsymbol{X}_i = \boldsymbol{x}\big) = \boldsymbol{x}^\top\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$ and (almost all) $\boldsymbol{x} \in \mathcal{X}$.

$\equiv$ Correct regression function $m(\boldsymbol{z}) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}$, $\boldsymbol{z} \in \mathcal{Z}$, correct choice of transformation $\boldsymbol{t}$ of the original covariates leading to linearity of the (conditional) response expectation.

(A2) $\mathsf{var}\big(Y_i \,\big|\, \boldsymbol{X}_i = \boldsymbol{x}\big) = \sigma^2$ for some σ^2 irrespective of (almost all) values of $\boldsymbol{x} \in \mathcal{X}$.

$\equiv$ The conditional response variance is constant (does not depend on the covariates or other factors) $\equiv$*homoscedasticity*[1] of the response.

(A3) $\mathsf{cov}\big(Y_i,\, Y_l \,\big|\, \mathbb{X} = \mathbf{x}\big) = 0$, $i \neq l$, for (almost all) $\mathbf{x} \in \mathcal{X}^n$.

$\equiv$ The responses are conditionally uncorrelated.

Some of our results (especially those shown in Chapter 3) were derived while additionally assuming normality of the response, i.e., while assuming

(A4) $Y_i \,|\, \boldsymbol{X}_i = \boldsymbol{x} \sim \mathcal{N}\big(\boldsymbol{x}^\top\boldsymbol{\beta},\, \sigma^2\big)$, for (almost all) $\boldsymbol{x} \in \mathcal{X}$.

$\equiv$ Normalityof the response.

If we take the error terms of the linear model, i.e., the vector $\big(\varepsilon_1,\, \ldots,\, \varepsilon_n\big)^\top = \boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} = \big(Y_1 - \boldsymbol{X}_1^\top\boldsymbol{\beta},\, \ldots,\, Y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\big)^\top$, the above assumptions can also be stated as saying that there exists $\boldsymbol{\beta} \in \mathbb{R}^k$ for which the error terms satisfy the following.

(A1) $\mathbb{E}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i = \boldsymbol{x}\big) = 0$ for (almost all) $\boldsymbol{x} \in \mathcal{X}$, and consequently also $\mathbb{E}\big(\varepsilon_i\big) = 0$, $i = 1, \ldots, n$.

$\equiv$ This again means that a structural part of the model stating that $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$ is correctly specified, or in other words, that the regression function of the model is correctly specified.

(A2) $\mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i = \boldsymbol{x}\big) = \sigma^2$ for some σ^2 which is constant irrespective of (almost all) values of $\boldsymbol{x} \in \mathcal{X}$. Consequently also $\mathsf{var}\big(\varepsilon_i\big) = 0$, $i = 1, \ldots, n$.

$\equiv$ The error variance is constant $\equiv$ *homoscedasticity* of the errors.

(A3) $\mathsf{cov}\big(\varepsilon_i,\, \varepsilon_l \,\big|\, \mathbb{X} = \mathbf{x}\big) = 0$, $i \neq l$, for (almost all) $\mathbf{x} \in \mathcal{X}^n$. Consequently also $\mathsf{cov}\big(\varepsilon_i,\, \varepsilon_l\big) = 0$, $i \neq l$.

$\equiv$ The errors are uncorrelated.

Possible assumption of *normality* is transferred into the errors as

(A4) $\varepsilon_i \,\big|\, \boldsymbol{X}_i = \boldsymbol{x} \sim \mathcal{N}\big(0,\, \sigma^2\big)$ for (almost all) $\boldsymbol{x} \in \mathcal{X}$ and consequently also $\varepsilon_i \sim \mathcal{N}\big(0,\, \sigma^2\big)$, $i = 1, \ldots, n$.

$\equiv$ The errors are normally distributed and owing to previous assumptions, $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma^2)$.

[1] *homoskedasticita*

Remember now that many important results, especially those already derived in Chapter 2, are valid even without assuming normality of the response/errors. Moreover, we shall show in Chapter 13 that also majority of inferential tools based on results of Chapters 3 and 5 are, under certain conditions, asymptotically valid even if normality does not hold.

In general, if inferential tools based on a statistical model with certain properties (assumptions) are to be used, we should verify, at least into some extent, validity of those assumptions with a particular dataset. In a context of regression models, the tools to verify the model assumptions are usually referred to as *regression diagnostic*[2] tools. In this chapter, we provide only the most basic graphical methods. Additional, more advanced tools of the regression diagnostics will be provided in Chapters 11 and 14.

As already mentioned above, the assumptions (A1)–(A4) are not equally important. Some of them are not needed to justify usage of a particular inferential tool (estimator, statistical test, ...), see assumptions and proofs of corresponding Theorems. This should be taken into account when using the regression diagnostics. It is indeed not necessary to verify those assumptions that are not needed for a specific task. It should finally be mentioned that with respect to the importance of the assumptions (A1)–(A4), far the most important is assumption (A1) concerning a correct specification of the regression function. Remember that practically all Theorems in this lecture that are related to the inference on the parameters of a linear model use in their proofs, in some sense, the assumption $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{X}) \in \mathcal{M}(\mathbb{X})$. Hence if this is not satisfied, majority of the traditional statistical inference is not correct. In other words, special attention in any data analysis should be devoted to verifying the assumption (A1) related to a correct specification of the regression function.

As we shall show, the assumptions of the linear model are basically checked through exploration of the properties of the residuals $\boldsymbol{U}$ of the model, where

$$\boldsymbol{U} = \mathbb{M}\boldsymbol{Y}, \qquad \mathbb{M} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}^\top\mathbb{X})^-\mathbb{X}^\top = (m_{i,l})_{i,l=1,\ldots,n}.$$

When doing so, it is exploited that each of assumptions (A1)–(A4) implies a certain property of the residuals stated earlier in Theorems 2.3 (Basic properties of the residuals and the residual sum of squares) and 3.2 (Properties of the LSE under the normality). It follows from those theorems (or their proofs) the following:

1. (A1) $\implies \mathbb{E}(\boldsymbol{U} \,|\, \mathbb{X}) = \mathbf{0}_n.$
2. (A1) & (A2) & (A3) $\implies \mathsf{var}(\boldsymbol{U} \,|\, \mathbb{X}) = \sigma^2\,\mathbb{M}.$
3. (A1) & (A2) & (A3) & (A4) $\implies \boldsymbol{U} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbf{0}_n,\, \sigma^2\,\mathbb{M}).$

Usually, the right-hand side of the implication is verified and if it is found not to be satisfied, we know that also the left-hand side of the implication (a particular assumption or a set of assumptions) is not fulfilled. Clearly, if we conclude that the right-hand side of the implication is fulfilled, we still do not know whether the left-hand side (a model assumption) is valid. Nevertheless, it is common to most of the statistical diagnostic tools that they are only able to reveal unsatisfied model assumptions but are never able to confirm their validity.

End of Lecture #6 (19/10/2016)
Start of Lecture #8 (26/10/2016)

An uncomfortable property of the residuals of the linear model is the fact that even if the errors ($\boldsymbol{\varepsilon}$) are homoscedastic ($\mathsf{var}(\varepsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$), the residuals $\boldsymbol{U}$ are, in general, *heteroscedastic* (having unequal variances). Indeed, even if the assumption (A2) if fulfilled, we have $\mathsf{var}(\boldsymbol{U} \,|\, \mathbb{X}) = \sigma^2\,\mathbb{M}$, $\mathsf{var}(U_i \,|\, \mathbb{X}) = \sigma^2\,m_{i,i}$ ($i = 1, \ldots, n$), where note that the residual projection matrix $\mathbb{M}$, in general, does not have a constant diagonal $m_{1,1}, \ldots, m_{n,n}$. Moreover, the matrix $\mathbb{M}$ is even not a diagonal matrix. That is, even if the errors $\varepsilon_1, \ldots, \varepsilon_n$ are uncorrelated, the residuals $U_1, \ldots, U_n$ are, in general, (coditionally given the regressors) correlated. This must be taken

[2] *regresní diagnostika*

into account when the residuals $\boldsymbol{U}$ are used to check validity of assumption (A2). The problem of heteroscedasticity of the residuals $\boldsymbol{U}$ is then partly solved be defining so called *standardized residuals*.

4.2 Standardized residuals

Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, with the vector or residuals $\boldsymbol{U} = \big(U_1, \ldots, U_n\big)$, the residual mean square MS_e, and the residual projection matrix $\mathbb{M}$ having a diagonal $\big(m_{1,1}, \ldots, m_{n.n}\big)$. The following definition is motivated by the facts following the properties of residuals shown in Theorem 2.3:

$$\mathbb{E}\big(\boldsymbol{U} \,|\, \mathbb{X}\big) = \mathbf{0}_n, \qquad \mathsf{var}\big(\boldsymbol{U} \,|\, \mathbb{X}\big) = \sigma^2\, \mathbb{M},$$
$$\mathbb{E}\big(U_i \,|\, \mathbb{X}\big) = 0, \qquad \mathsf{var}\big(U_i \,|\, \mathbb{X}\big) = \sigma^2\, m_{i,i}, \qquad i = 1, \ldots, n.$$

Definition 4.1 Standardized residuals.

The standardized residuals[3] *or the vector of standardized residuals of the model is a vector* $\boldsymbol{U}^{std} = \big(U_1^{std}, \ldots, U_n^{std}\big)$, *where*

$$U_i^{std} = \begin{cases} \dfrac{U_i}{\sqrt{\mathsf{MS}_e\, m_{i,i}}}, & m_{i,i} > 0, \\ \textit{undefined}, & m_{i,i} = 0, \end{cases} \qquad i = 1, \ldots, n.$$

Theorem 4.1 Moments of standardized residuals under normality.

Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$ *and let for chosen* $i \in \{1, \ldots, n\}$, $m_{i,i} > 0$. *Then*

$$\mathbb{E}\big(U_i^{std} \,|\, \mathbb{X}\big) = 0, \qquad \mathsf{var}\big(U_i^{std} \,|\, \mathbb{X}\big) = 1.$$

Proof. **Proof/calculations were available on the blackboard in KI.**

Lemma B.2 used in the proof.

Notes.

- Unfortunately, even in a normal linear model, the standardized residuals $U_1^{std}, \ldots, U_n^{std}$ are, in general,
  - neither normally distributed;
  - nor uncorrelated.
- In some literature (and some software packages), the standardized residuals are called *studentized residuals*[4].
- In other literature including those course notes (and many software packages including R), the term *studentized* residuals is reserved for a different quantity which we shall define in Chapter 14.

[3] *standardizovaná rezidua* [4] *studentizovaná rezidua*

4.3 Graphical tools of regression diagnostics

In the whole section, the columns of the model matrix $\mathbb{X}$ (the regressors), are denoted as $\boldsymbol{X}^0, \ldots, \boldsymbol{X}^{k-1}$, i.e.,

$$\mathbb{X} = \left(\boldsymbol{X}^0, \ldots, \boldsymbol{X}^{k-1}\right).$$

Remember that usually $\boldsymbol{X}^0 = \left(1, \ldots, 1\right)^\top$ is an intercept column. Further, in many situations, see Section 5.2 dealing with a submodel obtained by omitting some regressors, the current model matrix $\mathbb{X}$ is the model matrix of just a candidate submodel (playing the role of the model matrix $\mathbb{X}^0$ in Section 5.2) and perhaps additional regressors are available to model the response expectation $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right)$. Let us denote them as $\boldsymbol{V}^1, \ldots, \boldsymbol{V}^m$. That is, in the notation of Section 5.2,

$$\mathbb{X}^1 = \left(\boldsymbol{V}^1, \ldots, \boldsymbol{V}^m\right).$$

The reminder of this section provides purely an overview of basic residual plots that are used as basic diagnostic tools in the context of a linear regression. More explanation on use of those plots will be/was provided during the lecture and the exercise classes.

4.3.1 (A1) Correctness of the regression function

To detect:

Overall inappropriateness of the regression function

$\Rightarrow$ scatterplot $\left(\widehat{\boldsymbol{Y}}, \boldsymbol{U}\right)$ of residuals versus fitted values.

Nonlinearity of the regression function with respect to a particular regressor $\boldsymbol{X}^j$

$\Rightarrow$ scatterplot $\left(\boldsymbol{X}^j, \boldsymbol{U}\right)$ of residuals versus that regressor.

Possibly omitted regressor $\boldsymbol{V}$

$\Rightarrow$ scatterplot $\left(\boldsymbol{V}, \boldsymbol{U}\right)$ of residuals versus that regressor.

For all proposed plots, a slightly better insight is obtained if standardized residuals $\boldsymbol{U}^{std}$ are used instead of the raw residuals $\boldsymbol{U}$.

4.3.2 (A2) Homoscedasticity of the errors

To detect

Residual variance that depends on the response expectation

$\Rightarrow$ scatterplot $\left(\widehat{\boldsymbol{Y}}, \boldsymbol{U}\right)$ of residuals versus fitted values.

Residual variance that depends on a particular regressor $\boldsymbol{X}^j$

$\Rightarrow$ scatterplot $\left(\boldsymbol{X}^j, \boldsymbol{U}\right)$ of residuals versus that regressor.

Residual variance that depend on a regressor $\boldsymbol{V}$ not included in the model

$\Rightarrow$ scatterplot $\left(\boldsymbol{V}, \boldsymbol{U}\right)$ of residuals versus that regressor.

For all proposed plots, a better insight is obtained if standardized residuals $\boldsymbol{U}^{std}$ are used instead of the raw residuals $\boldsymbol{U}$. This due to the fact that even if homoscedasticity of the errors is fulfilled, the raw residuals $\boldsymbol{U}$ are not necessarily homoscedastic ($\mathsf{var}\big(\boldsymbol{U} \,\big|\, \mathbb{Z}\big) = \sigma^2\,\mathbb{M}$), but the standardized residuals are homoscedastic having all a unity variance if additionally normality of the response holds.

So called *scale-location* plots are obtained, if on the above proposed plots, the vector of raw residuals $\boldsymbol{U}$ is replaced by a vector

$$\Big(\sqrt{\big|U_1^{std}\big|}, \ldots, \sqrt{\big|U_n^{std}\big|}\Big).$$

4.3.3 (A3) Uncorrelated errors

Assumption of uncorrelated errors is often justified by the used data gathering mechanism (e.g., observations/measurements performed on clearly independently behaving units/individuals). In that case, it does not make much sense to verify this assumption. Two typical situation when uncorrelated errors cannot be taken for granted are

(i) repeated observations performed on N independently behaving units/subjects;

(ii) observations performed sequentially in time where the ith response value Y_i is obtained in time t_i and the observational occasions $t_1 < \cdots < t_n$ form an increasing (and often equidistant) sequence.

In the following, we will not discuss any further the case (i) of repeated observations. In that case, a simple linear model is in most cases fully inappropriate for a statistical inference and more advanced models and methods must be used, see the course *Advanced Regression Models (NMST432)*. In case (ii), the errors $\varepsilon_1, \ldots, \varepsilon_n$ can often be considered as a *time series*[5]. The assumptions (A1)–(A3) of the linear model then states that this time series (the errors of the model) forms a *white noise*[6]. Possible *serial correlation (autocorrelation)* between the error terms is then usually considered as possible violation of the assumption (A3) of uncorrelated errors.

As stated above, even if the errors are uncorrelated and assumption (A3) is fulfilled, the residuals $\boldsymbol{U}$ are in general correlated. Nevertheless, the correlation is usually rather low and the residuals are typically used to check assumption (A3) and possibly to detect a form of the serial correlation present in data at hand. See *Stochastic Processes 2 (NMSA409)* course for basic diagnostic methods that include:

- Autocorrelation and partial autocorrelation plot based on residuals $\boldsymbol{U}$.
- Plot of delayed residuals, that is a scatterplot based on points (U_1, U_2), (U_2, U_3), …, (U_{n-1}, U_n).

4.3.4 (A4) Normality

To detect possible non-normality of the errors, standard tools used to check normality of a random sample known from the course *Mathematical Statistics 1 (NMSA331)* are used, now with the vector of residuals $\boldsymbol{U}$ or standardized residuals $\boldsymbol{U}^{std}$ in place of the random sample which normality is to be checked. A basic graphical tool to check the normality of a sample is then

- the normal probability plot (the QQ plot).

Usage of both the raw residuals $\boldsymbol{U}$ and the standardized residuals $\boldsymbol{U}^{std}$ to check the normality assumption (A4) bears certain inconveniences. If all assumptions of the normal linear model are fulfilled, then

[5] *časová řada* [6] *bílý šum*

The raw residuals $\boldsymbol{U}$ satisfy $\boldsymbol{U} \,|\, \mathbb{Z} \sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\, \mathbb{M}\big)$. That is, they maintain the normality, nevertheless, they are, in general, not homoscedastic ($\mathsf{var}\big(U_i \,\big|\, \mathbb{Z}\big) = \sigma^2\, m_{i,i},\ i = 1, \ldots, n$). Hence seeming non-normality of a "sample" $U_1, \ldots, U_n$ might be caused by the fact that the residuals are imposed to different variability.

The standardized residuals $\boldsymbol{U}^{std}$ satisfy $\mathbb{E}\big(U_i^{std} \,\big|\, \mathbb{Z}\big) = 0$, $\mathsf{var}\big(U_i^{std} \,\big|\, \mathbb{Z}\big) = 1$ for all $i = 1, \ldots, n$. That is, the standardized residuals are homoscedastic (with a known variance of one), nevertheless, they are not necessarily normally distributed. On the other hand, deviation of the distributional shape of the standardized residuals from the distributional shape of the errors $\boldsymbol{\varepsilon}$ is usually rather minor and hence the standardized residuals are usually useful in detecting non-normality of the errors.

Chapter 5

Submodels

In this chapter, we will again consider the original response-covariate data being represented by n random vectors $\left(Y_i,\, \boldsymbol{Z}_i^\top\right)^\top$, $\boldsymbol{Z}_i = \left(Z_{i,1},\, \ldots,\, Z_{i,p}\right)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$. The main aim is still to find a suitable model to express the (conditional) response expectation $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$, where $\mathbb{Z}$ is a matrix with vectors $\boldsymbol{Z}_1$, …, $\boldsymbol{Z}_n$ in its rows. Suppose that $\boldsymbol{t}_0 : \mathbb{R}^p \longrightarrow \mathbb{R}^{k_0}$ and $\boldsymbol{t} : \mathbb{R}^p \longrightarrow \mathbb{R}^k$ are two transformations of the covariates leading to the model matrices

$$\mathbb{X}^0 = \begin{pmatrix} \boldsymbol{X}_1^{0\top} \\ \vdots \\ \boldsymbol{X}_n^{0\top} \end{pmatrix}, \quad \begin{array}{rcl} \boldsymbol{X}_1^0 &=& \boldsymbol{t}_0(\boldsymbol{Z}_1), \\ & \vdots & \\ \boldsymbol{X}_n^0 &=& \boldsymbol{t}_0(\boldsymbol{Z}_n), \end{array} \qquad \mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}, \quad \begin{array}{rcl} \boldsymbol{X}_1 &=& \boldsymbol{t}(\boldsymbol{Z}_1), \\ & \vdots & \\ \boldsymbol{X}_n &=& \boldsymbol{t}(\boldsymbol{Z}_n). \end{array} \tag{5.1}$$

Briefly, we will write

$$\mathbb{X}^0 = \boldsymbol{t}_0(\mathbb{Z}), \qquad \mathbb{X} = \boldsymbol{t}(\mathbb{Z}).$$

Let (almost surely),

$$\mathsf{rank}(\mathbb{X}^0) = r_0, \qquad \mathsf{rank}(\mathbb{X}) = r, \tag{5.2}$$

where $0 < r_0 \leq k_0 < n$, $0 < r \leq k < n$. We will now deal with a situation when the matrices $\mathbb{X}^0$ and $\mathbb{X}$ determine two linear models:

$$\begin{aligned} &\text{Model } \mathsf{M}_0 : \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\, \mathbf{I}_n\right), \\ &\text{Model } \mathsf{M} : \ \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\right), \end{aligned}$$

and the task is to decide on whether one of the two models fits “better” the data. In this chapter, we limit ourselves to a situation when M_0 is so called *submodel* of the model M.

5.1 Submodel

Definition 5.1 Submodel.
We say that the model M_0 *is the* submodel[1] *(or the* nested model[2]*) of the model* M *if*

$$\mathcal{M}(\mathbb{X}^0) \subset \mathcal{M}(\mathbb{X}) \quad \textit{with } r_0 < r.$$

Notation. Situation that a model M_0 is a submodel of a model M will be denoted as

$$\mathsf{M}_0 \subset \mathsf{M}.$$

Notes.
- Submodel provides a more parsimonious expression of the response expectation $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z})$.
- The fact that the submodel M_0 holds means $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z}) \in \mathcal{M}(\mathbb{X}^0) \subset \mathcal{M}(\mathbb{X})$. That is, if the submodel M_0 holds then also the larger model M holds. That is, there exist $\boldsymbol{\beta}^0 \in \mathbb{R}^{k_0}$ and $\boldsymbol{\beta} \in \mathbb{R}^k$ such that
$$\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z}) = \mathbb{X}^0\boldsymbol{\beta}^0 = \mathbb{X}\boldsymbol{\beta}.$$
- The fact that the submodel M_0 does not hold but the model M holds means that $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z}) \in \mathcal{M}(\mathbb{X}) \setminus \mathcal{M}(\mathbb{X}^0)$. That is, there exist *no* $\boldsymbol{\beta}^0 \in \mathbb{R}^{k_0}$ such that $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z}) = \mathbb{X}^0\boldsymbol{\beta}^0$.

5.1.1 Projection considerations

Decomposition of the n-dimensional Euclidean space

Since $\mathcal{M}(\mathbb{X}^0) \subset \mathcal{M}(\mathbb{X}) \subset \mathbb{R}^n$, it is possible to construct an orthonormal vector basis

$$\mathbb{P}_{n\times n} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$$

of the n-dimensional Euclidean space as

$$\mathbb{P} = (\mathbb{Q}^0, \mathbb{Q}^1, \mathbb{N}),$$

where

- $\mathbb{Q}^0_{n\times r_0}$: orthonormal vector basis of the submodel regression space, i.e.,
$$\mathcal{M}(\mathbb{X}^0) = \mathcal{M}(\mathbb{Q}^0).$$
- $\mathbb{Q}^1_{n\times(r-r_0)}$: orthonormal vectors such that $\mathbb{Q} := (\mathbb{Q}^0, \mathbb{Q}^1)$ is an orthonormal vector basis of the model regression space, i.e.,
$$\mathcal{M}(\mathbb{X}) = \mathcal{M}(\mathbb{Q}) = \mathcal{M}\Big((\mathbb{Q}^0, \mathbb{Q}^1)\Big).$$

[1] *podmodel* [2] *vnořený model*

- $\mathbb{N}_{n\times(n-r)}$: orthonormal vector basis of the model residual space, i.e.,

$$\mathcal{M}(\mathbb{X})^{\perp} = \mathcal{M}(\mathbb{N}).$$

Further,

- $\mathbb{N}^0_{n\times(n-r_0)} := (\mathbb{Q}^1,\ \mathbb{N})$: orthonormal vector basis of the submodel residual space, i.e.,

$$\mathcal{M}(\mathbb{X}^0)^{\perp} = \mathcal{M}(\mathbb{N}^0) = \mathcal{M}\Big((\mathbb{Q}^1,\ \mathbb{N})\Big).$$

It follows from the orthonormality of columns of the matrix $\mathbb{P}$:

$$\begin{aligned}\mathbf{I}_n = \mathbb{P}^{\top}\mathbb{P} = \mathbb{P}\,\mathbb{P}^{\top} &= \mathbb{Q}^0\,{\mathbb{Q}^0}^{\top} + \mathbb{Q}^1\,{\mathbb{Q}^1}^{\top} + \mathbb{N}\,\mathbb{N}^{\top}\\ &= \mathbb{Q}\,\mathbb{Q}^{\top} + \mathbb{N}\,\mathbb{N}^{\top}\\ &= \mathbb{Q}^0\,{\mathbb{Q}^0}^{\top} + \mathbb{N}^0\,{\mathbb{N}^0}^{\top}.\end{aligned}$$

Notation. In the following, let

$$\begin{aligned}\mathbb{H}^0 &= \mathbb{Q}^0\,{\mathbb{Q}^0}^{\top},\\ \mathbb{M}^0 &= \mathbb{N}^0\,{\mathbb{N}^0}^{\top} = \mathbb{Q}^1\,{\mathbb{Q}^1}^{\top} + \mathbb{N}\,\mathbb{N}^{\top}.\end{aligned}$$

Notes.

- Matrices $\mathbb{H}^0$ and $\mathbb{M}^0$ which are symmetric and idempotent, are projection matrices into the regression and the residual space, respectively, of the submodel.
- The hat matrix and the residual projection matrix of the model can now also be written as

$$\begin{aligned}\mathbb{H} &= \mathbb{Q}\,\mathbb{Q}^{\top} = \mathbb{Q}^0\,{\mathbb{Q}^0}^{\top} + \mathbb{Q}^1\,{\mathbb{Q}^1}^{\top} = \mathbb{H}^0 + \mathbb{Q}^1\,{\mathbb{Q}^1}^{\top},\\ \mathbb{M} &= \mathbb{N}\,\mathbb{N}^{\top} = \mathbb{M}^0 - \mathbb{Q}^1\,{\mathbb{Q}^1}^{\top}.\end{aligned}$$

Projections into subspaces of the n-dimensional Euclidean space

Let $\boldsymbol{y} \in \mathbb{R}^n$. We can then write

$$\begin{aligned}\boldsymbol{y} = \mathbf{I}_n\,\boldsymbol{y} &= \big(\mathbb{Q}^0{\mathbb{Q}^0}^{\top} + \mathbb{Q}^1{\mathbb{Q}^1}^{\top} + \mathbb{N}\mathbb{N}^{\top}\big)\boldsymbol{y}\\ &= \underbrace{\mathbb{Q}^0{\mathbb{Q}^0}^{\top}\boldsymbol{y} + \mathbb{Q}^1{\mathbb{Q}^1}^{\top}\boldsymbol{y}}_{\widehat{\boldsymbol{y}}} + \underbrace{\mathbb{N}\mathbb{N}^{\top}\boldsymbol{y}}_{\boldsymbol{u}}\\ &= \underbrace{\mathbb{Q}^0{\mathbb{Q}^0}^{\top}\boldsymbol{y}}_{\widehat{\boldsymbol{y}}^0} + \underbrace{\mathbb{Q}^1{\mathbb{Q}^1}^{\top}\boldsymbol{y} + \mathbb{N}\mathbb{N}^{\top}\boldsymbol{y}}_{\boldsymbol{u}^0}.\end{aligned}$$

We have

- $\widehat{\boldsymbol{y}} = \big(\mathbb{Q}^0\,{\mathbb{Q}^0}^{\top} + \mathbb{Q}^1\,{\mathbb{Q}^1}^{\top}\big)\boldsymbol{y} = \mathbb{H}\boldsymbol{y} \in \mathcal{M}(\mathbb{X})$.

- $\boldsymbol{u} = \mathbb{N}\,\mathbb{N}^\top \boldsymbol{y} = \mathbb{M}\boldsymbol{y} \in \mathcal{M}(\mathbb{X})^\perp$.
- $\widehat{\boldsymbol{y}}^0 := \mathbb{Q}^0\,\mathbb{Q}^{0\top}\boldsymbol{y} = \mathbb{H}^0\boldsymbol{y} \in \mathcal{M}(\mathbb{X}^0)$.
- $\boldsymbol{u}^0 := \big(\mathbb{Q}^1\,\mathbb{Q}^{1\top} + \mathbb{N}\,\mathbb{N}^\top\big)\boldsymbol{y} = \mathbb{M}^0\boldsymbol{y} \in \mathcal{M}(\mathbb{X}^0)^\perp$.
- $\boldsymbol{d} := \mathbb{Q}^1\,\mathbb{Q}^{1\top}\boldsymbol{y} = \widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}^0 = \boldsymbol{u}^0 - \boldsymbol{u}$.

5.1.2 Properties of submodel related quantities

***Notation** (Quantities related to a submodel).*

When dealing with a pair of a model and a submodel, quantities related to the submodel will be denoted by a superscript (or by a subscript) 0. In particular:

- $\widehat{\boldsymbol{Y}}^0 = \mathbb{H}^0\boldsymbol{Y} = \mathbb{Q}^0\mathbb{Q}^{0\top}\boldsymbol{Y}$: fitted values in the submodel (projection of $\boldsymbol{Y}$ into the submodel regression space).
- $\boldsymbol{U}^0 = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0 = \mathbb{M}^0\boldsymbol{Y} = \big(\mathbb{Q}^1\mathbb{Q}^{1\top} + \mathbb{N}\mathbb{N}^\top\big)\boldsymbol{Y}$: residuals of the submodel.
- $\mathsf{SS}_e^0 = \big\|\boldsymbol{U}^0\big\|^2$: residual sum of squares of the submodel.
- $\nu_e^0 = n - r_0$: submodel residual degrees of freedom.
- $\mathsf{MS}_e^0 = \dfrac{\mathsf{SS}_e^0}{\nu_e^0}$: submodel residual mean square.

Additionally, as $\boldsymbol{D}$, we denote projection of the response vector $\boldsymbol{Y}$ into the space $\mathcal{M}(\mathbb{Q}^1)$, i.e.,

$$\boldsymbol{D} = \mathbb{Q}^1\,\mathbb{Q}^{1\top}\boldsymbol{Y} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0 = \boldsymbol{U}^0 - \boldsymbol{U}. \tag{5.3}$$

Theorem 5.1 On a submodel.

Consider two linear models $\mathsf{M}:\ \boldsymbol{Y}\,|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$ *and* $\mathsf{M}_0:\ \boldsymbol{Y}\,|\,\mathbb{Z} \sim \big(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\big)$ *such that* $\mathsf{M}_0 \subset \mathsf{M}$. *Let the submodel* M_0 *holds, i.e., let* $\mathbb{E}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big) \in \mathcal{M}(\mathbb{X}^0)$. *Then*

(i) $\widehat{\boldsymbol{Y}}^0$ *is the best linear unbiased estimator (BLUE) of a vector parameter* $\boldsymbol{\mu}^0 = \mathbb{X}^0\boldsymbol{\beta}^0 = \mathbb{E}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big)$.

(ii) The submodel residual mean square MS_e^0 *is the unbiased estimator of the residual variance* σ^2.

(iii) Statistics $\widehat{\boldsymbol{Y}}^0$ *and* $\boldsymbol{U}^0$ *are conditionally, given* $\mathbb{Z}$, *uncorrelated.*

(iv) A random vector $\boldsymbol{D} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0 = \boldsymbol{U}^0 - \boldsymbol{U}$ *satisfies*

$$\big\|\boldsymbol{D}\big\|^2 \;=\; \mathsf{SS}_e^0 - \mathsf{SS}_e.$$

(v) If additionally, a normal linear model is assumed, i.e., if $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\big)$ *then the statistics* $\widehat{\boldsymbol{Y}}^0$ *and* $\boldsymbol{U}^0$ *are conditionally, given* $\mathbb{Z}$, *independent and*

$$F_0 \;=\; \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{r - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}} \;=\; \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{\nu_e^0 - \nu_e}}{\dfrac{\mathsf{SS}_e}{\nu_e}} \;\sim\; \mathcal{F}_{r-r_0,\,n-r} = \mathcal{F}_{\nu_e^0-\nu_e,\,\nu_e}. \tag{5.4}$$

Proof. **Proof/calculations were available on the blackboard in KI.** ❑

End of Lecture #8 (26/10/2016)
Start of Lecture #10 (02/11/2016)

5.1.3 Series of submodels

When looking for a suitable model to express $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z})$, often a series of submodels is considered. Let us now assume a series of models

$$\begin{aligned}
\text{Model } \mathsf{M}_0 &: \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\right),\\
\text{Model } \mathsf{M}_1 &: \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}^1\boldsymbol{\beta}^1,\, \sigma^2\,\mathbf{I}_n\right),\\
\text{Model } \mathsf{M} &: \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\right),
\end{aligned}$$

where, analogously to (5.1), an $n \times k_1$ matrix $\mathbb{X}^1$ is given as

$$\mathbb{X}^1 = \begin{pmatrix} \boldsymbol{X}_1^{1\top} \\ \vdots \\ \boldsymbol{X}_n^{1\top} \end{pmatrix}, \qquad \begin{aligned} \boldsymbol{X}_1^1 &= \boldsymbol{t}_1(\boldsymbol{Z}_1),\\ &\vdots \\ \boldsymbol{X}_n^1 &= \boldsymbol{t}_1(\boldsymbol{Z}_n), \end{aligned}$$

for some transformation $\boldsymbol{t}_1 : \mathbb{R}^p \longrightarrow \mathbb{R}^{k_1}$ of the original covariates $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, which we briefly write as

$$\mathbb{X}^1 = \boldsymbol{t}_1(\mathbb{Z}).$$

Analogously to (5.2), we will assume that for some $0 < r_1 \leq k_1 < n$,

$$\mathsf{rank}(\mathbb{X}^1) = r_1.$$

Finally, we will assume that the three considered models are mutually submodels. That is, we will assume that

$$\mathcal{M}(\mathbb{X}^0) \subset \mathcal{M}(\mathbb{X}^1) \subset \mathcal{M}(\mathbb{X}) \qquad \text{with } r_0 < r_1 < r,$$

which we denote as

$$\mathsf{M}_0 \subset \mathsf{M}_1 \subset \mathsf{M}.$$

Notation. Quantities derived while assuming a particular model will be denoted by the corresponding superscript (or by no superscript in case of the model M). That is:

- $\widehat{\boldsymbol{Y}}^0$, $\boldsymbol{U}^0$, SS_e^0, ν_e^0, MS_e^0: quantities based on the (sub)model M_0: $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\mathbf{I}_n\right)$;
- $\widehat{\boldsymbol{Y}}^1$, $\boldsymbol{U}^1$, SS_e^1, ν_e^1, MS_e^1: quantities based on the (sub)model M_1: $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}^1\boldsymbol{\beta}^1,\, \sigma^2\mathbf{I}_n\right)$;
- $\widehat{\boldsymbol{Y}}$, $\boldsymbol{U}$, SS_e, ν_e, MS_e: quantities based on the model M: $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\right)$.

Theorem 5.2 On submodels.

Consider three normal *linear models* $\mathsf{M}: \ \boldsymbol{Y} \,|\, \mathbb{Z} \sim \mathcal{N}_n\left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\right)$, $\mathsf{M}_1: \ \boldsymbol{Y} \,|\, \mathbb{Z} \sim \mathcal{N}_n\left(\mathbb{X}^1\boldsymbol{\beta}^1,\, \sigma^2\,\mathbf{I}_n\right)$,

M_0 : $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\big)$ *such that* $\mathsf{M}_0 \subset \mathsf{M}_1 \subset \mathsf{M}$. *Let the (smallest) submodel* M_0 *hold, i.e., let* $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}^0\big)$. *Then*

$$F_{0,1} \;=\; \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e^1}{r_1 - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}} \;=\; \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e^1}{\nu_e^0 - \nu_e^1}}{\dfrac{\mathsf{SS}_e}{\nu_e}} \;\sim\; \mathcal{F}_{r_1-r_0,\,n-r} = \mathcal{F}_{\nu_e^0-\nu_e^1,\,\nu_e}. \tag{5.5}$$

Proof. **Proof/calculations were available on the blackboard in KI.** ❑

Note. Both F-statistics (5.4) and (5.5) contain

- In the numerator: a difference in the residual sums of squares of the two models where one of them is a submodel of the other divided by the difference of the residual degrees of freedom of those two models.
- In the denominator: a residual sum of squares of the model which is larger or equal to any of the two models whose quantities appear in the numerator, divided by the corresponding degrees of freedom.
- To obtain an F-distribution of the F-statistics (5.4) or (5.5), the smallest model whose quantities appear in that F-statistic must hold which implies that any other larger model holds as well.

Notation *(Differences when dealing with a submodel).*

Let M_A and M_B are two models distinguished by symbols "A" and "B" such that $\mathsf{M}_A \subset \mathsf{M}_B$. Let $\widehat{\boldsymbol{Y}}^A$ and $\widehat{\boldsymbol{Y}}^B$, $\boldsymbol{U}^A$ and $\boldsymbol{U}^B$, SS_e^A and SS_e^B denote the fitted values, the vectors of residuals and the residual sums of squares based on models M_A and M_B, respectively. The following notation will be used if it becomes necessary to indicate which are the two model related to the vector $\boldsymbol{D}$ or to the difference in the sums of squares:

$$\boldsymbol{D}\big(\mathsf{M}_B \,\big|\, \mathsf{M}_A\big) = \boldsymbol{D}\big(B \,\big|\, A\big) := \widehat{\boldsymbol{Y}}^B - \widehat{\boldsymbol{Y}}^A = \boldsymbol{U}^A - \boldsymbol{U}^B.$$
$$\mathsf{SS}\big(\mathsf{M}_B \,\big|\, \mathsf{M}_A\big) = \mathsf{SS}\big(B \,\big|\, A\big) := \mathsf{SS}_e^A - \mathsf{SS}_e^B.$$

Notes.

- Both F-statistics (5.4) and (5.5) contain certain $\mathsf{SS}\big(B \,\big|\, A\big)$ in their numerators.
- Point (iv) of Theorem 5.1 gives

$$\mathsf{SS}\big(B \,\big|\, A\big) = \Big\|\boldsymbol{D}\big(B \,\big|\, A\big)\Big\|^2.$$

5.1.4 Statistical test to compare nested models

Theorems 5.1 and 5.2 provide a way to compare two nested models by the mean of a statistical test.

F-test on a submodel based on Theorem 5.1

Consider two *normal* linear models: Model M_0: $\boldsymbol{Y} \mid \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\big)$,

Model M: $\boldsymbol{Y} \mid \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$,

where $\mathsf{M}_0 \subset \mathsf{M}$, and a set of statistical hypotheses: H_0: $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}^0\big)$

H_1: $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}\big) \setminus \mathcal{M}\big(\mathbb{X}^0\big)$,

that aim in answering the questions:

- Is model M significantly better than model M_0?
- Does the (larger) regression space $\mathcal{M}\big(\mathbb{X}\big)$ provide a significantly better expression for $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ over the (smaller) regression space $\mathcal{M}\big(\mathbb{X}^0\big)$?

The F-statistic (5.4) from Theorem 5.1 now provides a way to test the above hypotheses as follows:

Test statistic: $$F_0 \;=\; \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{r - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}} \;=\; \frac{\dfrac{\mathsf{SS}\big(\mathsf{M} \,\big|\, \mathsf{M}_0\big)}{r - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}}.$$

Reject H_0 if $F_0 \geq \mathcal{F}_{r-r_0,n-r}(1-\alpha)$.

P-value when $F_0 = f_0$: $p = 1 - \mathsf{CDF}_{\mathcal{F},\,r-r_0,n-r}\big(f_0\big)$.

F-test on a submodel based on Theorem 5.2

Consider three *normal* linear models: Model M_0: $\boldsymbol{Y} \mid \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\big)$,

Model M_1: $\boldsymbol{Y} \mid \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}^1\boldsymbol{\beta}^1,\, \sigma^2\,\mathbf{I}_n\big)$,

Model M: $\boldsymbol{Y} \mid \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$,

where $\mathsf{M}_0 \subset \mathsf{M}_1 \subset \mathsf{M}$, and a set of statistical hypotheses: H_0: $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}^0\big)$

H_1: $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}^1\big) \setminus \mathcal{M}\big(\mathbb{X}^0\big)$,

that aim in answering the questions:

- Is model M_1 significantly better than model M_0?
- Does the (larger) regression space $\mathcal{M}\big(\mathbb{X}^1\big)$ provide a significantly better expression for $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ over the (smaller) regression space $\mathcal{M}\big(\mathbb{X}^0\big)$?

The F-statistic (5.5) from Theorem 5.2 now provides a way to test the above hypotheses as follows:

Test statistic: $$F_{0,1} \;=\; \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e^1}{r_1 - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}} \;=\; \frac{\dfrac{\mathsf{SS}\big(\mathsf{M}_1 \,\big|\, \mathsf{M}_0\big)}{r_1 - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}}.$$

Reject H_0 if $F_{0,1} \geq \mathcal{F}_{r_1-r_0,n-r}(1-\alpha)$.

P-value when $F_{0,1} = f_{0,1}$: $p = 1 - \mathsf{CDF}_{\mathcal{F},\,r_1-r_0,n-r}\big(f_{0,1}\big)$.

5.2 Omitting some regressors

The most common couple (model – submodel) is Model M: $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$,
Submodel M_0: $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\mathbf{I}_n\big)$,

where the submodel matrix $\mathbb{X}^0$ is obtained by omitting selected columns from the model matrix $\mathbb{X}$. In other words, some regressors are omitted from the original regressor vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ to get the submodel and the matrix $\mathbb{X}^0$. In the following, without the loss of generality, let

$$\mathbb{X} \,=\, \big(\mathbb{X}^0,\, \mathbb{X}^1\big), \qquad 0 < \mathsf{rank}\big(\mathbb{X}^0\big) = r_0 < r = \mathsf{rank}\big(\mathbb{X}\big) < n.$$

The corresponding submodel F-test then evaluates whether, given the knowledge of the regressors included in the submodel matrix $\mathbb{X}^0$, the regressors included in the matrix $\mathbb{X}^1$ has an impact on the response expectation.

Theorem 5.3 Effect of omitting some regressors.

Consider a couple (model – submodel), where the submodel is obtained by omitting some regressors from the model. The following then holds.

(i) If $\mathcal{M}\big(\mathbb{X}^1\big) \perp \mathcal{M}\big(\mathbb{X}^0\big)$ *then*

$$\boldsymbol{D} = \mathbb{X}^1\big(\mathbb{X}^{1\top}\mathbb{X}^1\big)^{-}\mathbb{X}^{1\top}\boldsymbol{Y} =: \widehat{\boldsymbol{Y}}^1,$$

which are the fitted values from a linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}^1\boldsymbol{\beta}^1,\, \sigma^2\mathbf{I}_n\big)$.

(ii) If for given $\mathbb{Z}$*, the conditional distribution* $\boldsymbol{Y} \,\big|\, \mathbb{Z}$ *is continuous, i.e., has a density with respect to the Lebesgue measure on* $\big(\mathbb{R}^n,\, \mathcal{B}_n\big)$ *then*

$$\boldsymbol{D} \neq \mathbf{0}_n \quad \textit{and} \quad \mathsf{SS}_e^0 - \mathsf{SS}_e > 0 \quad \textit{almost surely.}$$

Proof. Let

$$\mathbb{M}^0 \;:=\; \mathbf{I}_n \;-\; \mathbb{X}^0\big(\mathbb{X}^{0\top}\mathbb{X}^0\big)^{-}\mathbb{X}^{0\top}$$

be the projection matrix into the residual space $\mathcal{M}\big(\mathbb{X}^0\big)^{\perp}$ of the submodel. We then have

$$\mathbb{M}^0\mathbb{X}^1 \;=\; \mathbb{X}^1 \;-\; \mathbb{X}^0\big(\mathbb{X}^{0\top}\mathbb{X}^0\big)^{-}\mathbb{X}^{0\top}\mathbb{X}^1.$$

Hence

$$\mathcal{M}\big(\mathbb{X}^0,\, \mathbb{X}^1\big) \;=\; \mathcal{M}\big(\mathbb{X}^0,\, \mathbb{M}^0\,\mathbb{X}^1\big)$$

since both spaces are generated by columns of matrices $\mathbb{X}^0$ and $\mathbb{X}^1$. Due to the fact that $\mathbb{M}^0$ is the projection matrix into $\mathcal{M}\big(\mathbb{X}^0\big)^{\perp}$, all columns of the matrix $\mathbb{X}^0$ are orthogonal to all columns of the matrix $\mathbb{M}^0\,\mathbb{X}^1$. In other words

$$\mathcal{M}\big(\mathbb{X}^0\big) \;\perp\; \mathcal{M}\big(\mathbb{M}^0\,\mathbb{X}^1\big).$$

Let

$$\mathbb{P} \;=\; \big(\mathbb{Q}^0,\, \mathbb{Q}^1,\, \mathbb{N}\big)$$

be a matrix with the orthonormal basis of $\mathbb{R}^n$ in its columns such that

- $\mathbb{Q}^0_{n\times r_0}$: orthonormal basis of the submodel regression space $\mathcal{M}(\mathbb{X}^0)$, i.e.,
$$\mathcal{M}(\mathbb{X}^0) \;=\; \mathcal{M}(\mathbb{Q}^0).$$
- $\mathbb{Q}^1_{n\times(r-r_0)}$: orthonormal vectors such that $(\mathbb{Q}^0, \mathbb{Q}^1)$ is the orthonormal basis of the model regression space $\mathcal{M}(\mathbb{X}^0, \mathbb{X}^1)$, i.e.,
$$\mathcal{M}(\mathbb{X}^0, \mathbb{X}^1) \;=\; \mathcal{M}(\mathbb{Q}^0, \mathbb{Q}^1).$$
- $\mathbb{N}_{n\times n-r}$: orthonormal basis of the model residual space $\mathcal{M}(\mathbb{X}^0, \mathbb{X}^1)^\perp$.

Since $\mathcal{M}(\mathbb{X}^0, \mathbb{X}^1) \;=\; \mathcal{M}(\mathbb{X}^0, \mathbb{M}^0\,\mathbb{X}^1)$ and $\mathcal{M}(\mathbb{X}^0) \;\perp\; \mathcal{M}(\mathbb{M}^0\,\mathbb{X}^1)$, we also have that
$$\mathcal{M}(\mathbb{Q}^1) \;=\; \mathcal{M}(\mathbb{M}^0\,\mathbb{X}^1).$$

Vector $\boldsymbol{D}$ is a projection of the response vector $\boldsymbol{Y}$ into the space $\mathcal{M}(\mathbb{Q}^1) \;=\; \mathcal{M}(\mathbb{M}^0\,\mathbb{X}^1)$. The corresponding projection matrix, let say $\mathbb{H}^1$ can be calculated as (use Lemma 2.1 with $\mathbb{X} = \mathbb{M}^0\,\mathbb{X}^1$)
$$\mathbb{H}^1 \;=\; (\mathbb{M}^0\,\mathbb{X}^1)\,(\mathbb{X}^{1\top}\underbrace{\mathbb{M}^0\mathbb{M}^0}_{\mathbb{M}^0}\mathbb{X}^1)^-\,\mathbb{X}^{1\top}\mathbb{M}^0.$$

Then
$$\boldsymbol{D} \;=\; \mathbb{H}^1\,\boldsymbol{Y} \;=\; (\mathbb{M}^0\,\mathbb{X}^1)\,(\mathbb{X}^{1\top}\mathbb{M}^0\mathbb{X}^1)^-\,\mathbb{X}^{1\top}\underbrace{\mathbb{M}^0\boldsymbol{Y}}_{\boldsymbol{U}^0}. \tag{5.6}$$

That is,
$$\boldsymbol{D} \;=\; (\mathbb{M}^0\,\mathbb{X}^1)\,(\mathbb{X}^{1\top}\mathbb{M}^0\mathbb{X}^1)^-\,\mathbb{X}^{1\top}\boldsymbol{U}^0,$$
where $\boldsymbol{U}^0 = \mathbb{M}^0\boldsymbol{Y}$ are residuals of the submodel.

(i) If $\mathcal{M}(\mathbb{X}^1) \perp \mathcal{M}(\mathbb{X}^0)$, we have $\mathbb{M}^0\mathbb{X}^1 \;=\; \mathbb{X}^1$. Consequently,
$$\mathbb{X}^{1\top}\boldsymbol{U}^0 \;=\; \mathbb{X}^{1\top}\mathbb{M}^0\mathbb{X}^1 \;=\; \mathbb{X}^{1\top}\boldsymbol{Y}$$
and by (5.6), while realizing $\mathbb{M}^0 = \mathbb{M}^0\,\mathbb{M}^0$, we get
$$\boldsymbol{D} \;=\; \mathbb{X}^1\,(\mathbb{X}^{1\top}\mathbb{X}^1)^-\mathbb{X}^{1\top}\boldsymbol{Y}.$$

(ii) The vector $\boldsymbol{D}$, as a projection of vector $\boldsymbol{Y}$ into the vector space $\mathcal{M}(\mathbb{Q}^1) \;=\; \mathcal{M}(\mathbb{M}^0\mathbb{X}^1)$ (subspace of $\mathbb{R}^n$ of vector dimension $r - r_0$) is equal to the zero vector if and only if
$$\boldsymbol{Y} \;\in\; \mathcal{M}(\mathbb{Q}^1)^\perp,$$
where $\mathcal{M}(\mathbb{Q}^1)^\perp$ is a vector subspace of $\mathbb{R}^n$ of vector dimension $n - r + r_0 < n$. Hence, under our assumption of a continuous conditional distribution $\boldsymbol{Y}\,|\,\mathbb{Z}$,
$$\mathsf{P}(\boldsymbol{Y} \in \mathcal{M}(\mathbb{Q}^1)^\perp\,|\,\mathbb{Z}) = 0,$$
that is, $\boldsymbol{D} \neq \boldsymbol{0}_n$ almost surely.

Consequently, $\mathsf{SS}^0_e - \mathsf{SS}_e \;=\; \|\boldsymbol{D}\|^2 > 0$ almost surely.

❑

Note. If we take the residual sum of squares as a measure of a quality of the model, point (ii) of Theorem 5.3 says that the model is almost surely getting worse if some regressors are removed. Nevertheless, in practice, it is always a question whether this worsening is statistically significant (the submodel F-test answers this) or practically important (additional reasoning is needed).

5.3 Linear constraints

Suppose that a linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r$ is given and it is our aim to verify whether the response expectation $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ lies in a constrained regression space

$$\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big) := \big\{\boldsymbol{v} :\, \boldsymbol{v} = \mathbb{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^k,\ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big\}, \tag{5.7}$$

where $\mathbb{L}_{m\times k}$ is a given real matrix and $\boldsymbol{\theta}^0 \in \mathbb{R}^m$ is a given vector. In other words, verification of whether the response expectation lies in the space $\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big)$ corresponds to verification of whether the regression coefficients satisfy a linear constraint $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$.

Lemma 5.4 Regression space given by linear constraints.

Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \le k < n$. *Let* $\mathbb{L}_{m\times k}$ *be a real matrix with* $m \le r$ *rows such that*

(i) $\mathsf{rank}\big(\mathbb{L}\big) = m$ *(i.e.,* $\mathbb{L}$ *is a matrix with linearly independent rows);*

(ii) $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ *is* estimable *parameter of the considered linear model.*

The space $\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \mathbf{0}_m\big)$ *is then a* vector subspace *of dimension* $r - m$ *of the regression space* $\mathcal{M}\big(\mathbb{X}\big)$.

Proof. **Proof/calculations were available on the blackboard in Kl.** ❑

Notes.

- The space $\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big)$ is a vector space only if $\boldsymbol{\theta}^0 = \mathbf{0}_m$ since otherwise, $\mathbf{0}_n \notin \mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big)$. Nevertheless, for the purpose of the statistical analysis, it is possible (and in practice also necessary) to work also with $\boldsymbol{\theta}^0 \neq \mathbf{0}_m$.
- With $m = r$, $\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \mathbf{0}_m\big) = \big\{\mathbf{0}_n\big\}$.

Definition 5.2 Submodel given by linear constraints.

We say that the model M_0 *is a* submodel given by linear constraints[3] $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$ *of model* M*:* $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r$, *if matrix* $\mathbb{L}$ *satisfies conditions of Lemma 5.4,* $m < r$ *and the response expectation* $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ *under the model* M_0 *is assumed to lie in a space* $\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big)$.

Notation. A submodel given by linear constraints will be denoted as

$$\mathsf{M}_0 : \boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0.$$

[3] *podmodel zadaný lineárními omezeními*

Since with $\boldsymbol{\theta}^0 \neq \mathbf{0}_m$, the space $\mathcal{M}(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \mathbf{0}_m)$ is not a vector space, we in general cannot talk about projections in a sense of linear algebra when deriving the fitted values, the residuals and other quantities related to the submodel given by linear constraints. Hence we introduce the following definition.

Definition 5.3 Fitted values, residuals, residual sum of squares, rank of the model and residual degrees of freedom in a submodel given by linear constraints.

Let $\boldsymbol{b}^0 \in \mathbb{R}^k$ minimize $\mathsf{SS}(\boldsymbol{\beta}) = \left\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right\|^2$ over $\boldsymbol{\beta} \in \mathbb{R}^k$ subject to $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$. For the submodel $\mathsf{M}_0 : \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\right),\ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$, the following quantities are defined as follows:

Fitted values: $\widehat{\boldsymbol{Y}}^0 := \mathbb{X}\boldsymbol{b}^0$.

Residuals: $\boldsymbol{U}^0 := \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0$.

Residual sum of squares: $\mathsf{SS}_e^0 := \left\|\boldsymbol{U}^0\right\|^2$.

Rank of the model: $r_0 = r - m$.

Residual degrees of freedom: $\nu_e^0 := n - r_0$.

Note. The fitted values could also be defined as

$$\widehat{\boldsymbol{Y}}^0 = \underset{\widetilde{\boldsymbol{Y}} \in \mathcal{M}\left(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\right)}{\operatorname{argmin}} \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}\right\|^2.$$

That is, the fitted values are (still) the closest point to $\boldsymbol{Y}$ in the constrained regression space $\mathcal{M}(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0)$.

End of Lecture #10 (02/11/2016)
Start of Lecture #12 (10/11/2016)

Theorem 5.5 On a submodel given by linear constraints.

Let $\mathsf{M}_0 : \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\right),\ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$ be a submodel given by linear constraints of a model $\mathsf{M} : \boldsymbol{Y} \,|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\right)$. Then

(i) The fitted values $\widehat{\boldsymbol{Y}}^0$ and consequently also the residuals $\boldsymbol{U}^0$ and the residual sum of squares SS_e^0 are unique.

(ii) $\boldsymbol{b}^0$ minimizes $\mathsf{SS}(\boldsymbol{\beta}) = \left\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right\|^2$ subject to $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$ if and only if

$$\boldsymbol{b}^0 = \boldsymbol{b} - \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right),$$

where $\boldsymbol{b} = \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{X}^\top\boldsymbol{Y}$ is (any) solution to a system of normal equations $\mathbb{X}^\top\mathbb{X}\boldsymbol{b} = \mathbb{X}^\top\boldsymbol{Y}$.

(iii) The fitted values $\widehat{\boldsymbol{Y}}^0$ can be expressed as

$$\widehat{\boldsymbol{Y}}^0 = \widehat{\boldsymbol{Y}} - \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right).$$

(iv) The vector $\boldsymbol{D} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0$ satisfies

$$\left\|\boldsymbol{D}\right\|^2 = \mathsf{SS}_e^0 - \mathsf{SS}_e = \left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right). \tag{5.8}$$

Proof. First mention that under our assumptions, the matrix $\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{L}^\top$ is

(i) invertible;

(ii) does not depend on a choice of the pseudoinverse $\big(\mathbb{X}^\top\mathbb{X}\big)^-$.

This follows from Theorem 2.9 (Gauss–Markov for estimable vector parameter).

Second, try to look for $\widehat{\boldsymbol{Y}}^0 = \mathbb{X}\boldsymbol{b}^0$ such that $\boldsymbol{b}^0$ minimizes $\mathsf{SS}(\boldsymbol{\beta}) = \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big\|^2$ over $\boldsymbol{\beta} \in \mathbb{R}^k$ subject to $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$ by a method of Lagrange multipliers. Let

$$\begin{aligned}\varphi(\boldsymbol{\beta},\,\boldsymbol{\lambda}) &= \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big\|^2 \,+\, 2\boldsymbol{\lambda}^\top\big(\mathbb{L}\boldsymbol{\beta} - \boldsymbol{\theta}^0\big)\\ &= \big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big)^\top\big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big) \,+\, 2\boldsymbol{\lambda}^\top\big(\mathbb{L}\boldsymbol{\beta} - \boldsymbol{\theta}^0\big),\end{aligned}$$

where a factor of 2 in the second part of expression of the Lagrange function φ is only included to simplify subsequent expressions.

The first derivatives of φ are as follows:

$$\begin{aligned}\frac{\partial\varphi}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta},\,\boldsymbol{\lambda}) &= -2\,\mathbb{X}^\top\big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big) \,+\, 2\,\mathbb{L}^\top\boldsymbol{\lambda},\\ \frac{\partial\varphi}{\partial\boldsymbol{\lambda}}(\boldsymbol{\beta},\,\boldsymbol{\lambda}) &= 2\,\big(\mathbb{L}\boldsymbol{\beta} - \boldsymbol{\theta}^0\big).\end{aligned}$$

Realize now that $\dfrac{\partial\varphi}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta},\,\boldsymbol{\lambda}) = \mathbf{0}_k$ if and only if

$$\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta} \,=\, \mathbb{X}^\top\boldsymbol{Y} \,-\, \mathbb{L}^\top\boldsymbol{\lambda}. \tag{5.9}$$

Note that the linear system (5.9) is consistent for any $\boldsymbol{\lambda} \in \mathbb{R}^m$ and any $\boldsymbol{Y} \in \mathbb{R}^n$. This follows from the fact that due to estimability of a parameter $\mathbb{L}\boldsymbol{\beta}$, we have $\mathcal{M}\big(\mathbb{L}^\top\big) \subset \mathcal{M}\big(\mathbb{X}^\top\big)$ (Theorem 2.7). Hence the right-hand-side of the system (5.9) lies in $\mathcal{M}\big(\mathbb{X}^\top\big)$, for any $\boldsymbol{\lambda} \in \mathbb{R}^m$ and any $\boldsymbol{Y} \in \mathbb{R}^n$. The left-hand-side of the system (5.9) lies in $\mathcal{M}\big(\mathbb{X}^\top\mathbb{X}\big)$, for any $\boldsymbol{\beta} \in \mathbb{R}^k$. We already know that $\mathcal{M}\big(\mathbb{X}^\top\big) = \mathcal{M}\big(\mathbb{X}^\top\mathbb{X}\big)$ (Lemma 2.6) which proves that there always exist a solution to the linear system (5.9).

Let $\boldsymbol{b}^0(\boldsymbol{\lambda})$ be any solution to $\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta} \,=\, \mathbb{X}^\top\boldsymbol{Y} \,-\, \mathbb{L}^\top\boldsymbol{\lambda}$. That is,

$$\begin{aligned}\boldsymbol{b}^0(\boldsymbol{\lambda}) &= \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\boldsymbol{Y} \,-\, \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{L}^\top\boldsymbol{\lambda}\\ &= \boldsymbol{b} \,-\, \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{L}^\top\boldsymbol{\lambda},\end{aligned}$$

which depends on a choice of $\big(\mathbb{X}^\top\mathbb{X}\big)^-$.

Further, $\dfrac{\partial\varphi}{\partial\boldsymbol{\lambda}}(\boldsymbol{\beta},\,\boldsymbol{\lambda}) = \mathbf{0}_m$ if and only if

$$\begin{aligned}\mathbb{L}\boldsymbol{b}^0(\boldsymbol{\lambda}) &= \boldsymbol{\theta}^0\\ \mathbb{L}\boldsymbol{b} \,-\, \mathbb{L}\,\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{L}^\top\boldsymbol{\lambda} &= \boldsymbol{\theta}^0\\ \underbrace{\mathbb{L}\,\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{L}^\top}\boldsymbol{\lambda} &= \mathbb{L}\boldsymbol{b} \,-\, \boldsymbol{\theta}^0.\\ \text{invertible as we already know}\end{aligned}$$

That is,

$$\boldsymbol{\lambda} = \left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right).$$

Finally,

$$\boldsymbol{b}^0 = \boldsymbol{b} - \left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right),$$

$$\widehat{\boldsymbol{Y}}^0 = \mathbb{X}\boldsymbol{b}^0 = \widehat{\boldsymbol{Y}} - \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right).$$

Realize again that $\mathcal{M}\left(\mathbb{L}^\top\right) \subset \mathcal{M}\left(\mathbb{X}^\top\right)$. That is, there exist a matrix $\mathbb{A}$ such that

$$\mathbb{L}^\top = \mathbb{X}^\top\mathbb{A}^\top, \qquad \mathbb{L} = \mathbb{A}\mathbb{X}.$$

Under our assumptions, matrix $\mathbb{A}$ is even unique. The vector $\widehat{\boldsymbol{Y}}^0$ can now be written as

$$\widehat{\boldsymbol{Y}}^0 = \underbrace{\widehat{\boldsymbol{Y}}}_{\text{unique}} - \underbrace{\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top}_{\text{unique}}\underbrace{\mathbb{A}^\top}_{\text{unique}}\underbrace{\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}}_{\text{unique}}\left(\underbrace{\mathbb{L}\boldsymbol{b}}_{\text{unique}} - \boldsymbol{\theta}^0\right). \tag{5.10}$$

To show point (iv), use (5.10) in expressing the vector $\boldsymbol{D} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0$:

$$\boldsymbol{D} = \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top\mathbb{A}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right).$$

That is,

$$\left\|\boldsymbol{D}\right\|^2 = \left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\mathbb{A}\underbrace{\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top\mathbb{X}}_{\mathbb{X}\text{ by the five matrices rule}}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top\mathbb{A}^\top$$
$$\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)$$

$$= \left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\mathbb{A}\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{X}^\top\mathbb{A}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)$$

$$= \left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)$$

$$= \left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right).$$

It remains to be shown that $\left\|\boldsymbol{D}\right\|^2 = \mathsf{SS}_e^0 - \mathsf{SS}_e$. We have

$$\mathsf{SS}_e^0 = \left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0\right\|^2 = \Big\|\underbrace{\boldsymbol{Y} - \widehat{\boldsymbol{Y}}}_{\boldsymbol{U} \in \mathcal{M}(\mathbb{X})^\perp} + \underbrace{\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\left\{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0\right)}_{\boldsymbol{D} \in \mathcal{M}(\mathbb{X})}\Big\|^2$$

$$= \left\|\boldsymbol{U}\right\|^2 + \left\|\boldsymbol{D}\right\|^2 = \mathsf{SS}_e + \left\|\boldsymbol{D}\right\|^2.$$

❑

5.3.1 F-statistic to verify a set of linear constraints

Let us take the expression (5.8) for the difference between the residual sums of squares of the model and the submodel given by linear constraints and derive the submodel F-statistic (5.4):

$$F_0 = \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{r - r_0}}{\dfrac{\mathsf{SS}_e}{n-r}} = \frac{\dfrac{(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0)^\top \left\{\mathbb{L}(\mathbb{X}^\top\mathbb{X})^-\mathbb{L}^\top\right\}^{-1}(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0)}{m}}{\dfrac{\mathsf{SS}_e}{n-r}}$$

$$= \frac{1}{m}(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0)^\top \left\{\mathsf{MS}_e\,\mathbb{L}(\mathbb{X}^\top\mathbb{X})^-\mathbb{L}^\top\right\}^{-1}(\mathbb{L}\boldsymbol{b} - \boldsymbol{\theta}^0)$$

$$= \frac{1}{m}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^\top \left\{\mathsf{MS}_e\,\mathbb{L}(\mathbb{X}^\top\mathbb{X})^-\mathbb{L}^\top\right\}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0), \tag{5.11}$$

where $\widehat{\boldsymbol{\theta}} = \mathbb{L}\boldsymbol{b}$ is the LSE of the estimable vector parameter $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ in the linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$ without constraints. Note now that (5.11) is exactly equal to the Wald-type statistic Q_0 (see page 41) that we used in Section 3.2.2 to test the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ on an estimable vector parameter $\boldsymbol{\theta}$ in a *normal* linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$. If normality can be assumed, point (x) of Theorem 3.2 then provided that under the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, that is, under the validity of the submodel given by linear constraints $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$, the statistic F_0 follows the usual F-distribution $\mathcal{F}_{m,n-r}$. This shows that the Wald-type test on the estimable vector parameter in a normal linear model based on Theorem 3.2 is equivalent to the submodel F-test based on Theorem 5.1.

5.3.2 t-statistic to verify a linear constraint

Consider $\mathbb{L} = \mathbf{l}^\top$, $\mathbf{l} \in \mathbb{R}^k$, $\mathbf{l} \neq \mathbf{0}_k$ such that $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ is an estimable parameter of the normal linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$. Take $\theta^0 \in \mathbb{R}$ and consider the submodel given by $m = 1$ linear constraint $\mathbf{l}^\top\boldsymbol{\beta} = \theta^0$. Let $\widehat{\theta} = \mathbf{l}^\top\boldsymbol{b}$, where $\boldsymbol{b}$ is any solution to the normal equations in the model without constraints. The statistic (5.11) then takes the form

$$F_0 = \frac{1}{m}(\widehat{\theta} - \theta^0)\left\{\mathsf{MS}_e\,\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}\right\}^{-1}(\widehat{\theta} - \theta^0) = \left(\frac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\,\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}}}\right)^2 = T_0^2,$$

where

$$T_0 = \frac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\,\mathbf{l}^\top(\mathbb{X}^\top\mathbb{X})^-\mathbf{l}}}$$

is the Wald-type test statistic introduced in Section 3.2.2 (on page 40) to test the null hypothesis H_0: $\theta = \theta^0$ in a *normal* linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n)$. Point (viii) of Theorem 3.2 provided that under the null hypothesis H_0: $\theta = \theta^0$, the statistic T_0 follows the Student t-distribution t_{n-r} which is indeed in agreement with the fact that $T_0^2 = F_0$ follows the F-distribution $\mathcal{F}_{1,n-r}$.

5.4 Coefficient of determination

5.4.1 Intercept only model

Notation *(Response sample mean).*

The sample mean over the response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ will be denoted as $\overline{Y}$. That is,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\boldsymbol{Y}^\top \mathbf{1}_n.$$

Definition 5.4 Regression and total sums of squares in a linear model.

Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim (\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *The following expressions define the following quantities:*

(i) Regression sum of squares[4] *and corresponding degrees of freedom:*

$$\mathsf{SS}_R = \left\|\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}_n\right\|^2 = \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2, \quad \nu_R = r - 1,$$

(ii) Total sum of squares[5] *and corresponding degrees of freedom:*

$$\mathsf{SS}_T = \left\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\right\|^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2, \quad \nu_T = n - 1.$$

Lemma 5.6 Model with intercept only.

Let $\boldsymbol{Y} \sim (\mathbf{1}_n\gamma, \zeta^2\mathbf{I}_n)$. *Then*

(i) $\widehat{\boldsymbol{Y}} = \overline{Y}\mathbf{1}_n = \left(\overline{Y}, \ldots, \overline{Y}\right)^\top$.

(ii) $\mathsf{SS}_e = \mathsf{SS}_T$.

Proof. This is a full-rank model with $\mathbb{X} = \mathbf{1}_n$. Further,

$$\left(\mathbb{X}^\top\mathbb{X}\right)^{-1} = \left(\mathbf{1}_n^\top\mathbf{1}_n\right)^{-1} = \frac{1}{n}, \qquad \mathbb{X}^\top\boldsymbol{Y} = \mathbf{1}_n^\top\boldsymbol{Y} = \sum_{i=1}^{n} Y_i.$$

Hence $\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$ and $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\gamma} = \mathbf{1}_n\overline{Y} = \overline{Y}\mathbf{1}_n$.

❑

---

[4] *regresní součet čtverců* [5] *celkový součet čtverců*

5.4.2 Models with intercept

Lemma 5.7 Identity in a linear model with intercept.
Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *where* $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$. *Then*

$$\mathbf{1}_n^\top \boldsymbol{Y} = \sum_{i=1}^n Y_i = \sum_{i=1}^n \widehat{Y}_i = \mathbf{1}_n^\top \widehat{\boldsymbol{Y}}.$$

Proof.

- Follows directly from the normal equations if $\mathbf{1}_n$ is one of the columns of $\mathbb{X}$ matrix.
- General proof:
$$\mathbf{1}_n^\top \widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}^\top \mathbf{1}_n = \big(\mathbb{H}\boldsymbol{Y}\big)^\top \mathbf{1}_n = \boldsymbol{Y}^\top \mathbb{H}\mathbf{1}_n = \boldsymbol{Y}^\top \mathbf{1}_n,$$
since $\mathbb{H}\mathbf{1}_n = \mathbf{1}_n$ due to the fact that $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$.

❑

Theorem 5.8 Breakdown of the total sum of squares in a linear model with intercept.
Let $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *where* $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$. *Then*

$$\begin{array}{ccccc} \mathsf{SS}_T & = & \mathsf{SS}_e & + & \mathsf{SS}_R \\ \sum_{i=1}^n \big(Y_i - \overline{Y}\big)^2 & = & \sum_{i=1}^n \big(Y_i - \widehat{Y}_i\big)^2 & + & \sum_{i=1}^n \big(\widehat{Y}_i - \overline{Y}\big)^2. \end{array}$$

Proof. The identity $\mathsf{SS}_T = \mathsf{SS}_e + \mathsf{SS}_R$ follows trivially if $r = \mathsf{rank}\big(\mathbb{X}\big) = 1$ since then $\mathcal{M}\big(\mathbb{X}\big) = \mathcal{M}\big(\mathbf{1}_n\big)$ and hence (by Lemma 5.6) $\widehat{\boldsymbol{Y}} = \overline{Y}\mathbf{1}_n$. Then $\mathsf{SS}_T = \mathsf{SS}_e$, $\mathsf{SS}_R = 0$.

In the following, let $r = \mathsf{rank}\big(\mathbb{X}\big) > 1$. Then, model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbf{1}_n\beta^0,\, \sigma^2\mathbf{I}_n\big)$ is a submodel of the model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ and by Lemma 5.6, $\mathsf{SS}_T = \mathsf{SS}_e^0$. Further, from definition of SS_R, it equals to $\mathsf{SS}_R = \big\|\boldsymbol{D}\big\|^2$, where $\boldsymbol{D} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0$. By point (iv) of Theorem 5.1 (on a submodel), $\big\|\boldsymbol{D}\big\|^2 = \mathsf{SS}_e^0 - \mathsf{SS}_e$. In other words,

$$\mathsf{SS}_R = \mathsf{SS}_T - \mathsf{SS}_e.$$

❑

The identity $\mathsf{SS}_T = \mathsf{SS}_e + \mathsf{SS}_R$ can also be shown directly while using a little algebra. We have

$$\begin{aligned}
\mathsf{SS}_T &= \sum_{i=1}^{n} \big(Y_i - \overline{Y}\big)^2 = \sum_{i=1}^{n} \big(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y}\big)^2 \\
&= \sum_{i=1}^{n} \big(Y_i - \widehat{Y}_i\big)^2 + \sum_{i=1}^{n} \big(\widehat{Y}_i - \overline{Y}\big)^2 + 2 \sum_{i=1}^{n} \big(Y_i - \widehat{Y}_i\big)\big(\widehat{Y}_i - \overline{Y}\big) \\
&= \mathsf{SS}_e + \mathsf{SS}_R + \underbrace{2 \Big\{ \sum_{i=1}^{n} Y_i \widehat{Y}_i - \overline{Y} \sum_{i=1}^{n} Y_i + \overline{Y} \sum_{i=1}^{n} \widehat{Y}_i - \sum_{i=1}^{n} \widehat{Y}_i^2 \Big\}}_{0} \\
&= \mathsf{SS}_e + \mathsf{SS}_R
\end{aligned}$$

since $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$ and additionally

$$\sum_{i=1}^{n} Y_i \widehat{Y}_i = \boldsymbol{Y}^\top \widehat{\boldsymbol{Y}} = \boldsymbol{Y}^\top \mathbb{H} \boldsymbol{Y}, \qquad \sum_{i=1}^{n} \widehat{Y}_i^2 = \widehat{\boldsymbol{Y}}^\top \widehat{\boldsymbol{Y}} = \boldsymbol{Y}^\top \mathbb{H}\mathbb{H}\mathbb{H} \boldsymbol{Y} = \boldsymbol{Y}^\top \mathbb{H} \boldsymbol{Y}.$$

❑

5.4.3 Theoretical evaluation of a prediction quality of the model

One of the usual aims of regression modelling is so called prediction in which case the model based response mean is used as the predicted response value. In such situations, it is assumed that data $\big(Y_i,\, \boldsymbol{X}_i^\top\big)^\top$, $i = 1, \ldots, n$, are a random sample from some joint distribution of a generic random vector $\big(Y,\, \boldsymbol{X}^\top\big)^\top$, $\boldsymbol{X} = \big(X_0,\, \ldots,\, X_{k-1}\big)^\top$ and the conditional distribution $Y \,|\, \boldsymbol{X}$ can be described by a linear model, i.e.,

$$\mathbb{E}\big(Y \,\big|\, \boldsymbol{X}\big) = \boldsymbol{X}^\top \boldsymbol{\beta}, \qquad \mathsf{var}\big(Y \,\big|\, \boldsymbol{X}\big) = \sigma^2 \tag{5.12}$$

for some $\boldsymbol{\beta} = \big(\beta_0,\, \ldots,\, \beta_{k-1}\big)^\top \in \mathbb{R}^k$ and some $\sigma^2 > 0$, which leads to the linear model

$$\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big), \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \; \mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}$$

for the data. As usually, we assume $\mathsf{rank}\big(\mathbb{X}\big) = r \leq k < n$ (almost surely).

In the following, let $\gamma \in \mathbb{R}$ and $\zeta^2 > 0$ be the *marginal* mean and the variance, respectively, of the response random variable Y, i.e.,

$$\mathbb{E}\big(Y\big) = \gamma, \qquad \mathsf{var}\big(Y\big) = \zeta^2. \tag{5.13}$$

This corresponds to the only intercept linear model

$$\boldsymbol{Y} \sim \big(\mathbf{1}_n \gamma,\, \zeta^2 \mathbf{I}_n\big)$$

for the data with a model matrix $\mathbf{1}_n$ of rank 1.

Suppose now that all model parameters $(\boldsymbol{\beta},\ \gamma,\ \sigma^2,\ \zeta^2)$ related to the distribution of the random vector $\left(Y,\ \boldsymbol{X}^\top\right)^\top$ are known and the aim is to provide the prediction $\widehat{Y}$ of the response value Y. We could also say that we want to predict the Y-component of a not yet observed ("new") random vector $\left(Y_{new},\ \boldsymbol{X}_{new}^\top\right)^\top$ which is distributed as the generic vector $\left(Y,\ \boldsymbol{X}^\top\right)^\top$. Nevertheless, for simplicity of notation, we will not use the subscript $_{new}$ and will simply work with the random vector $\left(Y,\ \boldsymbol{X}^\top\right)^\top$ whose distribution satisfies (5.12) and (5.13).

Suppose further that the random vector $\left(Y,\ \boldsymbol{X}^\top\right)^\top$ is defined on a probability space $\left(\Omega,\ \mathcal{A},\ \mathsf{P}\right)$ and let $\sigma(\boldsymbol{X}) \subseteq \mathcal{A}$ be a σ-algebra generated by the random vector $\boldsymbol{X}$, $\mathsf{P}|_{\sigma(\boldsymbol{X})}$ be a probability measure restricted to this σ-algebra and $\mathsf{L}_2(\boldsymbol{X}) = \mathsf{L}_2\left(\Omega,\ \sigma(\boldsymbol{X}),\ \mathsf{P}|_{\sigma(\boldsymbol{X})}\right)$. Further, let $\sigma(\emptyset) = \left\{\emptyset,\ \Omega\right\}$ be a trivial σ-algebra on Ω, $\mathsf{P}|_{\sigma(\emptyset)}$ the related restricted probability measure and $\mathsf{L}_2(\emptyset) = \mathsf{L}_2\left(\Omega,\ \sigma(\emptyset),\ \mathsf{P}|_{\sigma(\emptyset)}\right)$.

A problem of prediction of a value of the random variable $Y \in \mathsf{L}_2\left(\Omega,\ \mathcal{A},\ \mathsf{P}\right)$ classically corresponds to looking for $\widehat{Y}$ which in a certain sense minimizes the mean squared error of prediction[6] (MSEP)

$$\mathsf{MSEP}\left(\widehat{Y}\right) = \mathbb{E}\left(\widehat{Y} - Y\right)^2.$$

We now distinguishes two situations:

(i) No exogenous information represented by the value of a random vector $\boldsymbol{X}$ is available to construct the prediction. In that case, we get (see also *Probability Theory 1 (NMSA333)* course)

$$\widehat{Y} = \underset{\widetilde{Y}\in\mathsf{L}_2(\emptyset)}{\operatorname{argmin}}\, \mathbb{E}\left(\widetilde{Y} - Y\right)^2 = \underset{\widetilde{Y}\in\mathbb{R}}{\operatorname{argmin}}\, \mathbb{E}\left(\widetilde{Y} - Y\right)^2 = \mathbb{E}\left(Y\right) = \gamma := \widehat{Y}^M.$$

In the following, we will call $\widehat{Y}^M$ as a *marginal* prediction of Y since it is based purely on the marginal distribution of the random variable Y. The MSEP is then

$$\mathsf{MSEP}\left(\widehat{Y}^M\right) = \mathbb{E}\left(\gamma - Y\right)^2 = \mathsf{var}\left(Y\right) = \zeta^2.$$

(ii) The value of a random vector $\boldsymbol{X}$ is available, which is mathematically represented by knowledge of the σ-algebra $\sigma(\boldsymbol{X})$ and the related probability measure $\mathsf{P}|_{\sigma(\boldsymbol{X})}$. This can be used to construct the prediction. Then (again, see *Probability Theory 1 (NMSA333)* course for details)

$$\widehat{Y} = \underset{\widetilde{Y}\in\mathsf{L}_2(\boldsymbol{X})}{\operatorname{argmin}}\ \mathbb{E}\left(\widetilde{Y} - Y\right)^2 = \mathbb{E}\left(Y \,\middle|\, \boldsymbol{X}\right) = \boldsymbol{X}^\top\boldsymbol{\beta} := \widehat{Y}^C,$$

which will be referred to as a *conditional* prediction of Y since it is based on the conditional distribution of Y given $\boldsymbol{X}$. Its MSEP is

$$\begin{aligned}\mathsf{MSEP}\left(\widehat{Y}^C\right) = \mathbb{E}\left(\boldsymbol{X}^\top\boldsymbol{\beta} - Y\right)^2 &= \mathbb{E}\left[\mathbb{E}\left\{\left(\boldsymbol{X}^\top\boldsymbol{\beta} - Y\right)^2 \,\middle|\, \boldsymbol{X}\right\}\right] \\ &= \mathbb{E}\left\{\mathsf{var}\left(Y \,\middle|\, \boldsymbol{X}\right)\right\} = \mathbb{E}\left(\sigma^2\right) = \sigma^2.\end{aligned}$$

In practice, the conditional prediction corresponds to a situation when covariates/regressors represented by the vector $\boldsymbol{X}$ are available to provide some information concerning the response Y. On the other hand, the marginal prediction corresponds to a situation when no exogenous information on Y is available.

[6] *střední čtvercová chyba predikce*

To compare the marginal and the conditional prediction, we introduce the ratio of the two MSEP's:

$$\frac{\mathsf{MSEP}\big(\widehat{Y}^C\big)}{\mathsf{MSEP}\big(\widehat{Y}^M\big)} = \frac{\sigma^2}{\zeta^2}.$$

That is, the ratio σ^2/ζ^2 quantifies advantage of using the prediction $\widehat{Y}^C$ based on the regression model and the covariate/regressor values $\boldsymbol{X}$ compared to using the prediction $\widehat{Y}^M$ which does not require any exogenous information and is equal to the marginal response expectation.

5.4.4 Coefficient of determination

In practice, data (the response vector $\boldsymbol{Y}$ and the model matrix $\mathbb{X}$) are available to estimate the unknown parameters using the linear models M_C: $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}\big) = r$ and M_M: $\boldsymbol{Y} \sim \big(\mathbf{1}_n\gamma,\, \zeta^2\,\mathbf{I}_n\big)$. The unbiased estimators of the conditional and the marginal variance are:

$$\widehat{\sigma}^2 = \frac{1}{n-r}\mathsf{SS}_e = \frac{1}{n-r}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2,$$

$$\widehat{\zeta}^2 = \frac{1}{n-1}\mathsf{SS}_T = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

where $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1,\, \ldots,\, \widehat{Y}_n\big)^\top$ are the fitted values from the model M_C. Note that $\widehat{\zeta}^2$ is a classical sample variance based on data given by the response vector $\boldsymbol{Y}$. That is, a suitable estimator of the ratio σ^2/ζ^2 is

$$\frac{\frac{1}{n-r}\,\mathsf{SS}_e}{\frac{1}{n-1}\mathsf{SS}_T} = \frac{n-1}{n-r}\,\frac{\mathsf{SS}_e}{\mathsf{SS}_T}. \tag{5.14}$$

Alternatively, if $Y_i \overset{\text{i.i.d.}}{\sim} Y$, $i = 1,\, \ldots,\, n$, $Y \sim \mathcal{N}(\gamma,\, \zeta^2)$, that is, if $Y_1, \ldots, Y_n$ is a random sample from $\mathcal{N}(\gamma,\, \zeta^2)$, it can be (it was) easily derived that a quantity

$$\frac{1}{n}\mathsf{SS}_T = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)^2$$

is the maximum-likelihood estimator[7] (MLE) of the marginal variance ζ^2. Analogously, if $Y \,|\, \boldsymbol{X} \sim \mathcal{N}\big(\boldsymbol{X}^\top\boldsymbol{\beta},\, \sigma^2\big)$, it can be derived (see the exercise class) that a quantity

$$\frac{1}{n}\mathsf{SS}_e = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \widehat{Y}\big)^2$$

is the MLE of the *conditional* variance σ^2. Alternative estimator of the ratio σ^2/ζ^2 is then

$$\frac{\frac{1}{n}\,\mathsf{SS}_e}{\frac{1}{n}\mathsf{SS}_T} = \frac{\mathsf{SS}_e}{\mathsf{SS}_T}. \tag{5.15}$$

Remember that in the model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ with intercept ($\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$), we have,

$$\underbrace{\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)^2}_{\mathsf{SS}_T} = \underbrace{\sum_{i=1}^{n}\big(Y_i - \widehat{Y}_i\big)^2}_{\mathsf{SS}_e} + \underbrace{\sum_{i=1}^{n}\big(\widehat{Y}_i - \overline{Y}\big)^2}_{\mathsf{SS}_R},$$

[7] *maximálně věrohodný odhad*

where the three sums of squares represent different sources of the response variability:

SS_T (total sum of squares):	original (marginal) variability of the response,
SS_e (residual sum of squares):	variability *not explained* by the regression model, (residual variability, conditional variability)
SS_R (regression sum of squares):	variability *explained* by the regression model.

Expressions (5.14) and (5.15) then motivate the following definition.

Definition 5.5 Coefficients of determination.
Consider a linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}) = r$ *where* $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$. *A value*

$$R^2 = 1 - \frac{\mathsf{SS}_e}{\mathsf{SS}_T}$$

is called the coefficient of determination[8] *of the linear model.*

A value

$$R^2_{adj} = 1 - \frac{n-1}{n-r}\,\frac{\mathsf{SS}_e}{\mathsf{SS}_T}$$

is called the adjusted coefficient of determination[9] *of the linear model.*

Notes.

- By Theorem 5.8, $\mathsf{SS}_T = \mathsf{SS}_e + \mathsf{SS}_R$ and at the same time $\mathsf{SS}_T \geq 0$. Hence

  $$0 \leq R^2 \leq 1, \qquad 0 \leq R^2_{adj} \leq 1,$$

  and R^2 can also be expressed as

  $$R^2 = \frac{\mathsf{SS}_R}{\mathsf{SS}_T}.$$

- Both R^2 and R^2_{adj} are often reported as $R^2 \cdot 100\%$ and $R^2_{adj} \cdot 100\%$ which can be interpreted as a percentage of the response variability *explained* by the regression model.
- Both R^2 and R^2_{adj} quantify a relative improvement of the quality of prediction if the regression model and the conditional distribution of response given the covariates is used compared to the prediction based on the marginal distribution of the response.
- Both coefficients of determination only quantifies the *predictive ability* of the model. They do not say much about the quality of the model with respect to the possibility to capture correctly the conditional mean $\mathbb{E}\big(Y \,|\, \boldsymbol{X}\big)$. Even a model with a low value of R^2 (R^2_{adj}) might be useful with respect to modelling the conditional mean $\mathbb{E}\big(Y \,|\, \boldsymbol{X}\big)$. The model is perhaps only useless for prediction purposes.

End of Lecture #12 (10/11/2016)

Start of Lecture #14 (16/11/2016)

5.4.5 Overall F-test

Lemma 5.9 Overall F-test.
Assume a normal *linear model* $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r > 1$ *where* $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$.

[8] *koeficient determinace* [9] *upravený koeficient determinace*

Let R^2 be its coefficient of determination. The submodel F-statistic to compare model $\mathsf{M}: \boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *and the only intercept model* $\mathsf{M}_0: \boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbf{1}_n\gamma,\, \sigma^2\mathbf{I}_n\big)$ *takes the form*

$$F_0 = \frac{R^2}{1-R^2} \cdot \frac{n-r}{r-1}. \tag{5.16}$$

Proof.

- $R^2 = 1 - \frac{\mathsf{SS}_e}{\mathsf{SS}_T}$ and according to Lemma 5.6: $\mathsf{SS}_T = \mathsf{SS}_e^0$.
- Hence

$$R^2 = 1 - \frac{\mathsf{SS}_e}{\mathsf{SS}_e^0} = \frac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{\mathsf{SS}_e^0}, \qquad 1 - R^2 = \frac{\mathsf{SS}_e}{\mathsf{SS}_e^0}.$$

- At the same time

$$F_0 = \frac{\frac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{r-1}}{\frac{\mathsf{SS}_e}{n-r}} = \frac{n-r}{r-1}\,\frac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{\mathsf{SS}_e} = \frac{n-r}{r-1}\,\frac{\frac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{\mathsf{SS}_e^0}}{\frac{\mathsf{SS}_e}{\mathsf{SS}_e^0}} = \frac{n-r}{r-1}\,\frac{R^2}{1-R^2}.$$

❑

Note. The F-test with the test statistic (5.16) is sometimes (especially in some software packages) referred to as an *overall goodness-of-fit* test. Nevertheless be cautious when interpreting the results of such test. It says practically nothing about the quality of the model and the "goodness-of-fit"!

Chapter 6

General Linear Model

We still assume that data are represented by a set of n random vectors $\big(Y_i,\, \boldsymbol{X}_i^\top\big)^\top$, $\boldsymbol{X}_i = \big(X_{i,0},\, \ldots,\, X_{i,k-1}\big)^\top$, $i = 1, \ldots, n$, and use symbols $\boldsymbol{Y}$ for a vector $\big(Y_1,\, \ldots,\, Y_n\big)^\top$ and $\mathbb{X}$ for an $n \times k$ matrix with rows given by the covariate/regressor vectors $\boldsymbol{X}_1,\, \ldots,\, \boldsymbol{X}_n$. In this chapter, we mildly extend a linear model by allowing for a (conditional) covariance matrix having different form than $\sigma^2\, \mathbf{I}_n$ assumed by now.

Definition 6.1 General linear model.

The data $\big(\boldsymbol{Y},\, \mathbb{X}\big)$ *satisfy a* general linear model[1] *if*

$$\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}, \qquad \mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \sigma^2\, \mathbb{W}^{-1},$$

where $\boldsymbol{\beta} \in \mathbb{R}^k$ *and* $0 < \sigma^2 < \infty$ *are unknown parameters and* $\mathbb{W}$ *is a* known *positive definite matrix.*

Notes.

- The fact that data follow a general linear model will be denoted as

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big).$$

- General linear model should not be confused with a *generalized linear model*[2] which is something different (see *Advanced Regression Models (NMST432)* course). In the literature, abbreviation "GLM" is used for (unfortunately) both general and generalized linear model. It must be clear from context which of the two is meant.

Example 6.1 (Regression based on sample means).

Suppose that data are represented by random vectors $\big(\widetilde{Y}_{1,1}, \ldots, \widetilde{Y}_{1,w_1},\, \boldsymbol{X}_1^\top\big)^\top$, $\ldots$, $\big(\widetilde{Y}_{n,1}, \ldots, \widetilde{Y}_{n,w_n},\, \boldsymbol{X}_n^\top\big)^\top$

such that for each $i = 1, \ldots, n$, *the random variables* $\widetilde{Y}_{i,1}, \ldots, \widetilde{Y}_{i,w_i}$ *are* uncorrelated *with a common conditional (given* $\boldsymbol{X}_i$*) variance* σ^2.

[1] *obecný lineární model* [2] *zobecněný lineární model*

Suppose that with respect to the response, we are only able to observe the sample means of the "$\widetilde{Y}$" variables leading to the response variables $Y_1, \ldots, Y_n$, where

$$Y_1 = \frac{1}{w_1}\sum_{j=1}^{w_1} \widetilde{Y}_{1,j}, \quad \ldots, \quad Y_n = \frac{1}{w_n}\sum_{j=1}^{w_n} \widetilde{Y}_{n,j}.$$

The covariance matrix (conditional given $\mathbb{X}$) of a random vector $\boldsymbol{Y} = \big(Y_1, \ldots, Y_n\big)^\top$ is then

$$\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \sigma^2 \underbrace{\begin{pmatrix} \frac{1}{w_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{w_n} \end{pmatrix}}_{\mathbb{W}^{-1}}.$$

Theorem 6.1 Generalized least squares.

Assume a general linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big)$, where $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \leq k < n$. The following then holds:

(i) A vector

$$\widehat{\boldsymbol{Y}}_G := \mathbb{X}\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$$

is the best linear unbiased estimator (BLUE) of a vector parameter $\boldsymbol{\mu} := \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}$, and

$$\mathsf{var}\big(\widehat{\boldsymbol{Y}}_G \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbb{X}\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}\mathbb{X}^\top.$$

Both $\widehat{\boldsymbol{Y}}_G$ and $\mathsf{var}\big(\widehat{\boldsymbol{Y}}_G \,\big|\, \mathbb{X}\big)$ do not depend on a choice of the pseudoinverse $\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}$.

If further $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big)$ then

$$\widehat{\boldsymbol{Y}}_G \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbb{X}\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}\mathbb{X}^\top\big).$$

(ii) Let $\mathbf{l} \in \mathbb{R}^k$, $\mathbf{l} \neq \mathbf{0}_k$, be such that $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ is an estimable parameter of the model and let

$$\boldsymbol{b}_G := \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

Then $\widehat{\theta}_G = \mathbf{l}^\top\boldsymbol{b}_G$ does not depend on a choice of the pseudoinverse used to calculate $\boldsymbol{b}_G$ and $\widehat{\theta}_G$ is the best linear unbiased estimator (BLUE) of θ with

$$\mathsf{var}\big(\widehat{\theta}_G \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}\mathbf{l},$$

which also does not depend on a choice of the pseudoinverse.

If further $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big)$ then

$$\widehat{\theta}_G \,\big|\, \mathbb{X} \sim \mathcal{N}\big(\theta,\, \sigma^2\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-}\mathbf{l}\big).$$

(iii) If further $r = k$ (full-rank general linear model), then

$$\widehat{\boldsymbol{\beta}}_G := \left(\mathbb{X}^\top \mathbb{W} \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbb{W} \boldsymbol{Y}$$

is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ with

$$\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X}\big) = \sigma^2 \left(\mathbb{X}^\top \mathbb{W} \mathbb{X}\right)^{-1}.$$

If additionally $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbb{W}^{-1}\big)$ then

$$\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X} \sim \mathcal{N}_k\big(\boldsymbol{\beta},\, \sigma^2 \left(\mathbb{X}^\top \mathbb{W} \mathbb{X}\right)^{-1}\big).$$

(iv) The statistic

$$\mathsf{MS}_{e,G} := \frac{\mathsf{SS}_{e,G}}{n - r},$$

where

$$\mathsf{SS}_{e,G} := \left\| \mathbb{W}^{\frac{1}{2}} \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\right) \right\|^2 = \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\right)^\top \mathbb{W} \big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\big),$$

is the unbiased estimator of the residual variance σ^2.
If additionally $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbb{W}^{-1}\big)$ then

$$\frac{\mathsf{SS}_{e,G}}{\sigma^2} \sim \chi^2_{n-r},$$

and the statistics $\mathsf{SS}_{e,G}$ and $\widehat{\boldsymbol{Y}}_G$ are conditionally, given $\mathbb{X}$, independent.

Proof. Matrices $\mathbb{W}^{-1}$ and $\mathbb{W}$ are positive definite. Hence there exist $\mathbb{W}^{\frac{1}{2}}$ such that

$$\mathbb{W} = \mathbb{W}^{\frac{1}{2}} \big(\mathbb{W}^{\frac{1}{2}}\big)^\top, \quad \text{e.g., Cholesky decomposition}$$
$$\mathbb{W}^{-1} = \mathbb{W}^{-\frac{1}{2}} \big(\mathbb{W}^{-\frac{1}{2}}\big)^\top.$$

(i) Let $\boldsymbol{Y}^\star = \mathbb{W}^{\frac{1}{2}} \boldsymbol{Y}$.
Then $\mathbb{E}\big(\boldsymbol{Y}^\star \,\big|\, \mathbb{X}\big) = \mathbb{W}^{\frac{1}{2}} \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{W}^{\frac{1}{2}} \mathbb{X} \boldsymbol{\beta}$,

$$\mathsf{var}\big(\boldsymbol{Y}^\star \,\big|\, \mathbb{X}\big) = \mathbb{W}^{\frac{1}{2}} \underbrace{\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)}_{\sigma^2 \mathbb{W}^{-1}} \big(\mathbb{W}^{\frac{1}{2}}\big)^\top = \sigma^2 \mathbf{I}_n.$$

That is, we have a linear model $\mathsf{M}^\star$

$$\mathsf{M}^\star\colon\ \boldsymbol{Y}^\star \,\big|\, \mathbb{X} \ \sim\ \big(\underbrace{\mathbb{W}^{\frac{1}{2}} \mathbb{X}}_{\mathbb{X}^\star} \boldsymbol{\beta},\ \sigma^2\, \mathbf{I}_n\big),$$

where $\mathsf{rank}\big(\mathbb{X}^\star\big) = \mathsf{rank}\big(\mathbb{W}^{\frac{1}{2}} \mathbb{X}\big) = \mathsf{rank}\big(\mathbb{X}\big) = r$.
The hat matrix for model $\mathsf{M}^\star$ is

$$\mathbb{H}^\star \ =\ \mathbb{W}^{\frac{1}{2}} \mathbb{X} \big(\mathbb{X}^\top \mathbb{W} \mathbb{X}\big)^{-} \mathbb{X}^\top \big(\mathbb{W}^{\frac{1}{2}}\big)^\top,$$

which does not depend on a choice of a pseudoinverse $\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top$. Note that due to regularity of the matrix $\mathbb{W}^{\frac{1}{2}}$, also expression $\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top$ does not depend on a choice of this pseudoinverse.

The fitted values in model $\mathsf{M}^\star$ are then calculated as

$$\widehat{\boldsymbol{Y}}^\star \;=\; \mathbb{H}^\star\boldsymbol{Y}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

By Gauss-Markov theorem (Theorem 2.2), the vector $\widehat{\boldsymbol{Y}}^\star$ is the best linear unbiased estimator (BLUE) of the vector $\mathbb{E}(\boldsymbol{Y}^\star \,|\, \mathbb{X}) \;=\; \mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta}$ with

$$\mathsf{var}(\widehat{\boldsymbol{Y}}^\star \,|\, \mathbb{X}) \;=\; \sigma^2\,\mathbb{H}^\star \;=\; \sigma^2\,\mathbb{W}^{\frac{1}{2}}\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top\big(\mathbb{W}^{\frac{1}{2}}\big)^\top.$$

By linearity, the vector

$$\widehat{\boldsymbol{Y}}_G \;:=\; \mathbb{W}^{-\frac{1}{2}}\widehat{\boldsymbol{Y}}^\star \;=\; \mathbb{X}\,(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$$

is the BLUE of the vector

$$\mathbb{W}^{-\frac{1}{2}}\mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta} \;=\; \mathbb{X}\boldsymbol{\beta} \;=\; \mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{X}),$$

and

$$\mathsf{var}(\widehat{\boldsymbol{Y}}_G \,|\, \mathbb{X}) \;=\; \mathbb{W}^{-\frac{1}{2}}\,\mathsf{var}(\widehat{\boldsymbol{Y}}^\star \,|\, \mathbb{X})\,\big(\mathbb{W}^{-\frac{1}{2}}\big)^\top \;=\; \sigma^2\,\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\,\mathbb{X},$$

which does not depend on a choice of a pseudoinverse $(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-$.

If additionally $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big)$, then by properties of a normal distribution (both $\widehat{\boldsymbol{Y}}^\star$ and $\widehat{\boldsymbol{Y}}_G$ are linear functions of $\boldsymbol{Y}$), we have

$$\widehat{\boldsymbol{Y}}^\star \,|\, \mathbb{X} \;\sim\; \mathcal{N}\Big(\mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta},\; \sigma^2\,\mathbb{W}^{\frac{1}{2}}\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top\big(\mathbb{W}^{\frac{1}{2}}\big)^\top\Big),$$
$$\widehat{\boldsymbol{Y}}_G \,|\, \mathbb{X} \;\sim\; \mathcal{N}\Big(\mathbb{X}\boldsymbol{\beta} \quad,\; \sigma^2\,\mathbb{X}(\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\mathbb{X}^\top\Big).$$

(ii) We are assuming that $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ is an estimable parameter of the model $\mathsf{M}:\;\; \boldsymbol{Y} \,|\, \mathbb{X} \;\sim\; \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbb{W}^{-1}\big)$. That is,

$$\forall\,\boldsymbol{\beta}_1,\,\boldsymbol{\beta}_2 \quad \mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 \quad \text{implies } \mathbf{l}^\top\boldsymbol{\beta}_1 = \mathbf{l}^\top\boldsymbol{\beta}_2.$$

Due to regularity of the matrix $\mathbb{W}^{\frac{1}{2}}$, condition $\mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2$ is equivalent to the condition

$$\underbrace{\mathbb{W}^{\frac{1}{2}}\mathbb{X}}_{\mathbb{X}^\star}\boldsymbol{\beta}_1 = \underbrace{\mathbb{W}^{\frac{1}{2}}\mathbb{X}}_{\mathbb{X}^\star}\boldsymbol{\beta}_2.$$

That is,

$$\forall\,\boldsymbol{\beta}_1,\,\boldsymbol{\beta}_2 \quad \mathbb{X}^\star\boldsymbol{\beta}_1 = \mathbb{X}^\star\boldsymbol{\beta}_2 \quad \text{implies } \mathbf{l}^\top\boldsymbol{\beta}_1 = \mathbf{l}^\top\boldsymbol{\beta}_2,$$

and hence parameter $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ is estimable also in model $\mathsf{M}^\star:\;\; \boldsymbol{Y}^\star \,|\, \mathbb{X} \;\sim\; \big(\mathbb{X}^\star\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$, where $\boldsymbol{Y}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\boldsymbol{Y}$, $\mathbb{X}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\mathbb{X}$.

By Theorem 2.5 (LSE and normal equations), we have that

$$\begin{aligned}
\widehat{\boldsymbol{Y}}^\star \;=\; \mathbb{X}^\star\boldsymbol{b}^\star \quad &\Longleftrightarrow \quad \boldsymbol{b}^\star \text{ solves normal equations in model } \mathsf{M}^\star \\
&\Longleftrightarrow \quad \boldsymbol{b}^\star \text{ solves } \mathbb{X}^{\star\top}\mathbb{X}^\star\boldsymbol{b} \;=\; \mathbb{X}^{\star\top}\boldsymbol{Y}^\star \\
&\Longleftrightarrow \quad \boldsymbol{b}^\star \text{ solves } \mathbb{X}^\top\mathbb{W}\mathbb{X}\boldsymbol{b} \;=\; \mathbb{X}^\top\mathbb{W}\boldsymbol{Y} \\
&\Longleftrightarrow \quad \boldsymbol{b}^\star \;=\; (\mathbb{X}^\top\mathbb{W}\mathbb{X})^-\,\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.
\end{aligned}$$

Remember that $\mathbb{X}^\star = \mathbb{W}^{\frac{1}{2}}\mathbb{X}$. Hence,

$$\widehat{\boldsymbol{Y}}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{b}^\star \quad \text{if and only if } \boldsymbol{b}^\star \;=\; \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\,\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

Further, remember that $\widehat{\boldsymbol{Y}}_G \;=\; \mathbb{W}^{-\frac{1}{2}}\widehat{\boldsymbol{Y}}^\star$. Hence,

$$\widehat{\boldsymbol{Y}}_G \;=\; \mathbb{W}^{-\frac{1}{2}}\mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{b}^\star \quad \text{if and only if } \boldsymbol{b}^\star \;=\; \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\,\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

That is,

$$\widehat{\boldsymbol{Y}}_G \;=\; \mathbb{X}\boldsymbol{b}_G \quad \text{if and only if } \boldsymbol{b}_G := \boldsymbol{b}^\star \;=\; \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\,\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

Then, by Gauss-Markov theorem (Theorem 2.8),

$$\widehat{\theta}_G \;:=\; \widehat{\theta}^\star \;=\; \mathbf{l}^\top\boldsymbol{b}_G$$

is BLUE of the parameter $\theta \;=\; \mathbf{l}^\top\boldsymbol{\beta}$, which does not depend on a choice of a pseudoinverse $\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-$. Furthermore,

$$\mathsf{var}\big(\widehat{\theta}_G \,\big|\, \mathbb{X}\big) \;=\; \mathsf{var}\big(\widehat{\theta}^\star \,\big|\, \mathbb{X}\big) \;=\; \sigma^2\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\mathbf{l},$$

which also does not depend on a choice of a pseudoinverse $\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-$. If additionally, $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big)$ then by properties of a normal distribution (only linear transformations are involved to calculate $\widehat{\theta}_G$ from $\boldsymbol{Y}$), we have

$$\widehat{\theta}_G \,\big|\, \mathbb{X} \;\sim\; \mathcal{N}\big(\theta,\, \sigma^2\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\mathbf{l}\big).$$

(iii) Suppose that for an $m \times k$ matrix, the parameter $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ is an estimable vector parameter of the model $\mathsf{M}\colon\ \boldsymbol{Y} \,\big|\, \mathbb{X} \;\sim\; \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbb{W}^{-1}\big)$. By analogous steps as in (ii), we show that

$$\widehat{\boldsymbol{\theta}}_G \;:=\; \mathbb{L}\boldsymbol{b}_G, \quad \boldsymbol{b}_G \;=\; \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\,\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$$

is BLUE of $\boldsymbol{\theta}$, which does not depend on a choice of a pseudoinverse $\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-$. Furthermore,

$$\mathsf{var}\big(\widehat{\boldsymbol{\theta}}_G \,\big|\, \mathbb{X}\big) \;=\; \sigma^2\,\mathbb{L}\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\mathbb{L}^\top,$$

and under assumption of normality,

$$\widehat{\boldsymbol{\theta}}_G \,\big|\, \mathbb{X} \;\sim\; \mathcal{N}_m\big(\boldsymbol{\theta},\, \sigma^2\,\mathbb{L}\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^-\mathbb{L}^\top\big).$$

Now, if $\mathsf{rank}\big(\mathbb{X}\big) = k$, the matrix $\mathbb{X}^\top\mathbb{W}\mathbb{X}$ is invertible and hence its only pseudoinverse is $\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^- \;=\; \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-1}$. Moreover, the vector parameter $\boldsymbol{\beta}$ is estimable and by taking $\mathbb{L} = \mathbf{I}_k$ we obtain that its BLUE is

$$\widehat{\boldsymbol{\beta}}_G \;:=\; \big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-1}\,\mathbb{X}^\top\mathbb{W}\boldsymbol{Y},$$

$$\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X}\big) \;=\; \sigma^2\,\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-1},$$

and under assumption of normality,

$$\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X} \;\sim\; \mathcal{N}_k\big(\boldsymbol{\beta},\, \sigma^2\,\big(\mathbb{X}^\top\mathbb{W}\mathbb{X}\big)^{-1}\big).$$

(iv) Let us first calculate the residual sum of squares of the model $\mathsf{M}^\star$: $\boldsymbol{Y}^\star \,\big|\, \mathbb{X} \sim \big(\mathbb{X}^\star\boldsymbol{\beta},\ \sigma^2\,\mathbf{I}_n\big)$, where $\boldsymbol{Y}^\star = \mathbb{W}^{\frac{1}{2}}\boldsymbol{Y}$, $\mathbb{X}^\star = \mathbb{W}^{\frac{1}{2}}\mathbb{X}$, $\mathsf{rank}\big(\mathbb{X}^\star\big) = \mathsf{rank}\big(\mathbb{X}\big) = r$. We have (remember further that $\widehat{\boldsymbol{Y}}^\star = \mathbb{W}^{\frac{1}{2}}\widehat{\boldsymbol{Y}}_G$)

$$\begin{aligned}\mathsf{SS}_e^\star \;&=\; \big(\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\big)^\top\big(\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\big) \;=\; \big(\mathbb{W}^{\frac{1}{2}}\boldsymbol{Y} - \mathbb{W}^{\frac{1}{2}}\widehat{\boldsymbol{Y}}_G\big)^\top\big(\mathbb{W}^{\frac{1}{2}}\boldsymbol{Y} - \mathbb{W}^{\frac{1}{2}}\widehat{\boldsymbol{Y}}_G\big)\\ &=\; \big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\big)^\top\mathbb{W}\big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\big) \;=:\; \mathsf{SS}_{e,G}.\end{aligned}$$

By Theorem 2.3, we have

$$\mathbb{E}\big(\mathsf{SS}_{e,G}\big) \;=\; \mathbb{E}\big(\mathsf{SS}_e^\star\big) \;=\; (n-r)\,\sigma^2 \;=\; \mathbb{E}\big(\mathsf{SS}_e^\star\,\big|\,\mathbb{X}\big) \;=\; \mathbb{E}\big(\mathsf{SS}_{e,G}\,\big|\,\mathbb{X}\big).$$

That is,

$$\mathsf{MS}_{e,G} \;:=\; \frac{\mathsf{SS}_{e,G}}{n-r}$$

is the unbiased estimator of the residual variance σ^2.

Furthermore, if normality is assumed, Theorem 3.2 applied to model $\mathsf{M}^\star$ provides that

$$\frac{\mathsf{SS}_e^\star}{\sigma^2} \;\sim\; \chi^2_{n-r}.$$

Since $\mathsf{SS}_e^\star = \mathsf{SS}_{e,G}$, we have directly

$$\frac{\mathsf{SS}_{e,G}}{\sigma^2} \;\sim\; \chi^2_{n-r}.$$

Finally, Theorem 3.2 also provides (conditional, given $\mathbb{X}$) independence of $\widehat{\boldsymbol{Y}}^\star$ and $\mathsf{SS}_e^\star$. Nevertheless, since $\widehat{\boldsymbol{Y}}_G = \mathbb{W}^{-\frac{1}{2}}\widehat{\boldsymbol{Y}}^\star$ and $\mathsf{SS}_{e,G} = \mathsf{SS}_e^\star$, we also have (conditional, given $\mathbb{X}$) independence of $\widehat{\boldsymbol{Y}}_G$ and $\mathsf{SS}_{e,G}$.

❑

Note. Mention also that as consequence of the above theorem, all classical tests, confidence intervals etc. work in the same way as in the OLS case.

Terminology *(Generalized fitted values, residual sum of squares, mean square, least square estimator).*

- The statistic $\widehat{\boldsymbol{Y}}_G = \mathbb{X}\left(\mathbb{X}^\top\mathbb{W}\mathbb{X}\right)^{-}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$ is called the vector of *the generalized fitted values.*[3]
- The statistic $\mathsf{SS}_{e,G} = \left\|\mathbb{W}^{\frac{1}{2}}\left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\right)\right\|^2 = \left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\right)^\top\mathbb{W}\left(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\right)$ is called *the generalized residual sum of squares.*[4]
- The statistic $\mathsf{MS}_{e,G} = \dfrac{\mathsf{SS}_{e,G}}{n-r}$ is called *the generalized mean square.*[5]
- The statistic $\widehat{\boldsymbol{\beta}}_G = \left(\mathbb{X}^\top\mathbb{W}\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$ in a full-rank general linear model is called *the generalized least squares (GLS) estimator*[6] of the regression coefficients.

[3] *zobecněné vyrovnané hodnoty* [4] *zobecněný reziduální součet čtverců* [5] *zobecněný střední čtverec* [6] *odhad metodou zobecněných nejmenších čtverců*

Note. The most common use of the generalized least squares is the situation described in Example 6.1, where

$$\mathbb{W}^{-1} = \begin{pmatrix} \frac{1}{w_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{w_n} \end{pmatrix}.$$

We then get

$$\mathbb{X}^\top \mathbb{W} \boldsymbol{Y} = \sum_{i=1}^{n} w_i Y_i \boldsymbol{X}_i, \qquad \mathbb{X}^\top \mathbb{W} \mathbb{X} = \sum_{i=1}^{n} w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

$$\mathsf{SS}_{e,G} = \sum_{i=1}^{n} w_i \big(Y_i - \widehat{Y}_{G,i}\big)^2.$$

The method of the generalized least squares is then usually referred to as the method of *the weighted least squares (WLS)*.[7]

Partial end of Lecture #14 (16/11/2016)

[7] *vážené nejmenší čtverce*

Chapter 7

Parameterizations of Covariates

7.1 Linearization of the dependence of the response on the covariates

Start of Lecture #5 (13/10/2016)

As it is usual in this lecture, we represent data by n random vectors $\left(Y_i,\, \boldsymbol{Z}_i^\top\right)^\top$, $\boldsymbol{Z}_i = \left(Z_{i,1},\, \ldots,\, Z_{i,p}\right)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$. The principal problem we consider is to find a suitable model to express the (conditional) response expectation $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right)$, where $\boldsymbol{Y} = \left(Y_1,\, \ldots,\, Y_n\right)^\top$ and $\mathbb{Z}$ is a matrix with vectors $\boldsymbol{Z}_1$, …, $\boldsymbol{Z}_n$ in its rows. To this end, we consider a linear model, where $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right)$ can be expressed as $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$, where

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}, \qquad \begin{aligned} \boldsymbol{X}_1 = \left(X_{1,0},\, \ldots,\, X_{1,k-1}\right)^\top &= \boldsymbol{t}(\boldsymbol{Z}_1), \\ &\vdots \\ \boldsymbol{X}_n = \left(X_{n,0},\, \ldots,\, X_{n,k-1}\right)^\top &= \boldsymbol{t}(\boldsymbol{Z}_n), \end{aligned}$$

and $\boldsymbol{t} : \mathcal{Z} \longrightarrow \mathcal{X} \subseteq \mathbb{R}^k$, $\boldsymbol{t}(\boldsymbol{z}) = \left(t_0(\boldsymbol{z}),\, \ldots,\, t_{k-1}(\boldsymbol{z})\right)^\top = \left(x_0,\, \ldots,\, x_{k-1}\right)^\top = \boldsymbol{x}$, is a suitable transformation of the original covariates that *linearize* the relationship between the response expectation and those covariates. The corresponding regression function is then

$$m(\boldsymbol{z}) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} = \beta_0\, t_0(\boldsymbol{z}) \,+\, \cdots \,+\, \beta_{k-1}\, t_{k-1}(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathcal{Z}. \tag{7.1}$$

One of the main problems of a regression analysis is to find a reasonable form of the transformation function $\boldsymbol{t}$ to obtain a model that is perhaps wrong but at least useful to capture sufficiently the form of $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right)$ and in general to express $\mathbb{E}\left(Y \,\middle|\, \boldsymbol{Z} = \boldsymbol{z}\right)$, $\boldsymbol{z} \in \mathcal{Z}$, for a generic response Y being generated, given the covariate value $\boldsymbol{Z} = \boldsymbol{z}$, by the same probabilistic mechanism as the original data.

7.2 Parameterization of a single covariate

In this and two following sections, we first limit ourselves to the situation of a single covariate, i.e., $p = 1$, $\mathcal{Z} \subseteq \mathbb{R}$, and show some classical choices of the transformations that are used in practical analyses when attempting to find a useful linear model.

7.2.1 Parameterization

Our aim is to propose transformations $\boldsymbol{t} : \mathcal{Z} \longrightarrow \mathbb{R}^k$, $\boldsymbol{t}(z) = \big(t_0(z), \ldots, t_{k-1}(z)\big)^\top$ such that a regression function (7.1) can possibly provide a useful model for the response expectation $\mathbb{E}\big(Y \,\big|\, Z = z\big)$. Furthermore, in most cases, we limit ourselves to transformations that lead to a linear model with intercept. In such cases, the regression function will be

$$m(z) = \beta_0 + \beta_1\, s_1(z) + \cdots + \beta_{k-1}\, s_{k-1}(z), \quad z \in \mathcal{Z}, \tag{7.2}$$

where the non-intercept part of the transformation $\boldsymbol{t}$ will be denoted as $\boldsymbol{s}$. That is, for $z \in \mathcal{Z}$,

$$s_j(z) = t_j(z), \qquad j = 1, \ldots, k-1,$$
$$\boldsymbol{s} : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}, \quad \boldsymbol{s}(z) = \big(s_1(z), \ldots, s_{k-1}(z)\big)^\top = \big(t_1(z), \ldots, t_{k-1}(z)\big)^\top.$$

Definition 7.1 Parameterization of a covariate.

Let $Z_1, \ldots, Z_n$ be values of a given univariate covariate $Z \in \mathcal{Z} \subseteq \mathbb{R}$. By a parameterization *of this covariate we mean*

(i) a function $\boldsymbol{s} : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}$, $\boldsymbol{s}(z) = \big(s_1(z), \ldots, s_{k-1}(z)\big)^\top$, $z \in \mathcal{Z}$, where all $s_1, \ldots, s_{k-1}$ are non-constant functions on $\mathcal{Z}$, and

(ii) an $n \times (k-1)$ matrix $\mathbb{S}$, where

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{s}^\top(Z_1) \\ \vdots \\ \boldsymbol{s}^\top(Z_n) \end{pmatrix} = \begin{pmatrix} s_1(Z_1) & \ldots & s_{k-1}(Z_1) \\ \vdots & \vdots & \vdots \\ s_1(Z_n) & \ldots & s_{k-1}(Z_n) \end{pmatrix}.$$

Terminology (Reparameterizing matrix, regressors).

Matrix $\mathbb{S}$ from Definition 7.1 is called *reparameterizing matrix*[1] of a covariate. Its columns, i.e., vectors

$$\boldsymbol{X}^1 = \begin{pmatrix} s_1(Z_1) \\ \vdots \\ s_1(Z_n) \end{pmatrix}, \ \ldots, \ \boldsymbol{X}^{k-1} = \begin{pmatrix} s_{k-1}(Z_1) \\ \vdots \\ s_{k-1}(Z_n) \end{pmatrix}$$

determine the *regressors* of the linear model based on the covariate values $Z_1, \ldots, Z_n$.

[1] *reparametrizační matice*

Notes.

- A model matrix $\mathbb{X}$ of the model with the regression function (7.2) is

$$\mathbb{X} = \Big(\mathbf{1}_n, \ \mathbb{S}\Big) = \Big(\mathbf{1}_n, \ \boldsymbol{X}^1, \ \ldots, \ \boldsymbol{X}^{k-1}\Big) = \begin{pmatrix} 1 & X_{1,1} & \ldots & X_{1,k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \ldots & X_{n,k-1} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{X}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_n^\top \end{pmatrix},$$

$$\boldsymbol{X}_i = \boldsymbol{s}(Z_i), \qquad X_{i,j} = s_j(Z_i), \ i = 1, \ldots, n, \ j = 1, \ldots, k-1.$$

- Definition 7.1 is such that an intercept vector $\mathbf{1}_n$ (or a vector $c\,\mathbf{1}_n$, $c \in \mathbb{R}$) is (with a positive probability provided a non-degenerated covariate distribution) not included in the reparameterizing matrix $\mathbb{S}$. Nevertheless, it will be useful in some situations to consider such parameterizations that (almost surely) include an intercept term in the space generated by the columns of the reparameterizing matrix $\mathbb{S}$ itself. That is, for some parameterizations (see the regression splines in Section 7.3.4), we will have $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{S}\big)$.

7.2.2 Covariate types

The covariate space $\mathcal{Z}$ and the corresponding univariate covariates $Z_1, \ldots, Z_n$ are *usually* of one of the two types and different parameterizations are useful depending on the covariate type which are the following.

Numeric covariates

Numeric[2] covariates are such covariates where a ratio of the two covariate values makes sense and a unity increase of the covariate value has an unambiguous meaning. The numeric covariate is then *usually* of one of the two following subtypes:

(i) *continuous*, in which case $\mathcal{Z}$ is mostly an interval in $\mathbb{R}$. Such covariates have usually a physical interpretation and some units whose choice must be taken into account when interpreting the results of the statistical analysis. The continuous numeric covariates are *mostly* (but not necessarily) represented by continuous random variables.

(ii) *discrete*, in which case $\mathcal{Z}$ is infinite countable or finite (but "large") subset of $\mathbb{R}$. The most common situation of a discrete numeric covariate is a *count*[3] with $\mathcal{Z} \subseteq \mathbb{N}_0$. The numeric discrete covariates are represented by discrete random variables.

Categorical covariates

Categorical[4] covariates (in the R software referred to as *factors*), are such covariates where the ratio of the two covariate values does not necessarily make sense and a unity increase of the covariate value does not necessarily have an unambiguous meaning. The sample space $\mathcal{Z}$ is a finite (and mostly "small") set, i.e.,

$$\mathcal{Z} = \big\{\omega_1, \ \ldots, \ \omega_G\big\},$$

where the values $\omega_1 < \cdots < \omega_G$ are somehow arbitrarily chosen *labels* of categories purely used to obtain a mathematical representation of the covariate values. The categorical covariate is always represented by a discrete random variable. Even for categorical covariates, it is useful to distinguish the two subtypes:

[2] *numerické, příp. kvantitativní* [3] *počet* [4] *kategoriální, příp. kvalitativní*

(i) *nominal*[5] where from a practical point of view, chosen values $\omega_1, \ldots, \omega_G$ are completely arbitrary. Consequently, practically interpretable results and conclusions of any sensible statistical analysis should be invariant towards the choice of $\omega_1, \ldots, \omega_G$. The nominal categorical covariate mostly represents a pertinence to some group (a group label), e.g., region of residence.

(ii) *ordinal*[6] where *ordering* $\omega_1 < \cdots < \omega_G$ makes sense also from a practical point of view. An example is a school grade.

Notes.

- From the practical point of view, it is mainly important to distinguish *numeric* and *categorical* covariates.
- Often, *ordinal categorical* covariate can be viewed also as a *discrete numeric*. Whatever in this lecture that will be applied to the discrete numeric covariate can also be applied to the ordinal categorical covariate if it makes sense to interprete, at least into some extent, its unity increase (and not only the ordering of the covariate values).

[5] *nominální* [6] *ordinální*

7.3 Numeric covariate

It is now assumed that $Z_i \in \mathcal{Z} \subseteq \mathbb{R}$, $i = 1, \ldots, n$, are *numeric* covariates. Our aim is now to propose their sensible parameterizations.

7.3.1 Simple transformation of the covariate

The regression function is

$$m(z) = \beta_0 + \beta_1 s(z), \quad z \in \mathcal{Z}, \tag{7.3}$$

where $s : \mathcal{Z} \longrightarrow \mathbb{R}$ is a suitable *non-constant* function. The corresponding reparameterizing matrix is

$$\mathbb{S} = \begin{pmatrix} s(Z_1) \\ \vdots \\ s(Z_n) \end{pmatrix}.$$

Due to interpretability issues, "simple" functions like: identity, logarithm, exponential, square root, reciprocal, ..., are considered in place of the transformation s.

Evaluation of the effect of the original covariate

Advantage of a model with the regression function (7.3) is the fact that a single regression coefficient β_1 (the slope in a model with the regression line in $x = s(z)$) quantifies the effect of the covariate on the response expectation which can then be easily summarized by a single point estimate and a confidence interval. Evaluation of a statistical significance of the effect of the original covariate on the response expectation is achieved by testing the null hypothesis

$$\mathrm{H}_0 : \ \beta_1 = 0.$$

A possible test procedure was introduced in Section 3.2.

Interpretation of the regression coefficients

Disadvantage is the fact that the slope β_1 expresses the change of the response expectation that corresponds to a unity change of the transformed covariate $X = s(Z)$, i.e., for $z \in \mathcal{Z}$:

$$\beta_1 = \mathbb{E}\big(Y \,\big|\, X = s(z) + 1\big) - \mathbb{E}\big(Y \,\big|\, X = s(z)\big),$$

which is not always easily interpretable.

Moreover, unless the transformation s is a linear function, the change in the response expectation that corresponds to a unity change of the original covariate is a function of that covariate:

$$\mathbb{E}\big(Y \,\big|\, Z = z + 1\big) - \mathbb{E}\big(Y \,\big|\, Z = z\big) = \beta_1\big\{s(z+1) - s(z)\big\}, \qquad z \in \mathcal{Z}.$$

In other words, a model with the regression function (7.3) and a non-linear transformation s expresses the fact that the original covariate has different influence on the response expectation depending on the value of this covariate.

Note. It is easily seen that if $n > k = 2$, the transformation s is strictly monotone and the data contain at least two different values among $Z_1, \ldots, Z_n$ (which has a probability of one if the covariates Z_i are sampled from a continuous distribution), the model matrix $\mathbb{X} = \big(\mathbf{1}_n, \mathbb{S}\big)$ is of a full-rank $r = k = 2$.

7.3.2 Raw polynomials

The regression function is polynomial of a chosen degree $k-1$, i.e.,

$$m(z) = \beta_0 + \beta_1 z + \cdots + \beta_{k-1} z^{k-1}, \quad z \in \mathcal{Z}. \tag{7.4}$$

The parameterization is

$$\boldsymbol{s} : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}, \qquad \boldsymbol{s}(z) = \big(z, \ldots, z^{k-1}\big)^\top, \; z \in \mathcal{Z}$$

and the corresponding reparameterizing matrix is

$$\mathbb{S} = \begin{pmatrix} Z_1 & \ldots & Z_1^{k-1} \\ \vdots & \vdots & \vdots \\ Z_n & \ldots & Z_n^{k-1} \end{pmatrix}.$$

Evaluation of the effect of the original covariate

The effect of the original covariate on the response expectation is now quantified by a set of $k-1$ regression coefficients $\boldsymbol{\beta}^Z := \big(\beta_1, \ldots, \beta_{k-1}\big)^\top$. To evaluate a statistical significance of the effect of the original covariate on the response expectation we have to test the null hypothesis

$$\mathrm{H}_0 : \; \boldsymbol{\beta}^Z = \mathbf{0}_{k-1}.$$

An appropriate test procedure was introduced in Section 3.2.

Interpretation of the regression coefficients

With $k > 2$ (at least a quadratic regression function), the single regression coefficients $\beta_1, \ldots, \beta_{k-1}$ only occasionally have a direct reasonable interpretation. Analogously to simple non-linear transformation of the covariate, the change in the response expectation that corresponds to a unity change of the original covariate is a function of that covariate:

$$\begin{aligned} &\mathbb{E}\big(Y \,\big|\, Z = z+1\big) \; - \; \mathbb{E}\big(Y \,\big|\, Z = z\big) \\ &\qquad = \beta_1 + \beta_2 \big\{(z+1)^2 - z^2\big\} + \cdots + \beta_{k-1} \big\{(z+1)^{k-1} - z^{k-1}\big\}, \qquad z \in \mathcal{Z}. \end{aligned}$$

Note. It is again easily seen that if $n > k$ and the data contain at least k different values among among $Z_1, \ldots, Z_n$ (which has a probability of one if the covariates Z_i are sampled from a continuous distribution), the model matrix $\big(\mathbf{1}_n, \mathbb{S}\big)$ is of a full-rank $r = k$.

Degree of a polynomial

Test on a subset of regression coefficients (Section 3.2) or a submodel test (Section 5.2) can be used to infer on the degree of a polynomial in the regression function (7.4). The null hypothesis expressing, for $d < k$, belief that the regression function is a polynomial of degree $d-1$ corresponds to the null hypothesis

$$\mathrm{H}_0 : \; \beta_d = 0 \;\; \& \;\; \ldots \;\; \& \;\; \beta_{k-1} = 0.$$

7.3.3 Orthonormal polynomials

The regression function is again polynomial of a chosen degree $k-1$, nevertheless, a different basis of the regression space, i.e., a different parameterization of the polynomial is used. Namely, the regression function is

$$m(z) = \beta_0 + \beta_1 P^1(z) + \cdots + \beta_{k-1} P^{k-1}(z), \quad z \in \mathcal{Z}, \tag{7.5}$$

where P^j is an *orthonormal polynomial* of degree j, $j = 1, \ldots, k-1$ built above a set of the covariate datapoints $Z_1, \ldots, Z_n$. That is,

$$P^j(z) = a_{j,0} + a_{j,1}\, z + \cdots + a_{j,j}\, z^j, \qquad j = 1, \ldots, k-1, \tag{7.6}$$

and the polynomial coefficients $a_{j,l}$, $j = 1, \ldots, k-1$, $l = 0, \ldots, j$ are such that vectors

$$\boldsymbol{P}^j = \begin{pmatrix} P^j(Z_1) \\ \vdots \\ P^j(Z_n) \end{pmatrix}, \quad j = 1, \ldots, k-1,$$

are all *orthonormal* and also *orthogonal* to an intercept vector $\boldsymbol{P}^0 = \big(1, \ldots, 1\big)^\top$. The corresponding reparameterizing matrix is

$$\mathbb{S} = \Big(\boldsymbol{P}^1, \quad \ldots, \quad \boldsymbol{P}^{k-1}\Big) = \begin{pmatrix} P^1(Z_1) & \ldots & P^{k-1}(Z_1) \\ \vdots & \vdots & \vdots \\ P^1(Z_n) & \ldots & P^{k-1}(Z_n) \end{pmatrix}, \tag{7.7}$$

which leads to the model matrix $\mathbb{X} = \big(\mathbf{1}_n, \ \mathbb{S}\big)$ which have all columns mutually *orthogonal* and the non-intercept columns having even a unity norm. For methods of calculation of the coefficients of the polynomials (7.6), see lectures on linear algebra. It can only be mentioned here that as soon as the data contain at least k different values among $Z_1, \ldots, Z_n$, those polynomial coefficients exist and are unique.

Note. For given dataset and given polynomial degree $k-1$, the model matrix $\mathbb{X} = \big(\mathbf{1}_n, \ \mathbb{S}\big)$ based on the orthonormal polynomial provide the same regression space as the model matrix based on the raw polynomials. Hence, the two model matrices determine two equivalent linear models.

Advantages of orthonormal polynomials compared to raw polynomials

- All non-intercept columns of the model matrix have the same (unity) norm. Consequently, all non-intercept regression coefficients $\beta_1, \ldots, \beta_{k-1}$ have the same scale. This may be helpful when evaluating a practical (not statistical!) importance of higher-order degree polynomial terms.
- Matrix $\mathbb{X}^\top\mathbb{X}$ is a diagonal matrix $\mathsf{diag}(n, 1, \ldots, 1)$. Consequently, the covariance matrix $\mathsf{var}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{X}\big)$ is also a diagonal matrix, i.e., the LSE of the regression coefficients are uncorrelated.

Evaluation of the effect of the original covariate

The effect of the original covariate on the response expectation is again quantified by a set of $k-1$ regression coefficients $\boldsymbol{\beta}^Z := \big(\beta_1, \ldots, \beta_{k-1}\big)^\top$. To evaluate a statistical significance of the effect of the original covariate on the response expectation we have to test the null hypothesis

$$\mathrm{H}_0: \ \boldsymbol{\beta}^Z = \mathbf{0}_{k-1}.$$

See Section 3.2 for a possible test procedure.

Interpretation of the regression coefficients

The single regression coefficients $\beta_1, \ldots, \beta_{k-1}$ do not usually have a direct reasonable interpretation.

Degree of a polynomial

Test on a subset of regression coefficients/test on submodels (were introduced in Sections 3.2 and 5.2) can again be used to infer on the degree of a polynomial in the regression function (7.5) in the same way as with the raw polynomials. The null hypothesis expressing, for $d < k$, belief that the regression function is a polynomial of degree $d-1$ corresponds to the null hypothesis

$$\mathrm{H}_0: \ \beta_d = 0 \ \ \& \ \ \ldots \ \ \& \ \ \beta_{k-1} = 0.$$

7.3.4 Regression splines

Basis splines

The advantage of a polynomial regression function introduced in Sections 7.3.2 and 7.3.3 is that it is *smooth* (have continuous derivatives of all orders) on the whole real line. Nevertheless, with the least squares estimation, each data point affects *globally* the fitted regression function. This often leads to undesirable boundary effects when the fitted regression function only poorly approximates the response expectation $\mathbb{E}\big(Y \,\big|\, Z = z\big)$ for the values of z being close to the boundaries of the covariate space $\mathcal{Z}$. This can be avoided with so-called *regression splines*.

Definition 7.2 Basis spline with distinct knots.

Let $d \in \mathbb{N}_0$ and $\boldsymbol{\lambda} = \big(\lambda_1, \ldots, \lambda_{d+2}\big)^\top \in \mathbb{R}^{d+2}$, where $-\infty < \lambda_1 < \cdots < \lambda_{d+2} < \infty$. The basis spline of degree d with distinct knots[7] *$\boldsymbol{\lambda}$ is such a function $B^d(z; \boldsymbol{\lambda})$, $z \in \mathbb{R}$ that*

(i) $B^d(z; \boldsymbol{\lambda}) = 0$, for $z \le \lambda_1$ and $z \ge \lambda_{d+2}$;

(ii) On each of the intervals $(\lambda_j, \lambda_{j+1})$, $j = 1, \ldots, d+1$, $B^d(\cdot; \boldsymbol{\lambda})$ is a polynomial of degree d;

(iii) $B^d(\cdot; \boldsymbol{\lambda})$ has continuous derivatives up to an order $d-1$ on $\mathbb{R}$.

Notes.

- The basis spline with distinct knots is *piecewise*[8] polynomial of degree d on $(\lambda_1, \lambda_{d+2})$.
- The polynomial pieces are connected smoothly (of order $d-1$) at inner knots $\lambda_2, \ldots, \lambda_{d+1}$.
- On the boundary (λ_1 and λ_{d+2}), the polynomial pieces are connected smoothly (of order $d-1$) with a constant zero.

Definition 7.3 Basis spline with coincident left boundary knots.

Let $d \in \mathbb{N}_0$, $1 < r < d+2$ and $\boldsymbol{\lambda} = \big(\lambda_1, \ldots, \lambda_{d+2}\big)^\top \in \mathbb{R}^{d+2}$, where $-\infty < \lambda_1 = \cdots = \lambda_r < \cdots < \lambda_{d+2} < \infty$. The basis spline of degree d with r coincident left boundary knots[9] *$\boldsymbol{\lambda}$ is such a function $B^d(z; \boldsymbol{\lambda})$, $z \in \mathbb{R}$ that*

[7] *bazický spline [čti splajn] stupně d se vzájemně různými uzly* [8] *po částech* [9] *bazický spline stupně d s r překrývajícími se levými uzly*

(i) $B^d(z;\, \boldsymbol{\lambda}) = 0$, *for* $z \leq \lambda_r$ *and* $z \geq \lambda_{d+2}$;

(ii) On each of the intervals $(\lambda_j,\, \lambda_{j+1})$, $j = r,\, \ldots,\, d+1$, $B^d(\cdot;\, \boldsymbol{\lambda})$ *is a polynomial of degree* d;

(iii) $B^d(\cdot;\, \boldsymbol{\lambda})$ *has continuous derivatives up to an order* $d-1$ *on* $(\lambda_r,\, \infty)$;

(iv) $B^d(\cdot;\, \boldsymbol{\lambda})$ *has continuous derivatives up to an order* $d-r$ *in* λ_r.

Notes.

- The only qualitative difference between the basis spline with coincident left boundary knots and the basis spline with distinct knots is the fact that the basis spline with coincident left boundary knots is at the left boundary smooth of order only $d-r$ compared to order $d-1$ in case of the basis spline with distinct knots.
- By mirroring Definition 7.3 to the right boundary, basis spline with coincident right boundary knots is defined.

Basis B-splines

There are many ways on how to construct the basis splines that satisfy conditions of Definitions 7.2 and 7.3, see *Fundamentals of Numerical Mathematics (NMNM201)* course. In statistics, so called *B-splines* have proved to be extremely useful for regression purposes. It goes beyond the scope of this lecture to explain in detail their construction which is fully covered by two landmark books de Boor (1978, 2001); Dierckx (1993) or in a compact way, e.g., by a paper Eilers and Marx (1996). For the purpose of this lecture it is assumed that a routine is available to construct the basis B-splines of given degree with given knots (e.g., the R function `bs` from the recommended package splines).

An important property of the basis B-splines is that they are positive inside their support interval (general basis splines can also attain negative values inside the support interval). That is, if $\boldsymbol{\lambda} = \left(\lambda_1,\, \ldots,\, \lambda_{d+2}\right)^\top$ is a set of knots (either distinct or coincident left or right) and $B^d(\cdot,\, \boldsymbol{\lambda})$ is a basis B-spline of degree d built above the knots $\boldsymbol{\lambda}$ then

$$B^d(z,\, \boldsymbol{\lambda}) > 0, \qquad \lambda_1 < z < \lambda_{d+2},$$
$$B^d(z,\, \boldsymbol{\lambda}) = 0, \qquad z \leq \lambda_1,\ z \geq \lambda_{d+2}.$$

Spline basis

Definition 7.4 Spline basis.

Let $d \in \mathbb{N}_0$, $k \geq d+1$ *and* $\boldsymbol{\lambda} = \left(\lambda_1,\, \ldots,\, \lambda_{k-d+1}\right)^\top \in \mathbb{R}^{k-d+1}$, *where* $-\infty < \lambda_1 < \ldots < \lambda_{k-d+1} < \infty$. *The* spline basis[10] *of degree* d *with knots* $\boldsymbol{\lambda}$ *is a set of basis splines* $B_1,\, \ldots,\, B_k$, *where for* $z \in \mathbb{R}$,

$$B_1(z) = B^d(z;\, \underbrace{\lambda_1,\, \ldots,\, \lambda_1}_{(d+1)\times},\, \lambda_2),$$

$$B_2(z) = B^d(z;\, \underbrace{\lambda_1,\, \ldots,\, \lambda_1}_{d\times},\, \lambda_2,\, \lambda_3),$$

$$\vdots$$

[10] *splinová báze*

$$
\begin{aligned}
B_d(z) &= B^d(z;\, \underbrace{\lambda_1,\, \lambda_1}_{2\times},\, \lambda_2,\, \ldots,\, \lambda_{d+1}),\\
B_{d+1}(z) &= B^d(z;\, \lambda_1,\, \lambda_2,\, \ldots,\, \lambda_{d+2}),\\
B_{d+2}(z) &= B^d(z;\, \lambda_2,\, \ldots,\, \lambda_{d+3}),\\
&\vdots\\
B_{k-d}(z) &= B^d(z;\, \lambda_{k-2d},\, \ldots,\, \lambda_{k-d+1}),\\
B_{k-d+1}(z) &= B^d(z;\, \lambda_{k-2d+1},\, \ldots,\, \underbrace{\lambda_{k-d+1},\, \lambda_{k-d+1}}_{2\times}),\\
&\vdots\\
B_{k-1}(z) &= B^d(z;\, \lambda_{k-d-1},\, \lambda_{k-d}\ldots,\, \underbrace{\lambda_{k-d+1},\, \ldots,\, \lambda_{k-d+1}}_{d\times}),\\
B_k(z) &= B^d(z;\, \lambda_{k-d}\ldots,\, \underbrace{\lambda_{k-d+1},\, \ldots,\, \lambda_{k-d+1}}_{(d+1)\times}).
\end{aligned}
$$

End of Lecture #5 (13/10/2016)
Start of Lecture #7 (20/10/2016)

Properties of the B-spline basis

If $k \geq d+1$, a set of knots $\boldsymbol{\lambda} = \left(\lambda_1,\, \ldots,\, \lambda_{k-d+1}\right)^\top$, $-\infty < \lambda_1 < \ldots < \lambda_{k-d+1} < \infty$ is given and $B_1,\, \ldots,\, B_k$ is the spline basis of degree d with knots $\boldsymbol{\lambda}$ composed of basis B-splines $k \geq d+1$, a set of knots $\boldsymbol{\lambda} = \left(\lambda_1,\, \ldots,\, \lambda_{k-d+1}\right)^\top$, $-\infty < \lambda_1 < \ldots < \lambda_{k-d+1} < \infty$ is given and $B_1,\, \ldots,\, B_k$ is the spline basis of degree d with knots $\boldsymbol{\lambda}$ composed of basis B-splines then

(a) $$\sum_{j=1}^{k} B_j(z) = 1 \qquad \text{for all } z \in \left(\lambda_1,\, \lambda_{k-d+1}\right); \tag{7.8}$$

(b) for each $m \leq d$ there exist a set of coefficients $\gamma_1^m,\, \ldots,\, \gamma_k^m$ such that

$$\sum_{j=1}^{k} \gamma_j^m\, B_j(z) \text{ is on } (\lambda_1,\, \lambda_{k-d+1}) \text{ a polynomial in } z \text{ of degree } m. \tag{7.9}$$

Regression spline

It will now be assumed that the covariate space is a bounded interval, i.e., $\mathcal{Z} = \left(z_{min},\, z_{max}\right)$, $-\infty < z_{min} < z_{max} < \infty$. The regression function that exploits the regression splines is

$$m(z) = \beta_1\, B_1(z) \,+\, \cdots \,+\, \beta_k\, B_k(z), \quad z \in \mathcal{Z}, \tag{7.10}$$

where $B_1,\, \ldots,\, B_k$ is the spline basis of chosen degree $d \in \mathbb{N}_0$ composed of basis B-splines built above a set of chosen knots $\boldsymbol{\lambda} = \left(\lambda_1,\, \ldots,\, \lambda_{k-d+1}\right)^\top$, $z_{min} = \lambda_1 < \ldots < \lambda_{k-d+1} = z_{max}$. The corresponding reparameterizing matrix coincided with the model matrix and is

$$\mathbb{X} = \mathbb{S} = \begin{pmatrix} B_1(Z_1) & \ldots & B_k(Z_1) \\ \vdots & \vdots & \vdots \\ B_1(Z_n) & \ldots & B_k(Z_n) \end{pmatrix} =: \mathbb{B}. \tag{7.11}$$

Notes.

- It follows from (7.8) that
$$\mathbf{1}_n \in \mathcal{M}(\mathbb{B}).$$
This is also the reason why we do not explicitely include the intercept term in the regression function since it is implicitely included in the regression space. Due to clarity of notation, the regression coefficients are now indexed from 1 to k. That is, the vector of regression coefficients is $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$.
- It also follows from (7.9) that for any $m \leq d$, a linear model with the regression function based on either raw or orthonormal polynomials of degree m is a submodel of the linear model with the regression function given by a regression spline and the model matrix $\mathbb{B}$.
- With $d = 0$, the regression spline (7.10) is simply a piecewise constant function.
- In practice, not much attention is paid to the choice of the degree d of the regression spline. Usually $d = 2$ (quadratic spline) or $d = 3$ (cubic spline) is used which provides continuous first or second derivatives, respectively, of the regression function inside the covariate domain $\mathcal{Z}$.
- On the other hand, the placement of knots (selection of the values of $\lambda_1, \ldots, \lambda_{k-d+1}$) is quite important to obtain the regression function that sufficiently well approximates the response expectations $\mathbb{E}(Y \mid Z = z)$, $z \in \mathcal{Z}$. Unfortunately, only relatively ad-hoc methods towards selection of the knots will be demonstrated during this lecture as profound methods of the knots selection go far beyond the scope of this course.

Advantages of the regression splines compared to raw/orthogonal polynomials

- Each data point influences the LSE of the regression coefficients and hence the fitted regression function only *locally*. Indeed, only the LSE of those regression coefficients that correspond to the basis splines whose supports cover a specific data point are influenced by those data points.
- Regression splines of even a low degree d (2 or 3) are, with a suitable choice of knots, able to approximate sufficiently well even functions with a highly variable curvature and that globally on the whole interval $\mathcal{Z}$.

Evaluation of the effect of the original covariate

To evaluate a statistical significance of the effect of the original covariate on the response expectation we have to test the null hypothesis

$$\mathrm{H}_0 : \ \beta_1 = \cdots = \beta_k.$$

Due to the property (7.8), this null hypothesis corresponds to assuming that $\mathbb{E}(\boldsymbol{Y} \mid \mathbb{Z}) \in \mathcal{M}(\mathbf{1}_n) \subset \mathcal{M}(\mathbb{B})$. Consequently, it is possible to use a test on submodel that was introduced in Section 5.1 to test the above null hypothesis.

Interpretation of the regression coefficients

The single regression coefficients $\beta_1, \ldots, \beta_k$ do not usually have a direct reasonable interpretation.

7.4 Categorical covariate

In this Section, it is assumed that $Z_i \in \mathcal{Z}$, $i = 1, \ldots, n$, are values of a *categorical* covariate. That is, the covariate sample space $\mathcal{Z}$ is finite and its elements are only understood as *labels*. Without loss of generality, we will use, unless stated otherwise, a simple sequence $1, \ldots, G$ for those labels, i.e.,

$$\mathcal{Z} = \{1, \ldots, G\}.$$

Unless explicitely stated (in Section 7.4.4), even the ordering of the labels $1 < \cdots < G$ will not be used for any but notational purposes and the methodology described below is then suitable for both *nominal* and *ordinal* categorical covariates.

The regression function, $m : \mathcal{Z} \longrightarrow \mathbb{R}$ is now a function defined on a finite set aiming in parameterizing just G (conditional) response expectations $\mathbb{E}(Y \mid Z = 1), \ldots, \mathbb{E}(Y \mid Z = G)$. For some clarity in notation, we will also use symbols $m_1, \ldots, m_G$ for those expectations, i.e.,

$$\begin{array}{rcll} m(1) & = & \mathbb{E}(Y \mid Z = 1) & =: \quad m_1, \\ & \vdots & & \vdots \\ m(G) & = & \mathbb{E}(Y \mid Z = G) & =: \quad m_G. \end{array}$$

Notation and terminology (One-way classified group means).

Since a categorical covariate often indicates pertinence to one of G groups, we will call $m_1, \ldots, m_G$ as *group means*[11] or *one-way classified group means*. A vector

$$\boldsymbol{m} = (m_1, \ldots, m_G)^\top$$

will be called a vector of *group means*,[12] or a vector of *one-way classified group means*.

Note. Perhaps appealing simple regression function of the form

$$m(z) = \beta_0 + \beta_1 z, \qquad z = 1, \ldots, G,$$

is in most cases fully inappropriate. First, it orders ad-hoc the group means to form a monotone sequence (increasing if $\beta_1 > 0$, decreasing if $\beta_1 < 0$). Second, it ad-hoc assumes a linear relationship between the group means. Both those properties also depend on the ordering or even the values of the labels ($1, \ldots, G$ in our case) assigned to the G categories at hand. With a *nominal* categorical covariate, none of it is justifiable, with an *ordinal* categorical covariate, such assumptions should, at least, never be taken for granted and used without proper verification.

7.4.1 Link to a G-sample problem

For following considerations, we will additionally assume (again without loss of generality) that the data $(Y_i, Z_i)^\top$, $i = 1, \ldots, n$, are sorted according to the covariate values $Z_1, \ldots, Z_n$. Furthermore, we will also exchangeably use a double subscript with the response where the first subscript

[11] *skupinové střední hodnoty* [12] *vektor skupinových středních hodnot*

will indicate the covariate value, i.e.,

$$\mathbb{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_{n_1} \\ - - - \\ \vdots \\ - - - \\ Z_{n-n_G+1} \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -- \\ \vdots \\ -- \\ G \\ \vdots \\ G \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} n_1\text{-times} \\ \\ \left.\vphantom{\begin{matrix} G \\ \vdots \\ G \end{matrix}}\right\} n_G\text{-times} \end{matrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ - - - - \\ \vdots \\ - - - - \\ Y_{n-n_G+1} \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ - - - \\ \vdots \\ - - - \\ Y_{G,1} \\ \vdots \\ Y_{G,n_G} \end{pmatrix}.$$

Finally, let

$$\boldsymbol{Y}_g = \big(Y_{g,1},\, \ldots,\, Y_{g,n_g}\big)^\top, \qquad g = 1, \ldots, G,$$

denote a subvector of the response vector that corresponds to observations with the covariate value being equal to g. That is,

$$\boldsymbol{Y} = \big(Y_1,\, \ldots,\, Y_n\big)^\top = \big(\boldsymbol{Y}_1^\top,\, \ldots,\, \boldsymbol{Y}_G^\top\big)^\top.$$

Suppose now that the data $\big(Y_i,\, Z_i\big)^\top \overset{\text{i.i.d.}}{\sim} \big(Y,\, Z\big)^\top$ and a linear model holds with $\mathbb{E}\big(Y \,\big|\, Z = g\big) = m_g$, $\mathsf{var}\big(Y \,\big|\, Z = g\big) = \sigma^2$, $g = 1, \ldots, G$. In that case, for given $g \in \{1, \ldots, G\}$, the random variables $Y_{g,1}, \ldots, Y_{g,n_G}$ (elements of the vector $\boldsymbol{Y}_g$) are i.i.d. from a distribution of $Y \,|\, Z = g$ whose mean is m_g and the variance is σ^2. That is,

$$\begin{aligned} Y_{1,1}, \ldots, Y_{1,n_1} &\overset{\text{i.i.d.}}{\sim} \big(m_1,\, \sigma^2\big), \\ &\vdots \\ Y_{G,1}, \ldots, Y_{G,n_G} &\overset{\text{i.i.d.}}{\sim} \big(m_G,\, \sigma^2\big). \end{aligned} \tag{7.12}$$

Having all elements of the response random vector independent, (7.12) describes a classical G sample problem where the samples are assumed to be homoscedastic (having the same variance).

Notes.

- If the covariates $Z_1, \ldots, Z_G$ are random then also $n_1, \ldots, n_G$ are random.
- In the following, it is always assumed that $n_1 > 0, \ldots, n_G > 0$ (almost surely).

7.4.2 Linear model parameterization of one-way classified group means

As usual, let $\boldsymbol{\mu}$ be the (conditional) response expectation, i.e.,

$$\mathbb{E}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big) = \boldsymbol{\mu} := \begin{pmatrix} \mu_{1,1} \\ \vdots \\ \mu_{1,n_1} \\ -- \\ \vdots \\ -- \\ \mu_{G,1} \\ \vdots \\ \mu_{G,n_G} \end{pmatrix} = \begin{pmatrix} m_1 \\ \vdots \\ m_1 \\ -- \\ \vdots \\ -- \\ m_G \\ \vdots \\ m_G \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{-times} \\ \\ \\ \\ \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} m_1\,\mathbf{1}_{n_1} \\ \vdots \\ m_G\,\mathbf{1}_{n_G} \end{pmatrix}. \tag{7.13}$$

Notation and terminology *(Regression space of a categorical covariate).*

A vector space

$$\left\{ \begin{pmatrix} m_1\,\mathbf{1}_{n_1} \\ \vdots \\ m_G\,\mathbf{1}_{n_G} \end{pmatrix} : \; m_1, \ldots, m_G \in \mathbb{R} \right\} \subseteq \mathbb{R}^n$$

will be called *the regression space of a categorical covariate (factor)* with levels frequencies $n_1, \ldots, n_G$ and will be denoted as $\mathcal{M}_F(n_1, \ldots, n_G)$.

Note. Obviously, with $n_1 > 0, \ldots, n_G > 0$, a vector dimension of $\mathcal{M}_F(n_1, \ldots, n_G)$ is equal to G and a possible (orthogonal) vector basis is

$$\mathbb{Q} = \begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \vdots & \vdots \\ 1 & \ldots & 0 \\ - & - & -- \\ \vdots & \vdots & \vdots \\ - & - & -- \\ 0 & \ldots & 1 \\ \vdots & \vdots & \vdots \\ 0 & \ldots & 1 \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{-times} \\ \\ \\ \\ \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \big(1, \ldots, 0\big) \\ \vdots \\ \mathbf{1}_{n_G} \otimes \big(0, \ldots, 1\big) \end{pmatrix}. \tag{7.14}$$

When using the linear model, we are trying to allow for expressing the response expectation $\boldsymbol{\mu}$, i.e., a vector from $\mathcal{M}_F(n_1, \ldots, n_G)$ as a linear combination of columns of a suitable $n \times k$ matrix $\mathbb{X}$, i.e., as

$$\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^k.$$

It is obvious that any model matrix that parameterizes the regression space $\mathcal{M}_F(n_1, \ldots, n_G)$

must have at least G columns, i.e., $k \geq G$ and must be of the type

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_1^\top \\ -- \\ \vdots \\ -- \\ \boldsymbol{x}_G^\top \\ \vdots \\ \boldsymbol{x}_G^\top \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_1^\top \end{matrix}}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\vphantom{\begin{matrix} \boldsymbol{x}_G^\top \\ \vdots \\ \boldsymbol{x}_G^\top \end{matrix}}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{x}_1^\top \\ \vdots \\ \mathbf{1}_{n_G} \otimes \boldsymbol{x}_G^\top \end{pmatrix}, \tag{7.15}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_G \in \mathbb{R}^k$ are suitable vectors.

Problem of parameterizing a categorical covariate with G levels thus simplifies into selecting a $G \times k$ matrix $\widetilde{\mathbb{X}}$ such that

$$\widetilde{\mathbb{X}} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_G^\top \end{pmatrix}.$$

Clearly,

$$\mathsf{rank}\big(\mathbb{X}\big) = \mathsf{rank}\big(\widetilde{\mathbb{X}}\big).$$

Hence to be able to parameterize the regression space $\mathcal{M}_F(n_1, \ldots, n_G)$ which has a vector dimension of G, the matrix $\widetilde{\mathbb{X}}$ must satisfy

$$\mathsf{rank}\big(\widetilde{\mathbb{X}}\big) = G.$$

The group means then depend on a vector $\boldsymbol{\beta} = \big(\beta_0, \ldots, \beta_{k-1}\big)^\top$ of the regression coefficients as

$$\begin{aligned} m_g &= \boldsymbol{x}_g^\top \boldsymbol{\beta}, \qquad g = 1, \ldots, G, \\ \boldsymbol{m} &= \widetilde{\mathbb{X}} \boldsymbol{\beta}. \end{aligned}$$

A possible (full-rank) linear model parameterization of regression space of a categorical covariate uses matrix $\mathbb{Q}$ from (7.14) as a model matrix $\mathbb{X}$. In that case, $\widetilde{\mathbb{X}} = \mathbf{I}_G$ and we have

$$\begin{aligned} \boldsymbol{\mu} &= \mathbb{Q}\,\boldsymbol{\beta}, \\ \boldsymbol{m} &= \boldsymbol{\beta}. \end{aligned} \tag{7.16}$$

Even though parameterization (7.16) seems appealing since the regression coefficients are directly equal to the group means, it is only rarely considered in practice for reasons that will become clear later on. Still, it is useful for some of theoretical derivations.

7.4.3 ANOVA parameterization of one-way classified group means

In practice and especially in the area of designed experiments, the group means are parameterized as

$$\begin{aligned} m_g &= \alpha_0 + \alpha_g, \qquad g = 1, \ldots, G, \\ \boldsymbol{m} = \big(\mathbf{1}_G,\, \mathbf{I}_G\big)\boldsymbol{\alpha} &= \alpha_0\, \mathbf{1}_G \,+\, \boldsymbol{\alpha}^Z, \end{aligned} \tag{7.17}$$

where $\boldsymbol{\alpha} = \big(\alpha_0,\, \alpha_1,\, \ldots,\, \alpha_G\big)^\top$ is a vector of regression coefficients and $\boldsymbol{\alpha}^Z = \big(\alpha_1,\, \ldots,\, \alpha_G\big)^\top$ is its non-intercept subvector. That is, the matrix $\widetilde{\mathbb{X}}$ is

$$\widetilde{\mathbb{X}} = \big(\mathbf{1}_G,\, \mathbf{I}_G\big).$$

The model matrix is then

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ldots & 0 \\ - & - & - & - \\ \vdots & \vdots & \vdots & \vdots \\ - & - & - & - \\ 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_1\text{-times} \\ \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \big(1,\, 1,\, \ldots,\, 0\big) \\ \vdots \\ \mathbf{1}_{n_G} \otimes \big(1,\, 0,\, \ldots,\, 1\big) \end{pmatrix}, \tag{7.18}$$

which has $G + 1$ columns but its rank is G (as required). That is, the linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\alpha},\, \sigma^2\mathbf{I}_n\big)$ is less-than-full rank. In other words, for given $\boldsymbol{\mu} \in \mathcal{M}\big(\mathbb{X}\big) = \mathcal{M}_F(n_1,\, \ldots,\, n_G)$, there exists infinitely many vectors $\boldsymbol{\alpha} \in \mathbb{R}^{G+1}$ such that $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\alpha}$. Consequently, also a solution to related normal equations is not unique. Nevertheless, unique solution can be obtained if suitable *indetifying constraints*[13] are imposed on the vector of regression coefficients $\boldsymbol{\alpha}$.

Terminology *(Effects of a categorical covariate).*

Values of $\alpha_1,\, \ldots,\, \alpha_G$ (a vector $\boldsymbol{\alpha}^Z$) are called *effects of a categorical covariate*.

Note. Effects of a categorical covariate are not unique. Hence their interpretation depends on chosen identifying constraints.

Identification in less-than-full-rank linear model

In the following, a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r$ will be assumed (in our general notation), where $r < k$. We shall consider linear constraints on a vector of regression coefficients, i.e., constraints of the type

$$\mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m,$$

where $\mathbb{A}$ is an $m \times k$ matrix.

[13] *identifikační omezení*

Definition 7.5 Identifying constraints.

We say that a constraint

$$\mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m$$

identifies *a vector* $\boldsymbol{\beta}$ *in a linear model* $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *if and only if for each* $\boldsymbol{\mu} \in \mathcal{M}\big(\mathbb{X}\big)$ *there exists* only one *vector* $\boldsymbol{\beta}$ *which satisfies at the same time*

$$\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta} \quad \textit{and} \quad \mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m.$$

Note. If a matrix $\mathbb{A}$ determines the identifying constraints, then, due to Theorem 2.5 (least squares and normal equations), it also uniquely determines the solution to normal equations. That is, there is a unique solution $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}$ that jointly solves linear systems

$$\mathbb{X}^\top\mathbb{X}\boldsymbol{b} = \mathbb{X}^\top\boldsymbol{Y}, \quad \mathbb{A}\boldsymbol{b} = \mathbf{0}_m,$$

or written differently, there is a unique solution to a linear system

$$\begin{pmatrix} \mathbb{X}^\top\mathbb{X} \\ \mathbb{A} \end{pmatrix} \boldsymbol{b} = \begin{pmatrix} \mathbb{X}^\top\boldsymbol{Y} \\ \mathbf{0}_m \end{pmatrix}.$$

The question is now, what are the conditions for a matrix $\mathbb{A}$ to determine an identifying constraint. Remember (Theorem 2.7): If a matrix $\mathbb{L}_{m\times k}$ satisfies $\mathcal{M}\big(\mathbb{L}^\top\big) \subset \mathcal{M}\big(\mathbb{X}^\top\big)$ then a parameter vector $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ is estimable which also means that for all real vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ the following holds:

$$\mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2 \quad \Longrightarrow \quad \mathbb{L}\boldsymbol{\beta}_1 = \mathbb{L}\boldsymbol{\beta}_2.$$

That is, if two different solutions of normal equations are taken and one of them satisfies the constraint then do the both. It was also shown in Section 5.3 that if further $\mathbb{L}$ has linearly independent rows then a set of linear constraints $\mathbb{L}\boldsymbol{\beta} = \mathbf{0}$ determines a so called submodel (Lemma 5.4). It follows from above that for identification, we cannot use such a matrix $\mathbb{L}$ for identification.

Theorem 7.1 Scheffé on identification in a linear model.

Constraint $\mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m$ *with a real matrix* $\mathbb{A}_{m\times k}$ *identifies a vector* $\boldsymbol{\beta}$ *in a linear model* $\boldsymbol{Y} \,|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r < k < n$ *if and only if*

$$\mathcal{M}\big(\mathbb{A}^\top\big) \cap \mathcal{M}\big(\mathbb{X}^\top\big) = \{\mathbf{0}\},$$
$$\mathsf{rank}(\mathbb{X}) + \mathsf{rank}(\mathbb{A}) = k.$$

Proof. We have to show that, for any $\boldsymbol{\mu} \in \mathcal{M}\big(\mathbb{X}\big)$, the conditions stated in the theorem are equivalent to existence of the unique solution to a linear system $\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ that satisfies $\mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m$.

Existence of the solution

$$\Leftrightarrow \quad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \text{ there exists a vector } \boldsymbol{\beta} \in \mathbb{R}^k \text{ such that}$$
$$\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu} \quad \& \quad \mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m.$$

$$\Leftrightarrow \quad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \text{ there exists a vector } \boldsymbol{\beta} \in \mathbb{R}^k \text{ such that}$$
$$\underbrace{\begin{pmatrix} \mathbb{X}_{n\times k} \\ \mathbb{A}_{m\times k} \end{pmatrix}}_{\mathbb{D}} \boldsymbol{\beta} \;=\; \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0}_m \end{pmatrix}.$$

$$\Leftrightarrow \quad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \text{ there exists a vector } \boldsymbol{\beta} \in \mathbb{R}^k \text{ such that}$$
$$\mathbb{D}\,\boldsymbol{\beta} \;=\; \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0}_m \end{pmatrix}.$$

$$\Leftrightarrow \quad \Big\{ \big(\boldsymbol{\mu}^\top,\, \mathbf{0}_m^\top\big)^\top,\; \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \Big\} \subseteq \mathcal{M}(\mathbb{D}).$$

$$\Leftrightarrow \quad \Big\{ \mathcal{M}(\mathbb{D}) \Big\}^\perp \subseteq \Big\{ \big(\boldsymbol{\mu}^\top,\, \mathbf{0}_m^\top\big)^\top,\; \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \Big\}^\perp.$$

$$\Leftrightarrow \quad \Big\{ \forall \boldsymbol{v}_1 \in \mathbb{R}^n,\, \boldsymbol{v}_2 \in \mathbb{R}^m \quad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X})$$
$$\big(\boldsymbol{v}_1^\top,\, \boldsymbol{v}_2^\top\big)\,\mathbb{D} = \mathbf{0}_k^\top \quad \Rightarrow \quad \big(\boldsymbol{v}_1^\top,\, \boldsymbol{v}_2^\top\big) \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0}_m \end{pmatrix} = 0 \Big\}.$$

$$\Leftrightarrow \quad \Big\{ \forall \boldsymbol{v}_1 \in \mathbb{R}^n,\, \boldsymbol{v}_2 \in \mathbb{R}^m \quad \forall \boldsymbol{\beta} \in \mathbb{R}^k$$
$$\big(\boldsymbol{v}_1^\top,\, \boldsymbol{v}_2^\top\big)\,\mathbb{D} = \mathbf{0}_k^\top \quad \Rightarrow \quad \big(\boldsymbol{v}_1^\top,\, \boldsymbol{v}_2^\top\big) \begin{pmatrix} \mathbb{X}\boldsymbol{\beta} \\ \mathbf{0}_m \end{pmatrix} = 0 \Big\}.$$

$$\Leftrightarrow \quad \Big\{ \forall \boldsymbol{v}_1 \in \mathbb{R}^n,\, \boldsymbol{v}_2 \in \mathbb{R}^m \quad \forall \boldsymbol{\beta} \in \mathbb{R}^k$$
$$\boldsymbol{v}_1^\top \mathbb{X} \;=\; -\boldsymbol{v}_2^\top \mathbb{A} \quad \Rightarrow \quad \boldsymbol{v}_1^\top \mathbb{X}\boldsymbol{\beta} = 0 \Big\}.$$

$$\Leftrightarrow \quad \Big\{ \forall \boldsymbol{v}_1 \in \mathbb{R}^n,\, \boldsymbol{v}_2 \in \mathbb{R}^m \quad \boldsymbol{v}_1^\top \mathbb{X} \;=\; -\boldsymbol{v}_2^\top \mathbb{A} \quad \Rightarrow \quad \boldsymbol{v}_1^\top \mathbb{X} = \mathbf{0}_k^\top \Big\}.$$

$$\Leftrightarrow \quad \Big\{ \forall \boldsymbol{v}_1 \in \mathbb{R}^n,\, \boldsymbol{v}_2 \in \mathbb{R}^m \quad \underbrace{\mathbb{X}^\top \boldsymbol{v}_1}_{\boldsymbol{u}} = -\mathbb{A}^\top \boldsymbol{v}_2 \quad \Rightarrow \quad \mathbb{X}^\top \boldsymbol{v}_1 = \mathbf{0}_k \Big\}.$$

$$\Leftrightarrow \quad \Big\{ \forall \boldsymbol{u} \in \mathbb{R}^k \quad \boldsymbol{u} \in \mathcal{M}(\mathbb{X}^\top) \,\cap\, \mathcal{M}(\mathbb{A}^\top) \quad \Rightarrow \quad \boldsymbol{u} = \mathbf{0}_k \Big\}.$$

$$\Leftrightarrow \quad \mathcal{M}(\mathbb{A}^\top) \cap \mathcal{M}(\mathbb{X}^\top) \;=\; \{\mathbf{0}\}.$$

Uniqueness of the solution

$$\Leftrightarrow \quad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \text{ there exists a unique vector } \boldsymbol{\beta} \in \mathbb{R}^k \text{ such that}$$
$$\mathbb{X}\boldsymbol{\beta} = \boldsymbol{\mu} \quad \& \quad \mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m.$$

$$\Leftrightarrow \quad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbb{X}) \text{ there exists a unique vector } \boldsymbol{\beta} \in \mathbb{R}^k \text{ such that}$$

$$\underbrace{\begin{pmatrix}\mathbb{X}_{n\times k}\\ \mathbb{A}_{m\times k}\end{pmatrix}}_{\mathbb{D}}\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}_m\end{pmatrix}.$$

$\Leftrightarrow$ $\mathsf{rank}(\mathbb{D}) = k$.

$\Leftrightarrow$ $\mathbb{A}$ has rows such that $\dim\left\{\mathcal{M}\left(\mathbb{A}^\top\right)\right\} = k - r$ (since $\mathsf{rank}(\mathbb{X}) = r$) and all rows in $\mathbb{A}$ are linearly independent with rows in $\mathbb{X}$.

$\Leftrightarrow$ $\mathsf{rank}(\mathbb{A}) = k - r$ (since we already have a condition $\mathcal{M}\left(\mathbb{X}^\top\right) \cap \mathcal{M}\left(\mathbb{A}^\top\right) = \{\mathbf{0}\}$ needed for existence of the solution).

❑

End of Lecture #7 (20/10/2016)
Start of Lecture #9 (27/10/2016)

Notes.

1. Matrix $\mathbb{A}_{m\times k}$ used for identification must satisfy $\mathsf{rank}(\mathbb{A}) = k - r$. In practice, the number of identifying constraints (the number of rows of the matrix $\mathbb{A}$) is usually the lowest possible, i.e., $m = k - r$.

2. Theorem 7.1 further states that the matrix $\mathbb{A}$ must be such that a vector parameter $\boldsymbol{\theta} = \mathbb{A}\boldsymbol{\beta}$ is *not* estimable in a given model.

3. In practice, a vector $\boldsymbol{\mu}$, for which we look for a unique $\boldsymbol{\beta}$ such that $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$, $\mathbb{A}\boldsymbol{\beta} = \mathbf{0}_m$ is equal to the vector of fitted values $\widehat{\boldsymbol{Y}}$. That is, we look for a unique $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^k$ (that, since being unique, can be considered as the LSE of the regression coefficients) such that

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{Y}} \qquad \& \qquad \mathbb{A}\widehat{\boldsymbol{\beta}} = \mathbf{0}_m.$$

   By Theorem 2.5 (Least squares and normal equations), $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$ if and only if $\widehat{\boldsymbol{\beta}}$ solves normal equations $\mathbb{X}^\top\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\top\boldsymbol{Y}$.

   Suppose now that $\mathsf{rank}\left(\mathbb{A}\right) = m = k - r$, i.e., the regression parameters are identified by a set of $m = k - r$ linearly independent linear constraints. To get $\widehat{\boldsymbol{\beta}}$, we have to solve a linear system

$$\mathbb{X}^\top\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\top\boldsymbol{Y},$$
$$\mathbb{A}\widehat{\boldsymbol{\beta}} = \mathbf{0}_m,$$

   which can be solved by solving

$$\mathbb{X}^\top\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\top\boldsymbol{Y},$$
$$\mathbb{A}^\top\mathbb{A}\widehat{\boldsymbol{\beta}} = \mathbf{0}_m,$$

   or using a linear system

$$\left(\mathbb{X}^\top\mathbb{X} + \mathbb{A}^\top\mathbb{A}\right)\widehat{\boldsymbol{\beta}} = \mathbb{X}^\top\boldsymbol{Y},$$

   which written differently is

$$\mathbb{D}^\top\mathbb{D}\widehat{\boldsymbol{\beta}} = \mathbb{X}^\top\boldsymbol{Y},$$

with

$$\mathbb{D} = \begin{pmatrix} \mathbb{X} \\ \mathbb{A} \end{pmatrix}.$$

Matrix $\mathbb{D}^\top\mathbb{D}$ is now an *invertible* $k \times k$ matrix and hence the unique solution is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbb{D}^\top\mathbb{D}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y}.$$

Identification in a one-way ANOVA model

As example of use of Scheffé's Theorem 7.1, consider a model matrix $\mathbb{X}$ given by (7.18) that provides an ANOVA parameterization of a single categorical covariate, i.e., a linear model for the one-way classified group means parameterized as

$$m_g = \alpha_0 + \alpha_g, \qquad g = 1, \ldots, G.$$

We have $\mathsf{rank}\left(\mathbb{X}_{n\times(G+1)}\right) = G$ with a vector $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_1, \ldots, \alpha_G\right)^\top$ of the regression coefficients. The smallest matrix $\mathbb{A}_{m\times(G+1)}$ that identifies $\boldsymbol{\alpha}$ with respect to the regression space $\mathcal{M}\left(\mathbb{X}\right) = \mathcal{M}_F(n_1, \ldots, n_G)$ is hence a non-zero matrix with $m = 1$ row, i.e.,

$$\mathbb{A} = \boldsymbol{a}^\top = \left(a_0, a_1, \ldots, a_G\right)^\top \neq \mathbf{0}_{G+1}$$

such that $\boldsymbol{a} \notin \mathcal{M}\left(\mathbb{X}^\top\right)$, i.e., such that $\theta = \boldsymbol{a}^\top\boldsymbol{\alpha}$ is not estimable in the linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\alpha}, \sigma^2\,\mathbf{I}_n\right)$.

It is seen from a structure of the matrix $\mathbb{X}$ given by (7.18) that

$$\boldsymbol{a} \in \mathcal{M}\left(\mathbb{X}^\top\right) \quad \Longleftrightarrow \quad \boldsymbol{a} = \left(\sum_{g=1}^{G} c_g,\ c_1, \ldots, c_G\right)^\top$$

for some $\boldsymbol{c} = \left(c_1, \ldots, c_G\right)^\top \in \mathbb{R}^G$, $\boldsymbol{c} \neq \mathbf{0}_G$. That is, for identification of $\boldsymbol{\alpha}$ in the linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\alpha}, \sigma^2\,\mathbf{I}_n\right)$ with the model matrix (7.18), we can use any vector $\boldsymbol{a} = \left(a_0, a_1, \ldots, a_G\right)^\top \neq \mathbf{0}_{G-1}$ that satisfy

$$a_0 \neq \sum_{g=1}^{G} a_g.$$

Commonly used identifying constraints include:

Sum constraint:

$$\mathbb{A}_1 = \boldsymbol{a}_1^\top = \left(0, 1, \ldots, 1\right)^\top \quad \Longleftrightarrow \quad \sum_{g=1}^{G} \alpha_g = 0$$

that implies the following interpretation of the model parameters:

$$\begin{aligned} \alpha_0 &= \frac{1}{G}\sum_{g=1}^{G} m_g \qquad =: \overline{m}, \\ \alpha_1 &= m_1 - \overline{m}, \\ &\vdots \\ \alpha_G &= m_G - \overline{m}. \end{aligned}$$

Weighted sum constraint:

$$\mathbb{A}_2 = \boldsymbol{a}_2^\top = \big(0,\, n_1,\, \ldots,\, n_G\big)^\top \quad \Longleftrightarrow \quad \sum_{g=1}^{G} n_g\, \alpha_g = 0$$

that implies

$$\begin{aligned} \alpha_0 &= \frac{1}{n}\sum_{g=1}^{G} n_g\, m_g \qquad =: \overline{m}_W, \\ \alpha_1 &= m_1 - \overline{m}_W, \\ &\vdots \\ \alpha_G &= m_G - \overline{m}_W. \end{aligned}$$

Reference group constraint $(l \in \{1,\, \ldots,\, G\})$**:**

$$\mathbb{A}_3 = \boldsymbol{a}_3^\top = \big(0,\, \underbrace{0, \ldots, 1, \ldots, 0}_{1 \text{ on } l\text{th place}}\big) \quad \Longleftrightarrow \quad \alpha_l = 0,$$

which corresponds to omitting one of the non-intercept columns in the model matrix $\mathbb{X}$ given by (7.18) and using the resulting full-rank parameterization. It implies

$$\begin{aligned} \alpha_0 &= m_l, \\ \alpha_1 &= m_1 - m_l, \\ &\vdots \\ \alpha_G &= m_G - m_l. \end{aligned}$$

No intercept:

$$\mathbb{A}_4 = \boldsymbol{a}_4^\top = \big(1,\, 0,\, \ldots,\, 0\big)^\top \qquad \Longleftrightarrow \qquad \alpha_0 = 0,$$

which corresponds to omitting the intercept column in the model matrix $\mathbb{X}$ given by (7.18) and using the full-rank parameterization with the matrix $\mathbb{Q}$ given by (7.14). That is,

$$\begin{aligned} \alpha_0 &= 0, \\ \alpha_1 &= m_1, \\ &\vdots \\ \alpha_G &= m_G. \end{aligned}$$

Note. Identifying constraints given by vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$ (sum, weighted sum and reference group constraint) correspond to one of commonly used full-rank parameterizations that will be introduced in Section 7.4.4 where we shall also discuss interpretation of the effects $\boldsymbol{\alpha}^Z = \big(\alpha_1,\, \ldots,\, \alpha_G\big)^\top$ if different identifying constraints are used.

7.4.4 Full-rank parameterization of one-way classified group means

In the following, we limit ourselves to *full-rank* parameterizations that involve an *intercept* column. That is, the model matrix will be an $n \times G$ matrix

$$\mathbb{X} = \begin{pmatrix} 1 & \boldsymbol{c}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{c}_1^\top \\ - & - & - \\ \vdots & \vdots \\ - & - & - \\ 1 & \boldsymbol{c}_G^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{c}_G^\top \end{pmatrix} \begin{matrix} \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \big(1,\, \boldsymbol{c}_1^\top\big) \\ \vdots \\ \mathbf{1}_{n_G} \otimes \big(1,\, \boldsymbol{c}_G^\top\big) \end{pmatrix},$$

where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_G \in \mathbb{R}^{G-1}$ are suitable vectors. In the following, let $\mathbb{C}$ be an $G \times (G-1)$ matrix with those vectors as rows, i.e.,

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix}.$$

A matrix $\widetilde{\mathbb{X}}$ is thus a $G \times G$ matrix

$$\widetilde{\mathbb{X}} = \big(\mathbf{1}_G,\, \mathbb{C}\big).$$

If $\boldsymbol{\beta} = \big(\beta_0, \ldots, \beta_{G-1}\big)^\top \in \mathbb{R}^G$ denote, as usual, a vector of regression coefficients, the group means $\boldsymbol{m}$ are parameterized as

$$\begin{aligned} m_g &= \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, \qquad g = 1, \ldots, G, \\ \boldsymbol{m} = \widetilde{\mathbb{X}}\boldsymbol{\beta} = \big(\mathbf{1}_G,\, \mathbb{C}\big)\boldsymbol{\beta} &= \beta_0\, \mathbf{1}_G + \mathbb{C}\boldsymbol{\beta}^Z, \end{aligned} \tag{7.19}$$

where $\boldsymbol{\beta}^Z = \big(\beta_1, \ldots, \beta_{G-1}\big)^\top$ is a non-intercept subvector of the regression coefficients. As we know,

$$\mathsf{rank}\big(\mathbb{X}\big) = \mathsf{rank}\big(\widetilde{\mathbb{X}}\big) = \mathsf{rank}\big((\mathbf{1}_G,\, \mathbb{C})\big).$$

Hence, to get the model matrix $\mathbb{X}$ of a full-rank ($\mathsf{rank}\big(\mathbb{X}\big) = G$), the matrix $\mathbb{C}$ must satisfy $\mathsf{rank}\big(\mathbb{C}\big) = G - 1$ and $\mathbf{1}_G \notin \mathcal{M}\big(\mathbb{C}\big)$. That is, the *columns* of $\mathbb{C}$ must be

(i) $(G-1)$ *linearly independent* vectors from $\mathbb{R}^G$;

(ii) being all linearly independent with a vector of ones $\mathbf{1}_G$.

Definition 7.6 Full-rank parameterization of a categorical covariate.

Full-rank parameterization of a categorical covariate with G levels ($G = card(\mathcal{Z})$) is a choice of the $G \times (G-1)$ matrix $\mathbb{C}$ that satisfies

$$\mathsf{rank}\big(\mathbb{C}\big) = G - 1, \qquad \mathbf{1}_G \notin \mathcal{M}\big(\mathbb{C}\big).$$

Terminology *((Pseudo)contrast matrix).*

Columns of matrix $\mathbb{C}$ are often chosen to form a set of $G-1$ contrasts from $\mathbb{R}^G$. In this case, we will call the matrix $\mathbb{C}$ as a *contrast matrix*.[14] In other cases, the matrix $\mathbb{C}$ will be called as a *pseudocontrast matrix*.[15]

Note. The (pseudo)contrast matrix $\mathbb{C}$ also determines parameterization of a categorical covariate according to Definition 7.1. Corresponding function $\boldsymbol{s}: \mathcal{Z} \longrightarrow \mathbb{R}^{G-1}$ is

$$\boldsymbol{s}(z) = \boldsymbol{c}_z^\top, \qquad z = 1, \ldots, G,$$

and the reparameterizing matrix $\mathbb{S}$ is an $n \times (G-1)$ matrix

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_1^\top \\ -- \\ \vdots \\ -- \\ \boldsymbol{c}_G^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_1^\top \end{matrix}}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\vphantom{\begin{matrix} \boldsymbol{c}_G^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{matrix}}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \mathbf{1}_{n_G} \otimes \boldsymbol{c}_G^\top \end{pmatrix}.$$

Evaluation of the effect of the categorical covariate

With a given full-rank parameterization of a categorical covariate, evaluation of a statistical significance of its effect on the response expectation corresponds to testing the null hypothesis

$$\mathrm{H}_0: \ \beta_1 = 0 \ \ \& \ \ \cdots \ \ \& \ \ \beta_{G-1} = 0, \tag{7.20}$$

or written concisely

$$\mathrm{H}_0: \ \boldsymbol{\beta}^Z = \mathbf{0}_{G-1}.$$

This null hypothesis indeed also corresponds to a submodel where only intercept is included in the model matrix. Finally, it can be mentioned that the null hypothesis (7.20) is indeed equivalent to the hypothesis of equality of the group means

$$\mathrm{H}_0: \ m_1 = \cdots = m_G. \tag{7.21}$$

If normality of the response is assumed, equivalently an F-test on a submodel (Theorem 5.1) or a test on a value of a subvector of the regression coefficients (F-test if $G \geq 2$, t-test if $G = 2$, see Theorem 3.2) can be used.

Notes. The following can be shown with only a little algebra:

- If $G = 2$, $\boldsymbol{\beta} = \left(\beta_0, \beta_1\right)^\top$. The (usual) t-statistic to test the hypothesis $\mathrm{H}_0: \beta_1 = 0$ using point (viii) of Theorem 3.2, i.e., the statistic based on the LSE of $\boldsymbol{\beta}$, is the same as a statistic of a standard two-sample t-test.

[14] *kontrastová matice* [15] *pseudokontrastová matice*

- If $G \geq 2$, the (usual) F-statistic to test the null hypothesis (7.20) using point (x) of Theorem 3.2 which is the same as the (usual) F-statistic on a submodel, where the submodel is the only-intercept model, is the same as an F-statistic used classically in one-way analysis of variance (ANOVA) to test the null hypothesis (7.21).

In the following, we introduce some of classically used (pseudo)contrast parameterizations which include: (i) reference group pseudocontrasts, (ii) sum contrasts, (iii) weighted sum contrasts, (iv) Helmert contrasts, and (v) orthonormal polynomial contrasts.

Reference group pseudocontrasts

$$\mathbb{C} = \begin{pmatrix} 0 & \dots & 0 \\ 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{G-1}^\top \\ \mathbf{I}_{G-1} \end{pmatrix} \tag{7.22}$$

The regression coefficients have the following interpretation

$$\begin{aligned} m_1 &= \beta_0, & \beta_0 &= m_1, \\ m_2 &= \beta_0 + \beta_1, & \beta_1 &= m_2 - m_1, \\ &\vdots & &\vdots \\ m_G &= \beta_0 + \beta_{G-1}, & \beta_{G-1} &= m_G - m_1. \end{aligned} \tag{7.23}$$

That is, the intercept β_0 is equal to the mean of the first (reference) group, the elements of $\boldsymbol{\beta}^Z = \big(\beta_1, \dots, \beta_{G-1}\big)^\top$ (the effects of Z) provide differences between the means of the remaining groups and the reference one. The regression function can be written as

$$m(z) = \beta_0 \, + \, \beta_1 \, \mathbb{I}(z = 2) \, + \, \cdots \, + \, \beta_{G-1} \, \mathbb{I}(z = G), \qquad z = 1, \dots, G.$$

It is seen from (7.23) that the full-rank parameterization using the reference group pseudocontrasts is equivalent to the less-than-full-rank (ANOVA) parameterization $m_g = \alpha_0 + \alpha_g$, $g = 1, \dots, G$, where $\boldsymbol{\alpha} = \big(\alpha_0, \alpha_1, \dots, \alpha_G\big)^\top$ is identified by the reference group constraint

$$\alpha_1 = 0.$$

Notes.

- With the pseudocontrast matrix $\mathbb{C}$ given by (7.22), a group labeled by $Z = 1$ is chosen as a reference for which the intercept β_0 provides the group mean. In practice, any other group can be taken as a reference by moving the zero row of the $\mathbb{C}$ matrix.
- In the R software, the reference group pseudocontrasts with the $\mathbb{C}$ matrix being of the form (7.22) are used by default to parameterize categorical covariates (`factors`). Explicitely this choice is indicated by the `contr.treatment` function. Alternatively, the `contr.SAS` function provides a pseudocontrast matrix in which the last Gth group serves as the reference, i.e., the $\mathbb{C}$ matrix has zeros on its last row.

Sum contrasts

$$\mathbb{C} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ -1 & \dots & -1 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{G-1} \\ -\mathbf{1}_{G-1}^\top \end{pmatrix} \tag{7.24}$$

Let

$$\overline{m} = \frac{1}{G} \sum_{g=1}^{G} m_g.$$

The regression coefficients have the following interpretation

$$\begin{aligned} & & \beta_0 &= \overline{m}, \\ m_1 &= \beta_0 + \beta_1, & \beta_1 &= m_1 - \overline{m}, \\ & \vdots & & \vdots \\ m_{G-1} &= \beta_0 + \beta_{G-1}, & \beta_{G-1} &= m_{G-1} - \overline{m}. \\ m_G &= \beta_0 - \sum_{g=1}^{G-1} \beta_g, & & \end{aligned} \tag{7.25}$$

The regression function can be written as

$$m(z) = \beta_0 + \beta_1 \mathbb{I}(z = 1) + \cdots + \beta_{G-1} \mathbb{I}(z = G-1) - \Big(\sum_{g=1}^{G-1} \beta_g \Big) \mathbb{I}(z = G),$$

$$z = 1, \dots, G.$$

If we consider the less-than-full-rank ANOVA parameterization of the group means as $m_g = \alpha_0 + \alpha_g$, $g = 1, \dots, G$, it is seen from (7.25) that the full-rank parameterization using the contrast matrix (7.24) links the regression coefficients of the two models as

$$\begin{aligned} \alpha_0 &= \beta_0 &&= \overline{m}, \\ \alpha_1 &= \beta_1 &&= \mu_1 - \overline{m}, \\ & \vdots & & \vdots \\ \alpha_{G-1} &= \beta_{G-1} &&= \mu_{G-1} - \overline{m}, \\ \alpha_G &= -\sum_{g=1}^{G-1} \beta_g &&= \mu_G - \overline{m}. \end{aligned} \tag{7.26}$$

At the same time, the vector $\boldsymbol{\alpha}$ satisfies

$$\sum_{g=1}^{G} \alpha_g = 0. \tag{7.27}$$

That is, the full-rank parameterization using the sum contrasts (7.25) is equivalent to the less-than-full-rank ANOVA parameterization, where the regression coefficients are identified by the sum constraint (7.27). The intercepts $\alpha_0 = \beta_0$ equal to the mean of the group means and the elements of $\boldsymbol{\beta}^Z = \big(\beta_1, \dots, \beta_{G-1}\big)^\top = \big(\alpha_1, \dots, \alpha_{G-1}\big)^\top$ are equal to the differences between

the corresponding group mean and the means of the group means. The same quantity for the last, Gth group, α_G is calculated from $\boldsymbol{\beta}^Z$ as $\alpha_G = -\sum_{g=1}^{G-1} \beta_g$.

Note. In the R software, the sum contrasts with the $\mathbb{C}$ matrix being of the form (7.24) can be used by the mean of the function `contr.sum`.

Weighted sum contrasts

$$\mathbb{C} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ -\dfrac{n_1}{n_G} & \dots & -\dfrac{n_{G-1}}{n_G} \end{pmatrix} \tag{7.28}$$

Let

$$\overline{m}_W = \frac{1}{n} \sum_{g=1}^{G} n_g \, m_g.$$

The regression coefficients have the following interpretation

$$\begin{aligned} & & \beta_0 &= \overline{m}_W, \\ m_1 &= \beta_0 + \beta_1, & \beta_1 &= m_1 - \overline{m}_W, \\ & \vdots & & \vdots \\ m_{G-1} &= \beta_0 + \beta_{G-1}, & \beta_{G-1} &= m_{G-1} - \overline{m}_W. \\ m_G &= \beta_0 - \sum_{g=1}^{G-1} \frac{n_g}{n_G} \beta_g, & & \end{aligned} \tag{7.29}$$

The regression function can be written as

$$m(z) = \beta_0 + \beta_1 \, \mathbb{I}(z = 1) + \dots + \beta_{G-1} \, \mathbb{I}(z = G-1) - \Big(\sum_{g=1}^{G-1} \frac{n_g}{n_G} \beta_g \Big) \, \mathbb{I}(z = G),$$

$$z = 1, \dots, G.$$

If we consider the less-than-full-rank ANOVA parameterization of the group means as $m_g = \alpha_0 + \alpha_g$, $g = 1, \dots, G$, it is seen from (7.29) that the full-rank parameterization using the contrast matrix (7.28) links the regression coefficients of the two models as

$$\begin{aligned} \alpha_0 &= \beta_0 &&= \overline{m}_W, \\ \alpha_1 &= \beta_1 &&= m_1 - \overline{m}_W, \\ & \vdots & & \vdots \\ \alpha_{G-1} &= \beta_{G-1} &&= m_{G-1} - \overline{m}_W, \\ \alpha_G &= -\sum_{g=1}^{G-1} \frac{n_g}{n_G} \beta_g &&= m_G - \overline{m}_W. \end{aligned}$$

At the same time, the vector $\boldsymbol{\alpha}$ satisfies

$$\sum_{g=1}^{G} n_g \, \alpha_g = 0. \tag{7.30}$$

That is, the full-rank parameterization using the weighted sum pseudocontrasts (7.29) is equivalent to the less-than-full-rank ANOVA parameterization, where the regression coefficients are identified by the weighted sum constraint (7.30). The intercepts $\alpha_0 = \beta_0$ equal to the *weighted* mean of the group means and the elements of $\boldsymbol{\beta}^Z = \big(\beta_1, \ldots, \beta_{G-1}\big)^\top = \big(\alpha_1, \ldots, \alpha_{G-1}\big)^\top$ are equal to the differences between the corresponding group mean and the *weighted* means of the group means. The same quantity for the last, Gth group, α_G is calculated from $\boldsymbol{\beta}^Z$ as $\alpha_G = -\sum_{g=1}^{G-1} \frac{n_g}{n_G}\beta_g$.

Helmert contrasts

$$\mathbb{C} = \begin{pmatrix} -1 & -1 & \ldots & -1 \\ 1 & -1 & \ldots & -1 \\ 0 & 2 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & G-1 \end{pmatrix} \tag{7.31}$$

The group means are obtained from the regression coefficients as

$$\begin{aligned} m_1 &= \beta_0 - \sum_{g=1}^{G-1} \beta_g, \\ m_2 &= \beta_0 + \beta_1 - \sum_{g=2}^{G-1} \beta_g, \\ m_3 &= \beta_0 + 2\,\beta_2 - \sum_{g=3}^{G-1} \beta_g, \\ &\vdots \\ m_{G-1} &= \beta_0 + (G-2)\,\beta_{G-2} - \beta_{G-1}, \\ m_G &= \beta_0 + (G-1)\,\beta_{G-1}. \end{aligned}$$

Inversely, the regression coefficients are linked to the group means as

$$\begin{aligned} \beta_0 &= \frac{1}{G}\sum_{g=1}^{G} m_g &&=: \overline{m}, \\ \beta_1 &= \frac{1}{2}(m_2 - m_1), \\ \beta_2 &= \frac{1}{3}\Big\{m_3 - \frac{1}{2}(m_1 + m_2)\Big\}, \\ \beta_3 &= \frac{1}{4}\Big\{m_4 - \frac{1}{3}(m_1 + m_2 + m_3)\Big\}, \\ &\vdots \\ \beta_{G-1} &= \frac{1}{G}\Big\{m_G - \frac{1}{G-1}\sum_{g=1}^{G-1} m_g\Big\}. \end{aligned}$$

which provide their (slightly awkward) interpretation: β_g, $g = 1, \ldots, G-1$, is $1/(g+1)$ times the difference between the mean of group $g+1$ and the mean of the means of the previous groups $1, \ldots, g$.

Note. In the R software, the Helmert contrasts with the $\mathbb{C}$ matrix being of the form (7.31) can be used by the mean of the function `contr.helmert`.

Orthonormal polynomial contrasts

$$\mathbb{C} = \begin{pmatrix} P^1(\omega_1) & P^2(\omega_1) & \dots & P^{G-1}(\omega_1) \\ P^1(\omega_2) & P^2(\omega_2) & \dots & P^{G-1}(\omega_2) \\ \vdots & \vdots & \vdots & \vdots \\ P^1(\omega_G) & P^2(\omega_G) & \dots & P^{G-1}(\omega_G) \end{pmatrix}, \tag{7.32}$$

where $\omega_1 < \cdots < \omega_G$ is an *equidistant (arithmetic)* sequence of the group labels and

$$P^j(z) = a_{j,0} + a_{j,1}\, z + \cdots + a_{j,j}\, z^j, \qquad j = 1, \dots, G-1,$$

are *orthonormal* polynomials of degree $1, \dots, G-1$ built above a sequence of the group labels.

Note. It can be shown that the polynomial coefficients $a_{j,l}$, $j = 1, \dots, G-1$, $l = 0, \dots, j$ and hence the $\mathbb{C}$ matrix (7.32) is for given G invariant (up to orientation) towards the choice of the group labels as soon as they form an *equidistant (arithmetic)* sequence. For example, for $G = 2, 3, 4$ the $\mathbb{C}$ matrix is

G = 2

$$\mathbb{C} = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix},$$

G = 3

$$\mathbb{C} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ 0 & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix},$$

G = 4

$$\mathbb{C} = \begin{pmatrix} -\frac{3}{2\sqrt{5}} & \frac{1}{2} & -\frac{1}{2\sqrt{5}} \\ -\frac{1}{2\sqrt{5}} & -\frac{1}{2} & \frac{3}{2\sqrt{5}} \\ \frac{1}{2\sqrt{5}} & -\frac{1}{2} & -\frac{3}{2\sqrt{5}} \\ \frac{3}{2\sqrt{5}} & \frac{1}{2} & \frac{1}{2\sqrt{5}} \end{pmatrix}.$$

The group means are then obtained as

$$\begin{aligned} m_1 &= m(\omega_1) = \beta_0 + \beta_1\, P^1(\omega_1) + \cdots + \beta_{G-1}\, P^{G-1}(\omega_1), \\ m_2 &= m(\omega_2) = \beta_0 + \beta_1\, P^1(\omega_2) + \cdots + \beta_{G-1}\, P^{G-1}(\omega_2), \\ &\vdots \\ m_G &= m(\omega_G) = \beta_0 + \beta_1\, P^1(\omega_G) + \cdots + \beta_{G-1}\, P^{G-1}(\omega_G), \end{aligned}$$

where

$$m(z) = \beta_0 + \beta_1\, P^1(z) + \cdots + \beta_{G-1}\, P^{G-1}(z), \qquad z \in \{\omega_1, \dots, \omega_G\}$$

is the regression function. The regression coefficients $\boldsymbol{\beta}$ now do not have any direct interpretation. That is why, even though the parameterization with the contrast matrix (7.32) can be used with the categorical *nominal* covariate, it is only rarely done so. Nevertheless, in case of the categorical

ordinal covariate where the *ordered* group labels $\omega_1 < \cdots < \omega_G$ have also practical interpretability, parameterization (7.32) can be used to reveal possible polynomial trends in the evolution of the group means $m_1, \ldots, m_G$ and to evaluate whether it may make sense to consider that covariate as *numeric* rather than *categorical*. Indeed, for $d < G$, the null hypothesis

$$\mathrm{H}_0 : \ \beta_d = 0 \ \ \& \ \ \ldots \ \ \& \ \ \beta_{G-1} = 0$$

corresponds to the hypothesis that the covariate at hand can be considered as numeric (with values $\omega_1, \ldots, \omega_G$ of the form of an equidistant sequence) and the evolution of the group means can be described by a polynomial of degree $d - 1$.

Note. In the R software, the orthonormal polynomial contrasts with the $\mathbb{C}$ matrix being of the form (7.32) can be used by the mean of the function `contr.poly`. It is also a default choice if the covariate is coded as categorical ordinal (`ordered`).

End of Lecture #9 (27/10/2016)

Chapter 8

Additivity and Interactions

8.1 Additivity and partial effect of a covariate

Start of Lecture #11 (03/11/2016)

Suppose now that the covariate vectors are

$$\left(Z_1, \boldsymbol{V}_1^\top\right)^\top, \ldots, \left(Z_n, \boldsymbol{V}_n^\top\right)^\top \in \mathcal{Z} \times \mathcal{V}, \qquad \mathcal{Z} \subseteq \mathbb{R},\ \mathcal{V} \subseteq \mathbb{R}^{p-1}, \qquad p \geq 2.$$

As usual, let Y denote a generic response variable, $\left(Z, \boldsymbol{V}^\top\right)^\top \in \mathcal{Z} \times \mathcal{V}$ a generic covariate vector, and let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, \qquad \mathbb{V} = \begin{pmatrix} \boldsymbol{V}_1^\top \\ \vdots \\ \boldsymbol{V}_n^\top \end{pmatrix}$$

be vectors/matrices covering the response variables and covariate values for the n data points.

8.1.1 Additivity

Definition 8.1 Additivity of the covariate effect.

We say that a covariate $Z \in \mathcal{Z}$ acts additively in the regression model with covariates $\left(Z, \boldsymbol{V}^\top\right)^\top \in \mathcal{Z} \times \mathcal{V}$, if the regression function is of the form

$$\mathbb{E}\big(Y \,\big|\, Z = z, \boldsymbol{V} = \boldsymbol{v}\big) = m(z, \boldsymbol{v}) = m_Z(z) + m_{\boldsymbol{V}}(\boldsymbol{v}), \qquad \left(z, \boldsymbol{v}^\top\right)^\top \in \mathcal{Z} \times \mathcal{V}, \tag{8.1}$$

where $m_Z : \mathcal{Z} \longrightarrow \mathbb{R}$ and $m_{\boldsymbol{V}} : \mathcal{V} \longrightarrow \mathbb{R}$ are some measurable functions.

8.1.2 Partial effect and conditional independence

If the effect of $Z \in \mathcal{Z}$ acts additively in a regression model, we have for *any* fixed $\boldsymbol{v} \in \mathcal{V}$:

$$\mathbb{E}\big(Y \,\big|\, Z = z+1, \boldsymbol{V} = \boldsymbol{v}\big) - \mathbb{E}\big(Y \,\big|\, Z = z, \boldsymbol{V} = \boldsymbol{v}\big) = m_Z(z+1) - m_Z(z), \quad z \in \mathcal{Z}. \tag{8.2}$$

That is, the influence (effect) of the covariate Z on the response expectation is the same with any value of $\boldsymbol{V} \in \mathcal{V}$.

Terminology *(Partial effect of a covariate).*

If a covariate $Z \in \mathcal{Z}$ acts additively in the regression model with covariates $\left(Z, \boldsymbol{V}^\top\right)^\top \in \mathcal{Z} \times \mathcal{V}$, quantity (8.2) expresses so called *partial effect*[1] of the covariate Z on the response given the value of $\boldsymbol{V}$.

Hypothesis of no partial effect of the Z covariate corresponds to testing

$$\mathrm{H}_0: \quad m_Z(z) = \text{const}, \qquad z \in \mathcal{Z}. \tag{8.3}$$

If it can be assumed that the covariates at hand (Z and $\boldsymbol{V}$) influence only the (conditional) response (Y) expectation and not other characteristics of the conditional distribution of the response given the covariates (as it is for example the case of a normal linear model), then the null hypothesis (8.3) corresponds to *conditional* independence between the response Y and the Z covariate given the remaining covariates $\boldsymbol{V}$ (given an arbitrary value of the remaining covariates).

8.1.3 Additivity in a linear model

In a context of a linear model, both m_Z and $m_{\boldsymbol{V}}$ are chosen to be linear in unknown (regression) parameters and the corresponding model matrix is decomposed as

$$\mathbb{X} = \left(\mathbb{X}^Z, \ \mathbb{X}^{\boldsymbol{V}}\right),$$

where $\mathbb{X}^Z$ corresponds to the regression function m_Z and depends only on the covariate values $\mathbb{Z}$, and $\mathbb{X}^{\boldsymbol{V}}$ corresponds to the regression function $m_{\boldsymbol{V}}$ and depends only on the covariate values $\mathbb{V}$. That is, the response expectation is assumed to be

$$\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}, \mathbb{V}\right) = \mathbb{X}^Z\boldsymbol{\beta} + \mathbb{X}^{\boldsymbol{V}}\boldsymbol{\gamma},$$

for some real vectors of regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Matrix $\mathbb{X}^Z$ and the regression function m_Z then correspond to parameterization of a single covariate for which any choice out of those introduced in Sections 7.3 and 7.4 (or others not discussed here) can be used plus possibly an intercept column. In most practical situations,

$$\mathsf{rank}\left(\mathbb{X}^Z\right) \geq 2, \qquad \mathbf{1}_n \in \mathcal{M}\left(\mathbb{X}^Z\right).$$

Model with the model matrix

$$\mathbb{X}^0 = \left(\mathbf{1}_n, \ \mathbb{X}^{\boldsymbol{V}}\right)$$

is then a submodel of the model with the model matrix $\mathbb{X} = \left(\mathbb{X}^Z, \ \mathbb{X}^{\boldsymbol{V}}\right)$ and corresponds to the regression function

$$m_0(z, \boldsymbol{v}) = \beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v}), \qquad \left(z, \boldsymbol{v}^\top\right)^\top \in \mathcal{Z} \times \mathcal{V}.$$

Hypothesis (8.3) of no partial effect of the Z covariate and conditional independence between the Z covariate and the response Y given the remaining covariates $\boldsymbol{V}$ corresponds to testing a submodel with the model matrix $\mathbb{X}^0 = \left(\mathbf{1}_n, \mathbb{X}^{\boldsymbol{V}}\right)$ againts a model with the model matrix $\mathbb{X} = \left(\mathbb{X}^Z, \ \mathbb{X}^{\boldsymbol{V}}\right)$.

[1] *parciální efekt*

8.2 Additivity of the effect of a numeric covariate

Remember that with additivity of the Z and the $\boldsymbol{V}$ covariates, the regression function $m(z, \boldsymbol{v})$, the related model matrix $\mathbb{X}$ of the linear model and expression of the conditional mean of the response vector $\boldsymbol{Y}$ are

$$\begin{aligned} \mathbb{E}\big(Y \,\big|\, Z = z, \boldsymbol{V} = \boldsymbol{v}\big) = m(z, \boldsymbol{v}) &= m_Z(z) + m_{\boldsymbol{V}}(\boldsymbol{v}), \\ \mathbb{X} &= \Big(\mathbb{X}^Z, \quad \mathbb{X}^{\boldsymbol{V}}\Big), \\ \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}, \mathbb{V}\big) &= \mathbb{X}^Z\boldsymbol{\beta} + \mathbb{X}^{\boldsymbol{V}}\boldsymbol{\gamma}, \end{aligned}$$

where $\mathbb{X}^Z = \boldsymbol{t}^Z(\mathbb{Z})$ is based on some transformation $\boldsymbol{t}^Z$ of the Z covariates, $\mathbb{X}^{\boldsymbol{V}} = \boldsymbol{t}^{\boldsymbol{V}}(\mathbb{V})$ is based on some transformation $\boldsymbol{t}^{\boldsymbol{V}}$ of the $\boldsymbol{V}$ covariates and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression coefficients.

Let us now assume that Z is a numeric covariate with $\mathcal{Z} \subseteq \mathbb{R}$. While limiting ourselves to parameterizations discussed in Section 7.3, the matrix $\mathbb{X}^Z$ can be

(i) $\mathbb{X}^Z = \big(\mathbf{1}_n, \mathbb{S}^Z\big)$, where $\mathbb{S}^Z$ is a reparameterizing matrix of a parameterization

$$\boldsymbol{s}_Z = \big(s_1, \ldots, s_{k-1}\big)^\top : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}$$

having a form of either

(a) a simple transformation (Section 7.3.1);

(b) raw polynomials (Section 7.3.2);

(c) orthonormal polynomials (Section 7.3.3).

If we denote the regression coefficients related to the model matrix $\mathbb{X}^Z$ as $\boldsymbol{\beta} = \big(\beta_0, \beta_1, \ldots, \beta_{k-1}\big)^\top$, the regression function is

$$m(z, \boldsymbol{v}) = \beta_0 + \beta_1 s_1(z) + \cdots + \beta_{k-1} s_{k-1}(z) + m_{\boldsymbol{V}}(\boldsymbol{v}), \qquad \big(z, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{V}, \quad (8.4)$$

which can also be interpreted as

$$m(z, \boldsymbol{v}) = \gamma_0(\boldsymbol{v}) + \beta_1 s_1(z) + \cdots + \beta_{k-1} s_{k-1}(z), \qquad \big(z, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{V},$$

where $\gamma_0(\boldsymbol{v}) = \beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v})$.

In other words, if a certain covariate acts additively and its effect on the response is described by parameterization $\boldsymbol{s}_Z$ then the remaining covariates $\boldsymbol{V}$ only modify an intercept term in the relationship between the response and the covariate Z.

(ii) $\mathbb{X}^Z = \mathbb{B}^Z$, where $\mathbb{B}^Z$ is a model matrix (7.11) of the regression splines $B_1, \ldots, B_k$. With the regression coefficients related to the model matrix $\mathbb{B}^Z$ being denoted as $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_k\big)^\top$, the regression function becomes

$$m(z, \boldsymbol{v}) = \beta_1 B_1(z) + \cdots + \beta_k B_k(z) + m_{\boldsymbol{V}}(\boldsymbol{v}), \qquad \big(z, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{V}, \quad (8.5)$$

where the term $m_{\boldsymbol{V}}(\boldsymbol{v})$ can again be interpreted as an intercept $\gamma_0(\boldsymbol{v}) = m_{\boldsymbol{V}}(\boldsymbol{v})$ in the relationship between response and the covariate Z whose value depends on the remaining covariates $\boldsymbol{V}$.

8.2.1 Partial effect of a numeric covariate

With the regression function (8.4), the partial effect of the Z covariate on the response is determined by a set of the non-intercept regression coefficients $\boldsymbol{\beta}^Z := \big(\beta_1, \ldots, \beta_{k-1}\big)^\top$. The null hypothesis

$$\mathrm{H}_0: \quad \boldsymbol{\beta}^Z = \mathbf{0}_{k-1}$$

then expresses the hypothesis that the covariate Z has, *conditionally* given a fixed (even though arbitrary) value of $\boldsymbol{V}$, no effect on the response expectation. That is, it is a hypothesis of no *partial* effect of the covariate Z on the response expectation.

With the spline-based regression function (8.5), the partial effect of the Z covariate is expressed by (all) spline-related regression coefficients $\beta_1, \ldots, \beta_k$. Nevertheless, due to the B-splines property (7.8), the null hypothesis of no partial effect of the Z covariate is now

$$\mathrm{H}_0: \quad \beta_1 = \cdots = \beta_k.$$

8.3 Additivity of the effect of a categorical covariate

Remember again that with additivity of the Z and the $\boldsymbol{V}$ covariates, the regression function $m(z, \boldsymbol{v})$, the related model matrix $\mathbb{X}$ of the linear model and expression of the conditional mean of the response vector $\boldsymbol{Y}$ are

$$\begin{aligned} \mathbb{E}\big(Y \,\big|\, Z = z, \boldsymbol{V} = \boldsymbol{v}\big) = m(z, \boldsymbol{v}) &= m_Z(z) + m_{\boldsymbol{V}}(\boldsymbol{v}), \\ \mathbb{X} &= \Big(\mathbb{X}^Z, \quad \mathbb{X}^{\boldsymbol{V}}\Big), \\ \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}, \mathbb{V}\big) &= \mathbb{X}^Z\boldsymbol{\beta} + \mathbb{X}^{\boldsymbol{V}}\boldsymbol{\gamma}, \end{aligned} \tag{8.6}$$

where $\mathbb{X}^Z = \boldsymbol{t}^Z(\mathbb{Z})$ is based on some transformation $\boldsymbol{t}^Z$ of the Z covariates, $\mathbb{X}^{\boldsymbol{V}} = \boldsymbol{t}^{\boldsymbol{V}}(\mathbb{V})$ is based on some transformation $\boldsymbol{t}^{\boldsymbol{V}}$ of the $\boldsymbol{V}$ covariates and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression coefficients.

Assume now that Z is a categorical covariate with $\mathcal{Z} = \{1, \ldots, G\}$ where $Z = g$, $g = 1, \ldots, G$, is repeated n_g-times in the data which are assumed (without loss of generality) to be sorted according to the values of this covariate. The group means used in Section 7.4 must now be understood as *conditional* group means, given a value of the covariates $\boldsymbol{V}$, and the regression function (8.6) parameterizes those conditional group means, i.e., for $\boldsymbol{v} \in \mathcal{V}$:

$$\begin{aligned} m(1, \boldsymbol{v}) &= \mathbb{E}\big(Y \,\big|\, Z = 1, \boldsymbol{V} = \boldsymbol{v}\big) &=: m_1(\boldsymbol{v}), \\ &\vdots & \vdots \\ m(G, \boldsymbol{v}) &= \mathbb{E}\big(Y \,\big|\, Z = G, \boldsymbol{V} = \boldsymbol{v}\big) &=: m_G(\boldsymbol{v}). \end{aligned} \tag{8.7}$$

Let

$$\boldsymbol{m}(\boldsymbol{v}) = \big(m_1(\boldsymbol{v}), \ldots, m_G(\boldsymbol{v})\big)^\top$$

be a vector of those conditional group means if the remaining covariates take a value $\boldsymbol{V} = \boldsymbol{v}$.

The matrix $\mathbb{X}^Z$ can be any of the model matrices discussed in Section 7.4. If we restrict ourselves to the full-rank parameterizations introduced in Section 7.4.4, the matrix $\mathbb{X}^Z$ is

$$\mathbb{X}^Z = \big(\mathbf{1}_n, \mathbb{S}^Z\big), \quad \mathbb{S}^Z = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \mathbf{1}_{n_G} \otimes \boldsymbol{c}_G^\top \end{pmatrix},$$

where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_G \in \mathbb{R}^{G-1}$ are rows of a chosen (pseudo)contrast matrix

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix}.$$

If $\boldsymbol{\beta} = \big(\beta_0, \beta_1, \ldots, \beta_{G-1}\big)^\top$ denotes the regression coefficients related to the model matrix $\mathbb{X}^Z = \big(\mathbf{1}_n, \mathbb{S}^Z\big)$ and we further denote $\boldsymbol{\beta}^Z = \big(\beta_1, \ldots, \beta_{G-1}\big)^\top$, the conditional group means are, for $\boldsymbol{v} \in \mathcal{V}$, given as

$$\begin{aligned} m_1(\boldsymbol{v}) &= \beta_0 + \boldsymbol{c}_1^\top\boldsymbol{\beta}^Z + m_{\boldsymbol{V}}(\boldsymbol{v}) &= \gamma_0(\boldsymbol{v}) + \boldsymbol{c}_1^\top\boldsymbol{\beta}^Z, \\ &\vdots & \vdots \\ m_G(\boldsymbol{v}) &= \beta_0 + \boldsymbol{c}_G^\top\boldsymbol{\beta}^Z + m_{\boldsymbol{V}}(\boldsymbol{v}) &= \gamma_0(\boldsymbol{v}) + \boldsymbol{c}_G^\top\boldsymbol{\beta}^Z, \end{aligned} \tag{8.8}$$

where $\gamma_0(\boldsymbol{v}) = \beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v})$, $\boldsymbol{v} \in \mathcal{V}$. In a matrix notation, (8.8) becomes

$$\boldsymbol{m}(\boldsymbol{v}) = \mathbf{1}_G\,\gamma_0(\boldsymbol{v}) + \mathbb{C}\boldsymbol{\beta}^Z. \tag{8.9}$$

8.3.1 Partial effects of a categorical covariate

In agreement with a general expression (8.2), we have for arbitrary $\boldsymbol{v} \in \mathcal{V}$ and arbitrary $g_1, g_2 \in \mathcal{Z}$:

$$\begin{aligned} \mathbb{E}\big(Y \,\big|\, Z = g_1,\, \boldsymbol{V} = \boldsymbol{v}\big) - \mathbb{E}\big(Y \,\big|\, Z = g_2,\, \boldsymbol{V} = \boldsymbol{v}\big) &= m_{g_1}(\boldsymbol{v}) - m_{g_2}(\boldsymbol{v}) \\ &= \big(\boldsymbol{c}_{g_1} - \boldsymbol{c}_{g_2}\big)^\top \boldsymbol{\beta}^Z, \end{aligned} \tag{8.10}$$

which does not depend on a value of $\boldsymbol{V} = \boldsymbol{v}$. That is, the difference between the two conditional group means is the same for all values of the covariates in $\boldsymbol{V}$.

Terminology *(Partial effects of a categorical covariate).*

If additivity of a categorical Z covariate and $\boldsymbol{V}$ covariates can be assumed, a vector of coefficients $\boldsymbol{\beta}^Z$ from parameterization of the conditional group means (Eqs. 8.8, 8.9) will be referred to as *partial effects* of the categorical covariate.

Note. It should be clear from (8.10) that interpretation of the partial effects of a categorical covariate depends on chosen parameterization (chosen (pseudo)contrast matrix $\mathbb{C}$).

If the Z covariate acts additively with the $\boldsymbol{V}$ covariate, it makes sense to ask a question whether all G conditional group means are, for a given $\boldsymbol{v} \in \mathcal{V}$, equal. That is, whether all partial effects of the Z covariate are equal to zero. In general, this corresponds to the null hypothesis

$$\mathrm{H}_0 : \; m_1(\boldsymbol{v}) = \cdots = m_G(\boldsymbol{v}), \qquad \boldsymbol{v} \in \mathcal{V}. \tag{8.11}$$

If the regression function is parameterized as (8.8), the null hypothesis (8.11) is expressed using the partial effects as

$$\mathrm{H}_0 : \; \boldsymbol{\beta}^Z = \mathbf{0}_{G-1}.$$

8.3.2 Interpretation of the regression coefficients

Note that (8.8) and (8.9) are basically the same expressions as those in (7.19) in Section 7.4.4. The only difference is dependence of the group means and the intercept term on the value of the covariates $\boldsymbol{V}$. Hence interpretation of the individual coefficients β_0 and $\boldsymbol{\beta}^Z = \big(\beta_1, \ldots, \beta_{G-1}\big)^\top$ depends on the chosen pseudocontrast matrix $\mathbb{C}$, nevertheless, it is basically the same as in case of a single categorical covariate in Section 7.4.4 with the only difference that

(i) The non-intercept coefficients in $\boldsymbol{\beta}^Z$ have the same interpretation as in Section 7.4.4 but always conditionally, given a chosen (even though arbitrary) value $\boldsymbol{v} \in \mathcal{V}$.

(ii) The intercept β_0 has interpretation given in Section 7.4.4 only for such $\boldsymbol{v} \in \mathcal{V}$ for which $m_{\boldsymbol{V}}(\boldsymbol{v}) = 0$. This follows from the fact that, again, for a chosen $\boldsymbol{v} \in \mathcal{V}$, the expression (8.8) of the conditional group means is the same as in Section 7.4.4. Nevertheless only for $\boldsymbol{v}$ such that $m_{\boldsymbol{V}}(\boldsymbol{v}) = 0$, we have $\beta_0 = \gamma_0(\boldsymbol{v})$.

Example 8.1 (Reference group pseudocontrasts).

If $\mathbb{C}$ is the reference group pseudocontrasts matrix (7.22), we obtain analogously to (7.23), but now for a chosen $\boldsymbol{v} \in \mathcal{V}$, the following

$$\begin{aligned} \beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v}) \;=\; \gamma_0(\boldsymbol{v}) &= m_1(\boldsymbol{v}), \\ \beta_1 &= m_2(\boldsymbol{v}) - m_1(\boldsymbol{v}), \\ &\vdots \\ \beta_{G-1} &= m_{G-1}(\boldsymbol{v}) - m_1(\boldsymbol{v}). \end{aligned}$$

Example 8.2 (Sum contrasts).

If $\mathbb{C}$ is the sum contrasts matrix (7.24), we obtain analogously to (7.25), but now for a chosen $\boldsymbol{v} \in \mathcal{V}$, the following

$$\begin{aligned} \beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v}) &= \gamma_0(\boldsymbol{v}) &&= \overline{m}(\boldsymbol{v}), \\ \beta_1 &= m_1(\boldsymbol{v}) - \overline{m}(\boldsymbol{v}), \\ &\vdots \\ \beta_{G-1} &= m_{G-1}(\boldsymbol{v}) - \overline{m}(\boldsymbol{v}), \end{aligned}$$

where

$$\overline{m}(\boldsymbol{v}) = \frac{1}{G}\sum_{g=1}^{G} m_g(\boldsymbol{v}), \qquad \boldsymbol{v} \in \mathcal{V}.$$

If we additionally define $\alpha_G = -\sum_{g=1}^{G-1} \beta_g$, we get, in agreement with (7.26),

$$\alpha_G \;=\; -\sum_{g=1}^{G-1} \beta_g \;=\; m_G(\boldsymbol{v}) - \overline{m}(\boldsymbol{v}).$$

8.4 Effect modification and interactions

8.4.1 Effect modification

From now onwards, we will assume that the covariate vectors are

$$\left(Z_1, W_1, \boldsymbol{V}_1^\top\right)^\top, \ldots, \left(Z_n, W_n, \boldsymbol{V}_n^\top\right)^\top \in \mathcal{Z}\times\mathcal{W}\times\mathcal{V}, \qquad \mathcal{Z} \subseteq \mathbb{R},\ \mathcal{W} \subseteq \mathbb{R},\ \mathcal{V} \subseteq \mathbb{R}^{p-2}, \qquad p \geq 2.$$

As usual, let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, \qquad \mathbb{W} = \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}, \qquad \mathbb{V} = \begin{pmatrix} \boldsymbol{V}_1^\top \\ \vdots \\ \boldsymbol{V}_n^\top \end{pmatrix}$$

denote vectors/matrices collecting the response variables and covariate values for the n data points. Finally, let (as always), Y denote a generic response variable and $\left(Z, W, \boldsymbol{V}^\top\right)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}$ be a generic covariate vector. In this and the following sections, we will assume that the regression function is

$$\mathbb{E}\left(Y \,\middle|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\right) = m(z, w, \boldsymbol{v}) = m_{ZW}(z, w) + m_{\boldsymbol{V}}(\boldsymbol{v}), \qquad \left(z, w, \boldsymbol{v}^\top\right)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}, \tag{8.12}$$

where $m_{\boldsymbol{V}} : \mathcal{V} \longrightarrow \mathbb{R}$ is some measurable function and $m_{ZW} : \mathcal{Z} \times \mathcal{W} \longrightarrow \mathbb{R}$ is a measurable function that cannot be factorized as $m_{ZW}(z, w) = m_Z(z) + m_W(w)$. We then have for *any* fixed $\boldsymbol{v} \in \mathcal{V}$.

$$\mathbb{E}\left(Y \,\middle|\, Z = z+1, W = w, \boldsymbol{V} = \boldsymbol{v}\right) - \mathbb{E}\left(Y \,\middle|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\right) = m_{ZW}(z+1, w) - m_{ZW}(z, w), \tag{8.13}$$

$$\mathbb{E}\left(Y \,\middle|\, Z = z, W = w+1, \boldsymbol{V} = \boldsymbol{v}\right) - \mathbb{E}\left(Y \,\middle|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\right) = m_{ZW}(z, w+1) - m_{ZW}(z, w), \qquad z \in \mathcal{Z},\ w \in \mathcal{W}, \tag{8.14}$$

where (8.13), i.e., the effect of a covariate Z on the response expectation possibly depends on the value of $W = w$ and also (8.14), i.e., the effect of a covariate W on the response expectation possibly depends on the value of $Z = z$. We then say that the covariates Z and W are mutual *effect modifiers*.[2]

8.4.2 Effect modification in a linear model

In a context of a linear model, both m_{ZW} and $m_{\boldsymbol{V}}$ are chosen to be linear in unknown (regression) parameters and the corresponding model matrix is decomposed as

$$\begin{aligned} \mathbb{X} &= \left(\mathbb{X}^{ZW}, \quad \mathbb{X}^{\boldsymbol{V}}\right), \\ \mathbb{E}\left(Y \,\middle|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\right) = m(z, w, \boldsymbol{v}) &= m_{ZW}(z, w) + m_{\boldsymbol{V}}(\boldsymbol{v}), \end{aligned}$$

where $\mathbb{X}^{ZW} = \boldsymbol{t}^{ZW}(\mathbb{Z}, \mathbb{W})$ corresponds to the regression function m_{ZW} and results from a certain transformation $\boldsymbol{t}^{ZW}$ of the Z and W covariates, and $\mathbb{X}^{\boldsymbol{V}} = \boldsymbol{t}^{\boldsymbol{V}}(\mathbb{V})$ corresponds to the regression

[2] *modifikátory efektu*

function $m_{\boldsymbol{V}}$ and results from a certain transformation $\boldsymbol{t}^{\boldsymbol{V}}$ of the V covariates. In the rest of this section and in Sections 8.5, 8.6 and 8.7, we show classical choices for the matrix $\mathbb{X}^{ZW}$ based on so called interactions derived from covariate parameterizations that we introduced in Sections 7.3 and 7.4.

End of Lecture #11 (03/11/2016)

Start of Lecture #13 (10/11/2016)

8.4.3 Interactions

Suppose that the covariate Z is parameterized using a parameterization

$$\boldsymbol{s}_Z = \left(s_1^Z, \ldots, s_{k-1}^Z\right)^\top : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}, \tag{8.15}$$

and the covariate W is parameterized using a parameterization

$$\boldsymbol{s}_W = \left(s_1^W, \ldots, s_{l-1}^W\right)^\top : \mathcal{W} \longrightarrow \mathbb{R}^{l-1}, \tag{8.16}$$

and let $\mathbb{S}^Z$ and $\mathbb{S}^W$ be the corresponding reparameterizing matrices:

$$\mathbb{S}^Z = \begin{pmatrix} \boldsymbol{s}_Z^\top(Z_1) \\ \vdots \\ \boldsymbol{s}_Z^\top(Z_n) \end{pmatrix} = \left(\boldsymbol{S}_Z^1, \ldots, \boldsymbol{S}_Z^{k-1}\right), \qquad \mathbb{S}^W = \begin{pmatrix} \boldsymbol{s}_W^\top(W_1) \\ \vdots \\ \boldsymbol{s}_W^\top(W_n) \end{pmatrix} = \left(\boldsymbol{S}_W^1, \ldots, \boldsymbol{S}_W^{l-1}\right).$$

Definition 8.2 Interaction terms.
Let $\left(Z_1, W_1\right)^\top, \ldots, \left(Z_n, W_n\right)^\top \in \mathcal{Z} \times \mathcal{W} \subseteq \mathbb{R}^2$ *be values of two covariates being parameterized using the reparameterizing matrices* $\mathbb{S}^Z$ *and* $\mathbb{S}^W$. *By* interaction terms[3] *based on the reparameterizing matrices* $\mathbb{S}^Z$ *and* $\mathbb{S}^W$ *we mean columns of a matrix*

$$\mathbb{S}^{ZW} := \mathbb{S}^Z : \mathbb{S}^W.$$

Note. See Definition A.5 for a definition of the columnwise product of two matrices. We have

$$\begin{aligned} \mathbb{S}^{ZW} = \mathbb{S}^Z : \mathbb{S}^W &= \left(\boldsymbol{S}_Z^1 : \boldsymbol{S}_W^1, \ldots, \boldsymbol{S}_Z^{k-1} : \boldsymbol{S}_W^1, \ldots, \boldsymbol{S}_Z^1 : \boldsymbol{S}_W^{l-1}, \ldots, \boldsymbol{S}_Z^{k-1} : \boldsymbol{S}_W^{l-1}\right) \\ &= \begin{pmatrix} \boldsymbol{s}_W^\top(W_1) \otimes \boldsymbol{s}_Z^\top(Z_1) \\ \vdots \\ \boldsymbol{s}_W^\top(W_n) \otimes \boldsymbol{s}_Z^\top(Z_n) \end{pmatrix} \\ &= \begin{pmatrix} s_1^Z(Z_1)\, s_1^W(W_1) & \cdots & s_{k-1}^Z(Z_1)\, s_1^W(W_1) & \cdots & s_1^Z(Z_1)\, s_{l-1}^W(W_1) & \cdots & s_{k-1}^Z(Z_1)\, s_{l-1}^W(W_1) \\ \vdots & & \vdots & & \vdots & & \vdots \\ s_1^Z(Z_n)\, s_1^W(W_n) & \cdots & s_{k-1}^Z(Z_n)\, s_1^W(W_n) & \cdots & s_1^Z(Z_n)\, s_{l-1}^W(W_n) & \cdots & s_{k-1}^Z(Z_n)\, s_{l-1}^W(W_n) \end{pmatrix}. \end{aligned}$$

[3] *interakční členy*

8.4.4 Linear model with interactions

Remember that we are now assuming that the regression function $m(z, \boldsymbol{v})$ is decomposed as $m(z, w, \boldsymbol{v}) = m_{ZW}(z, w) + m_{\boldsymbol{V}}(\boldsymbol{v})$. In context of a linear model, all factors of the regression function are chosen to be linear in the unknown model parameters and are determined by corresponding model matrices:

$$\begin{aligned} \mathbb{E}\big(Y \,\big|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\big) = m(z, w, \boldsymbol{v}) &= m_{ZW}(z, w) + m_{\boldsymbol{V}}(\boldsymbol{v}), \\ \mathbb{X} &= \Big(\mathbb{X}^{ZW}, \quad \mathbb{X}^{\boldsymbol{V}}\Big), \end{aligned} \tag{8.17}$$

where $\mathbb{X}^{ZW} = \boldsymbol{t}^{ZW}(\mathbb{Z}, \mathbb{W})$ is based on some transformation $\boldsymbol{t}^{ZW}$ of the Z and W covariates and $\mathbb{X}^{\boldsymbol{V}} = \boldsymbol{t}^{\boldsymbol{V}}(\mathbb{V})$ is based on some transformation $\boldsymbol{t}^{\mathbb{V}}$ of the $\boldsymbol{V}$ covariates.

Interaction terms in parameterization of the m_{ZW} part of the regression function, i.e., in the matrix $\mathbb{X}^{ZW}$ are used inside the linear model to express a certain form of the effect modification. If $\mathbf{1}_n \notin \mathbb{S}^Z$ and $\mathbf{1}_n \notin \mathbb{S}^W$, the matrix $\mathbb{X}^{ZW}$ from (8.17) is usually chosen as

$$\mathbb{X}^{ZW} = \big(\mathbf{1}_n, \mathbb{S}^Z, \mathbb{S}^W, \mathbb{S}^{ZW}\big), \tag{8.18}$$

which, as will be shown, corresponds to a certain form of the effect modification. Let the related regression coefficients be denoted as

$$\boldsymbol{\beta} = \big(\beta_0, \underbrace{\beta_1^Z, \ldots, \beta_{k-1}^Z}_{=:\,\boldsymbol{\beta}^Z}, \underbrace{\beta_1^W, \ldots, \beta_{l-1}^W}_{=:\,\boldsymbol{\beta}^W}, \underbrace{\beta_{1,1}^{ZW}, \ldots, \beta_{k-1,1}^{ZW}, \ldots, \beta_{1,l-1}^{ZW}, \ldots, \beta_{k-1,l-1}^{ZW}}_{=:\,\boldsymbol{\beta}^{ZW}}\big)^\top.$$

That is, the regression function is

$$\begin{aligned} m(z, w, \boldsymbol{v}) &= \beta_0 + \beta_1^Z\, s_1^Z(z) + \cdots + \beta_{k-1}^Z\, s_{k-1}^Z \\ &\quad + \beta_1^W\, s_1^W(w) + \cdots + \beta_{l-1}^W\, s_{l-1}^W \\ &\quad + \beta_{1,1}^{ZW}\, s_1^Z(z)\, s_1^W(w) + \cdots + \beta_{k-1,1}^{ZW}\, s_{k-1}^Z(z)\, s_1^W(w) \\ &\qquad + \cdots + \beta_{1,l-1}^{ZW}\, s_1^Z(z)\, s_{l-1}^W(w) + \cdots + \beta_{k-1,l-1}^{ZW}\, s_{k-1}^Z(z)\, s_{l-1}^W(w) \\ &\quad + m_{\boldsymbol{V}}(\boldsymbol{v}) \\ &= \beta_0 + \boldsymbol{s}_Z^\top(z)\,\boldsymbol{\beta}^Z + \boldsymbol{s}_W^\top(w)\,\boldsymbol{\beta}^W + \underbrace{\big(\boldsymbol{s}_W^\top(w) \otimes \boldsymbol{s}_Z^\top(z)\big)}_{\boldsymbol{s}_{ZW}^\top(z,\,w)}\,\boldsymbol{\beta}^{ZW} + m_{\boldsymbol{V}}(\boldsymbol{v}), \\ &\qquad \big(z, w, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}, \end{aligned} \tag{8.19}$$

where $\boldsymbol{s}_{ZW}(z, w) = \boldsymbol{s}_W(w) \otimes \boldsymbol{s}_Z(z)$, $\big(z, w\big)^\top \in \mathcal{Z} \times \mathcal{W}$.

Main and interaction effects

Coefficients in $\boldsymbol{\beta}^Z$ and $\boldsymbol{\beta}^W$ are called the *main effects*[4] of the covariate Z and W, respectively. Coefficients in $\boldsymbol{\beta}^{ZW}$ are called the *interaction effects*.[5]

[4] *hlavní efekty* [5] *interakční efekty*

The effects of the covariates Z or W, given the remaining covariates are then expressed as

$$\begin{aligned}&\mathbb{E}\big(Y \,\big|\, Z = z+1,\, W = w,\, \boldsymbol{V} = \boldsymbol{v}\big) - \mathbb{E}\big(Y \,\big|\, Z = z,\, W = w,\, \boldsymbol{V} = \boldsymbol{v}\big)\\ &\qquad = \big\{\boldsymbol{s}_Z(z+1) \,-\, \boldsymbol{s}_Z(z)\big\}^\top \boldsymbol{\beta}^Z \,+\, \big\{\boldsymbol{s}_{ZW}(z+1,\, w) \,-\, \boldsymbol{s}_{ZW}(z,\, w)\big\}^\top \boldsymbol{\beta}^{ZW}, \qquad (8.20)\\ &\mathbb{E}\big(Y \,\big|\, Z = z,\, W = w+1,\, \boldsymbol{V} = \boldsymbol{v}\big) - \mathbb{E}\big(Y \,\big|\, Z = z,\, W = w,\, \boldsymbol{V} = \boldsymbol{v}\big)\\ &\qquad = \big\{\boldsymbol{s}_W(w+1) \,-\, \boldsymbol{s}_W(w)\big\}^\top \boldsymbol{\beta}^W \,+\, \big\{\boldsymbol{s}_{ZW}(z,\, w+1) \,-\, \boldsymbol{s}_{ZW}(z,\, w)\big\}^\top \boldsymbol{\beta}^{ZW}, \qquad (8.21)\\ &\qquad\qquad z \in \mathcal{Z},\; w \in \mathcal{W}.\end{aligned}$$

That is, the effect (8.20) of the covariate Z is determined by the main effects $\boldsymbol{\beta}^Z$ of this covariate as well as by the interaction effects $\boldsymbol{\beta}^{ZW}$. Analogously, the effect (8.21) of the covariate W is determined by its main effects $\boldsymbol{\beta}^W$ as well as by the interaction effects $\boldsymbol{\beta}^{ZW}$.

Hypothesis of no effect modification

If factor $m_{ZW}(z,\, w)$ of the regression function (8.17) is parameterized by matrix $\mathbb{X}^{ZW}$ given by (8.18) then the hypothesis of no effect modification is expressed by considering a submodel in which matrix $\mathbb{X}^{ZW}$ is replaced by a matrix

$$\mathbb{X}^{Z+W} = \big(\mathbf{1}_n,\, \mathbb{S}^Z,\, \mathbb{S}^W\big).$$

8.4.5 Rank of the interaction model

Remind that we assume in the whole lecture that the number of rows n of all considered model matrices is higher than the number of columns and hence the rank of all such matrices is equal to their column rank.

Lemma 8.1 Rank of the interaction model.

(i) Let $\mathsf{rank}\Big(\big(\mathbb{S}^Z,\, \mathbb{S}^W\big)\Big) = k + l - 2$, *i.e., all columns from the matrices* $\mathbb{S}^Z$ *and* $\mathbb{S}^W$ *are linearly independent and the matrix* $\big(\mathbb{S}^Z,\, \mathbb{S}^W\big)$ *is of full-rank. Then the matrix* $\mathbb{S}^{ZW} = \mathbb{S}^Z : \mathbb{S}^W$ *is of full-rank as well, i.e.,*

$$\mathsf{rank}\big(\mathbb{S}^{ZW}\big) \;=\; (k-1)\,(l-1).$$

(ii) Let additionally $\mathbf{1}_n \notin \mathcal{M}\big(\mathbb{S}^Z\big)$, $\mathbf{1}_n \notin \mathcal{M}\big(\mathbb{S}^W\big)$. *Then also a matrix* $\mathbb{X}^{ZW} = \big(\mathbf{1}_n,\, \mathbb{S}^Z,\, \mathbb{S}^W,\, \mathbb{S}^{ZW}\big)$ *is of full-rank, i.e.,*

$$\mathsf{rank}\big(\mathbb{X}^{ZW}\big) \;=\; 1 + (k-1) + (l-1) + (k-1)\,(l-1) \;=\; k\,l.$$

Proof. Left as exercise in linear algebra.

Proof/calculations were skipped and are not requested for the exam.

❑

***Note** (Hypothesis of no effect modification).*

Under the conditions of Lemma 8.1, we have for $\mathbb{X}^{ZW} = \big(\mathbf{1}_n, \mathbb{S}^Z, \mathbb{S}^W, \mathbb{S}^{ZW}\big)$ and $\mathbb{X}^{Z+W} = \big(\mathbf{1}_n, \mathbb{S}^Z, \mathbb{S}^W\big)$:

$$\mathsf{rank}\big(\mathbb{X}^{ZW}\big) = k\,l, \qquad \mathsf{rank}\big(\mathbb{X}^{Z+W}\big) = 1 + (k-1) + (l-1) = k + l - 1,$$
$$\mathcal{M}\big(\mathbb{X}^{Z+W}\big) \subset \mathcal{M}\big(\mathbb{X}^{ZW}\big).$$

If the hypothesis of no effect modification is tested by a submodel F-test then its numerator degrees of freedom are

$$kl - k - l + 1 = (k-1)\cdot(l-1).$$

The corresponding null hypothesis can also be specified as a hypothesis on the zero value of an estimable vector of all interaction effects:

$$\mathrm{H}_0: \ \boldsymbol{\beta}^{ZW} = \mathbf{0}_{(k-1)\cdot(l-1)}.$$

8.4.6 Interactions with the regression spline

Either the Z covariate or/and the W covariate can also be parameterized by the regression splines. In that case, the interaction terms are defined in the same way as in Definition 8.2. As example, consider situation where the W covariate is parameterized by the regression splines

$$\boldsymbol{B}_W = \big(B_1^W, \ldots, B_l^W\big)^\top$$

with the related model matrix

$$\mathbb{B}^W = \begin{pmatrix} \boldsymbol{B}_W^\top(W_1) \\ \vdots \\ \boldsymbol{B}_W^\top(W_n) \end{pmatrix} = \big(\boldsymbol{B}_W^1, \ldots, \boldsymbol{B}_W^l\big),$$

and the Z covariates by the parameterization (8.15) and the reparameterizing matrix $\mathbb{S}^Z$ as usual. In that case, the interaction terms are columns of a matrix

$$\begin{aligned} \mathbb{B}^{ZW} = \mathbb{S}^Z : \mathbb{B}^W &= \big(\boldsymbol{S}_Z^1 : \boldsymbol{B}_W^1, \ \ldots, \ \boldsymbol{S}_Z^{k-1} : \boldsymbol{B}_W^1, \ \ldots, \ \boldsymbol{S}_Z^1 : \boldsymbol{B}_W^l, \ \ldots, \ \boldsymbol{S}_Z^{k-1} : \boldsymbol{B}_W^l\big) \\ &= \begin{pmatrix} \boldsymbol{B}_W^\top(W_1) \otimes \boldsymbol{s}_Z^\top(Z_1) \\ \vdots \\ \boldsymbol{B}_W^\top(W_n) \otimes \boldsymbol{s}_Z^\top(Z_n) \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} s_1^Z(Z_1)\,B_1^W(W_1) & \cdots & s_{k-1}^Z(Z_1)\,B_1^W(W_1) & \cdots & s_1^Z(Z_1)\,B_l^W(W_1) & \cdots & s_{k-1}^Z(Z_1)\,B_l^W(W_1) \\ \vdots & & \vdots & & \vdots & & \vdots \\ s_1^Z(Z_n)\,B_1^W(W_n) & \cdots & s_{k-1}^Z(Z_n)\,B_1^W(W_n) & \cdots & s_1^Z(Z_n)\,B_l^W(W_n) & \cdots & s_{k-1}^Z(Z_n)\,B_l^W(W_n) \end{pmatrix}.$$

Interaction model with regression splines

We have $\mathbf{1}_n \in \mathcal{M}(\mathbb{B}^W)$ and also $\mathcal{M}(\mathbb{S}^Z) \subset \mathcal{M}(\mathbb{B}^{ZW})$ for $\mathbb{B}^{ZW} = \mathbb{S}^Z : \mathbb{B}^W$ (see also Section 8.5.3). That is,

$$\mathcal{M}\Big((\mathbf{1}_n, \mathbb{S}^Z, \mathbb{B}^W, \mathbb{B}^{ZW})\Big) = \mathcal{M}\Big((\mathbb{B}^W, \mathbb{B}^{ZW})\Big).$$

It is thus sufficient (with respect to obtained regression space) to choose the matrix $\mathbb{X}^{ZW}$ as

$$\mathbb{X}^{ZW} = \big(\mathbb{B}^W, \mathbb{B}^{ZW}\big).$$

Let the related regression coefficients be denotes as

$$\boldsymbol{\beta} = \big(\underbrace{\beta_1^W, \ldots, \beta_l^W}_{=:\,\boldsymbol{\beta}^W}, \underbrace{\beta_{1,1}^{ZW}, \ldots, \beta_{k-1,1}^{ZW}, \ldots, \beta_{1,l}^{ZW}, \ldots, \beta_{k-1,l}^{ZW}}_{=:\,\boldsymbol{\beta}^{ZW}}\big)^\top.$$

The regression function is then

$$\begin{aligned} m(z, w, \boldsymbol{v}) \;&=\; \beta_1^W B_1^W(w) + \cdots + \beta_l^W B_l^W \\ &\quad + \; \beta_{1,1}^{ZW} s_1^Z(z) B_1^W(w) + \cdots + \beta_{k-1,1}^{ZW} s_{k-1}^Z(z) B_1^W(w) \\ &\qquad + \; \cdots \; + \; \beta_{1,l}^{ZW} s_1^Z(z) B_l^W(w) + \cdots + \beta_{k-1,l}^{ZW} s_{k-1}^Z(z) B_l^W(w) \\ &\quad + \; m_{\boldsymbol{V}}(\boldsymbol{v}) \\ &= \; \boldsymbol{B}_W^\top(w)\,\boldsymbol{\beta}^W \; + \; \underbrace{\big(\boldsymbol{B}_W^\top(w) \otimes \boldsymbol{s}_Z^\top(z)\big)}_{\boldsymbol{B}_{ZW}^\top(z,\,w)} \boldsymbol{\beta}^{ZW} \; + \; m_{\boldsymbol{V}}(\boldsymbol{v}), \end{aligned} \tag{8.22}$$

$$\big(z, w, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V},$$

where $\boldsymbol{B}_{ZW}(z, w) = \boldsymbol{B}_W(w) \otimes \boldsymbol{s}_Z(z), \quad \big(z, w\big)^\top \in \mathcal{Z} \times \mathcal{W}$.

No effect modification with the regression splines

The effect of the Z and W covariate, respectively, on the response expectation is

$$\begin{aligned} &\mathbb{E}\big(Y \,\big|\, Z = z+1, W = w, \boldsymbol{V} = \boldsymbol{v}\big) \;-\; \mathbb{E}\big(Y \,\big|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\big) \\ &\quad = \; \big\{\boldsymbol{B}_Z(z+1) \,-\, \boldsymbol{B}_Z(z)\big\}^\top \boldsymbol{\beta}^Z \; + \; \big\{\boldsymbol{B}_{ZW}(z+1, w) \,-\, \boldsymbol{B}_{ZW}(z, w)\big\}^\top \boldsymbol{\beta}^{ZW}, \\ &\mathbb{E}\big(Y \,\big|\, Z = z, W = w+1, \boldsymbol{V} = \boldsymbol{v}\big) \;-\; \mathbb{E}\big(Y \,\big|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\big) \\ &\quad = \; \big\{\boldsymbol{B}_{ZW}(z, w+1) \,-\, \boldsymbol{B}_{ZW}(z, w)\big\}^\top \boldsymbol{\beta}^{ZW}, \qquad \big(z, w, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}. \end{aligned}$$

Due to the fact that $\mathbf{1}_n \in \mathcal{M}(\mathbb{B}^Z)$, we have $\mathcal{M}(\mathbb{S}^W) \subset \mathcal{M}(\mathbb{B}^{ZW}) = \mathcal{M}(\mathbb{B}^Z : \mathbb{S}^W)$. Furthermore, let for $\boldsymbol{\beta}^W = \big(\beta_1^W, \ldots, \beta_{l-1}^W\big)^\top \in \mathbb{R}^{l-1}$, a vector $\boldsymbol{\beta}_0^{ZW}$ be defined as

$$\boldsymbol{\beta}_0^{ZW} = \big(\underbrace{\beta_1^W, \ldots, \beta_1^W}_{k\text{-times}}, \ldots, \underbrace{\beta_{l-1}^W, \ldots, \beta_{l-1}^W}_{k\text{-times}}\big)^\top. \tag{8.23}$$

Then (due to $\mathbf{1}_n \in \mathcal{M}(\mathbb{B}^Z)$), we have

$$\begin{aligned}
\big\{\boldsymbol{B}_{ZW}(z+1,\, w) \,-\, \boldsymbol{B}_{ZW}(z,\, w)\big\}^\top \boldsymbol{\beta}_0^{ZW} &= 0,\\
\big\{\boldsymbol{B}_{ZW}(z,\, w+1) \,-\, \boldsymbol{B}_{ZW}(z,\, w)\big\}^\top \boldsymbol{\beta}_0^{ZW} &= \big\{\boldsymbol{s}_W(w+1) \,-\, \boldsymbol{s}_W(w)\big\}^\top \boldsymbol{\beta}^W,\\
&\qquad\qquad (z,\, w)^\top \in \mathcal{Z} \times \mathcal{W}.
\end{aligned}$$

In terms of a linear model, situation when $\boldsymbol{\beta}^{ZW}$ has a form of (8.23) corresponds to replacing the $\mathbb{B}^{ZW}$ block in the model matrix $\mathbb{X}^{ZW}$, which parameterizes the m_{ZW} factor of the regression function by $\mathbb{S}^W$. That is, hypothesis of no effect modification corresponds to a submodel in which the matrix $\mathbb{X}^{ZW} = \big(\mathbb{B}^W,\, \mathbb{B}^{ZW}\big)$ is replaced by a matrix

$$\mathbb{X}^{Z+W} = \big(\mathbb{B}^W,\, \mathbb{S}^Z\big).$$

Rank of the models

With respect to the rank of the resulting models, analogous statements hold as those given in Lemma 8.1. Remember that the matrix $\mathbb{S}^Z$ has in general $k-1$ columns, the matrix $\mathbb{B}^W$ has l columns. Suppose that the matrices $\mathbb{S}^Z$ and $\mathbb{B}^W$ are such that

$$\mathsf{rank}\big(\mathbb{X}^{Z+W}\big) = \mathsf{rank}\Big(\big(\mathbb{B}^W,\, \mathbb{S}^Z\big)\Big) = l + k - 1,$$

that is, the columns from the matrices $\mathbb{S}^Z$ and $\mathbb{B}^W$ that parameterize the Z and the W covariate, respectively, are all linearly independent (which is satisfied in most practical situations). Analogously to Lemma 8.1, it can then be shown that both $\mathbb{B}^{ZW} = \mathbb{S}^Z : \mathbb{B}^W$ and $\mathbb{X}^{ZW} = \big(\mathbb{B}^W,\, \mathbb{B}^{ZW}\big)$ are of full-rank, i.e.,

$$\begin{aligned}
\mathsf{rank}\big(\mathbb{B}^{ZW}\big) &= \mathsf{rank}\big(\mathbb{S}^Z : \mathbb{B}^W\big) &&= (k-1)\, l,\\
\mathsf{rank}\big(\mathbb{X}^{ZW}\big) &= \mathsf{rank}\Big(\big(\mathbb{B}^W,\, \mathbb{B}^{ZW}\big)\Big) &&= l + (k-1)l \;=\; k\, l.
\end{aligned}$$

Test of hypothesis of no effect modification has then $k\, l - \big\{l + (k-1)\big\} = (k-1)\cdot(l-1)$ degrees of freedom.

8.5 Interaction of two numeric covariates

In this and two following sections, we discuss in more detail situation when the two covariates that are mutual effect modifiers are (i) both numeric, (ii) one of them numeric and one of them categorical (Section 8.6), (iii) both categorical (Section 8.7). We mainly focus on interpretation of the model parameters when the effect modification is modeled by interaction terms of a linear model.

First, we shall consider a situation when both Z and W are numeric covariates with $\mathcal{Z} \subseteq \mathbb{R}$, $\mathcal{W} \subseteq \mathbb{R}$. As in the whole chapter, there are possibly other covariates available, being given by the vector $\boldsymbol{V} \in \mathcal{V} \subseteq \mathbb{R}^{p-2}$, $p \geq 2$. The regression function is assumed to be

$$\mathbb{E}\big(Y \,\big|\, Z = z,\, W = w,\, \boldsymbol{V} = \boldsymbol{v}\big) = m(z,\, w,\, \boldsymbol{v}) = m_{ZW}(z,\, w) + m_{\boldsymbol{V}}(\boldsymbol{v}),$$
$$\big(z,\, w,\, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V},$$

leading to the model matrix

$$\mathbb{X} = \big(\mathbb{X}^{ZW},\, \mathbb{X}^{\boldsymbol{V}}\big), \tag{8.24}$$

where $\mathbb{X}^{ZW} = \boldsymbol{t}^{ZW}(\mathbb{Z},\, \mathbb{W})$ corresponds to the regression function m_{ZW} and results from a certain transformation $\boldsymbol{t}^{ZW}$ of the Z and W covariates, and $\mathbb{X}^{\boldsymbol{V}} = \boldsymbol{t}^{\boldsymbol{V}}(\mathbb{V})$ corresponds to the regression function $m_{\boldsymbol{V}}$ and results from a certain transformation $\boldsymbol{t}^{\boldsymbol{V}}$ of the V covariates. The response vector expectation is

$$\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z},\, \mathbb{W},\, \mathbb{V}\big) = \mathbb{X}^{ZW}\boldsymbol{\beta} + \mathbb{X}^{\boldsymbol{V}}\boldsymbol{\gamma},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression parameters.

8.5.1 Linear effect modification

Suppose first that $\mathbb{S}^Z$ is a reparameterizing matrix that corresponds to a simple *identity* transformation of the covariate Z. For the second covariate, W, assume that the matrix $\mathbb{S}^W$ is an $n \times (l-1)$ reparameterizing matrix that corresponds to the general parameterization (8.15), e.g., any of reparameterizing matrices discussed in Sections 7.3.1, 7.3.2 and 7.3.3. That is,

$$\begin{aligned} s_Z(z) &= z, && z \in \mathcal{Z}, \\ \boldsymbol{s}_W(w) &= \big(s_1^W(w),\, \ldots,\, s_{l-1}^W(w)\big)^\top, && w \in \mathcal{W}, \end{aligned}$$

$$\mathbb{S}^Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, \qquad \mathbb{S}^W = \begin{pmatrix} \boldsymbol{s}_W^\top(W_1) \\ \vdots \\ \boldsymbol{s}_W^\top(W_n) \end{pmatrix} = \begin{pmatrix} s_1^W(W_1) & \ldots & s_{l-1}^W(W_1) \\ \vdots & \vdots & \vdots \\ s_1^W(W_n) & \ldots & s_{l-1}^W(W_n) \end{pmatrix}, \tag{8.25}$$

If the mutual effect modification is expressed by the interactions, the matrix $\mathbb{X}^{ZW}$ from (8.24) is given as

$$\mathbb{X}^{ZW} = \begin{pmatrix} 1 & Z_1 & s_1^W(W_1) & \ldots & s_{l-1}^W(W_1) & Z_1\, s_1^W(W_1) & \ldots & Z_1\, s_{l-1}^W(W_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n & s_1^W(W_n) & \ldots & s_{l-1}^W(W_n) & Z_n\, s_1^W(W_n) & \ldots & Z_n\, s_{l-1}^W(W_n) \end{pmatrix}$$
$$\underbrace{}_{\mathbf{1}_n} \quad \underbrace{}_{\mathbb{S}^Z} \quad \underbrace{\phantom{s_1^W(W_1)\ldots s_{l-1}^W(W_1)}}_{\mathbb{S}^W} \quad \underbrace{\phantom{Z_1\, s_1^W(W_1)\ldots Z_1\, s_{l-1}^W(W_1)}}_{\mathbb{S}^{ZW}}$$

and the related regression coefficients are

$$\boldsymbol{\beta} = \big(\beta_0,\ \beta^Z,\ \underbrace{\beta_1^W, \ldots, \beta_{l-1}^W}_{\boldsymbol{\beta}^W},\ \underbrace{\beta_1^{ZW}, \ldots, \beta_{l-1}^{ZW}}_{\boldsymbol{\beta}^{ZW}}\big)^\top.$$

The regression function (8.19) then becomes

$$\begin{aligned} m(z, w, \boldsymbol{v}) &= \beta_0 \ + \ \beta^Z z \ + \ \beta_1^W s_1^W(w) \ + \cdots + \ \beta_{l-1}^W s_{l-1}^W(w) \\ &\quad + \ \beta_1^{ZW} z\, s_1^W(w) \ + \cdots + \ \beta_{l-1}^{ZW} z\, s_{l-1}^W(w) \ + \ m_{\boldsymbol{V}}(\boldsymbol{v}) \qquad (8.26) \\ &= \underbrace{\big\{\beta_0 \ + \ \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}^W \ + \ m_{\boldsymbol{V}}(\boldsymbol{v})\big\}}_{=:\ \gamma_0^Z(w,\, \boldsymbol{v})} \ + \ \underbrace{(\beta^Z \ + \ \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}^{ZW})}_{=:\ \gamma_1^Z(w)}\, z \qquad (8.27) \\ &= \underbrace{\big\{\beta_0 + \beta^Z z + m_{\boldsymbol{V}}(\boldsymbol{v})\big\}}_{=:\ \gamma_0^W(z,\, \boldsymbol{v})} \\ &\quad + \underbrace{(\beta_1^W + \beta_1^{ZW} z)}_{=:\ \gamma_1^W(z)}\, s_1^W(w) \ + \cdots + \ \underbrace{(\beta_{l-1}^W + \beta_{l-1}^{ZW} z)}_{=:\ \gamma_{l-1}^W(z)}\, s_{l-1}^W(w), \qquad (8.28) \end{aligned}$$

$$\big(z, w, \boldsymbol{v}\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}.$$

The regression function (8.26) can be interpreted twofold.

(i) Expression (8.27) shows that for any fixed $w \in \mathcal{W}$, the covariates Z and $\boldsymbol{V}$ act additively and the effect of Z on the response expectation is expressed by a line. Nevertheless, both intercept γ_0^Z and the slope γ_1^Z of this line depend on w and this dependence is described by the parameterization $\boldsymbol{s}_W$. The intercept is further additively modified by a factor $m_{\boldsymbol{V}}(\boldsymbol{v})$.

With respect to interpretation, this shows that the main effect β^Z has an interpretation of the slope of the line that for a given $\boldsymbol{V} = \boldsymbol{v}$ describes the influence of Z on the response if $W = w$ is such that $\boldsymbol{s}_W(w) = \mathbf{0}_{l-1}$. This also shows that a test of the null hypothesis

$$\mathrm{H}_0 : \ \beta^Z = 0$$

does not evaluate a statistical significance of the influence of the covariate Z on the response expectation. It only evaluates it for values of $W = w$ for which $\boldsymbol{s}_W(w) = \mathbf{0}_{l-1}$.

(ii) Analogously, expression (8.28) shows that for any fixed $z \in \mathcal{Z}$, the covariates W and $\boldsymbol{V}$ act additively and the effect of W on the response expectation is expressed by its parameterization $\boldsymbol{s}_W$. Nevertheless, the related coefficients $(\gamma_0^W, \gamma_1^W, \ldots, \gamma_{l-1}^W)$ depend in a *linear* way on z. The intercept term is γ_0^W further additively modified by a factor $m_{\boldsymbol{V}}(\boldsymbol{v})$.

With respect to interpretation, this shows that the main effects $\boldsymbol{\beta}^W$ has an interpretation of the coefficients of the influence of the W covariate on the response if $Z = 0$. This also shows that a test of the null hypothesis

$$\mathrm{H}_0 : \ \boldsymbol{\beta}^W = \mathbf{0}_{l-1}$$

does not evaluate a statistical significance of the influence of the covariate W on the response expectation. It only evaluates it under the condition of $Z = 0$.

8.5.2 More complex effect modification

More complex effect modifications can be obtained by choosing a more complex reparameterizing matrix $\mathbb{S}^Z$ for the Z covariate. Suppose that

$$\begin{aligned}\boldsymbol{s}_Z(z) &= \big(s_1^Z(z), \ldots, s_{k-1}^Z(z)\big)^\top, \qquad z \in \mathcal{Z},\\ \boldsymbol{s}_W(w) &= \big(s_1^W(w), \ldots, s_{l-1}^W(w)\big)^\top, \quad w \in \mathcal{W},\end{aligned}$$

$$\mathbb{S}^Z = \begin{pmatrix}\boldsymbol{s}_Z^\top(Z_1)\\ \vdots \\ \boldsymbol{s}_Z^\top(Z_n)\end{pmatrix} = \begin{pmatrix}s_1^Z(Z_1) & \ldots & s_{k-1}^Z(Z_1)\\ \vdots & \vdots & \vdots \\ s_1^Z(Z_n) & \ldots & s_{k-1}^Z(Z_n)\end{pmatrix},$$

$$\mathbb{S}^W = \begin{pmatrix}\boldsymbol{s}_W^\top(W_1)\\ \vdots \\ \boldsymbol{s}_W^\top(W_n)\end{pmatrix} = \begin{pmatrix}s_1^W(W_1) & \ldots & s_{l-1}^W(W_1)\\ \vdots & \vdots & \vdots \\ s_1^W(W_n) & \ldots & s_{l-1}^W(W_n)\end{pmatrix},$$

and as before, the matrix $\mathbb{X}^{ZW}$ is given as $\mathbb{X}^{ZW} = \big(\mathbf{1}_n, \mathbb{S}^Z, \mathbb{S}^W, \mathbb{S}^{ZW}\big)$, where $\mathbb{S}^{ZW} = \mathbb{S}^Z : \mathbb{S}^W$. The regression coefficients related to the matrix $\mathbb{X}^{ZW}$ are now

$$\boldsymbol{\beta} = \big(\beta_0, \underbrace{\beta_1^Z, \ldots, \beta_{k-1}^Z}_{\boldsymbol{\beta}^Z}, \underbrace{\beta_1^W, \ldots, \beta_{l-1}^W}_{\boldsymbol{\beta}^W}, \underbrace{\beta_{1,1}^{ZW}, \ldots, \beta_{k-1,1}^{ZW}, \ldots, \beta_{1,l-1}^{ZW}, \ldots, \beta_{k-1,l-1}^{ZW}}_{\boldsymbol{\beta}^{ZW}}\big)^\top,$$

and the regression function is

$$\begin{aligned}m(z, w, \boldsymbol{v}) \;=\; & \beta_0 + \beta_1^Z\, s_1^Z(z) + \cdots + \beta_{k-1}^Z\, s_{k-1}^Z(z)\\ & + \beta_1^W\, s_1^W(w) + \cdots + \beta_{l-1}^W\, s_{l-1}^W(w)\\ & + \beta_{1,1}^{ZW}\, s_1^Z(z)\, s_1^W(w) + \cdots + \beta_{k-1,1}^{ZW}\, s_{k-1}^Z(z)\, s_1^W(w) + \cdots\\ & + \beta_{1,l-1}^{ZW}\, s_1^Z(z)\, s_{l-1}^W(w) + \cdots + \beta_{k-1,l-1}^{ZW}\, s_{k-1}^Z(z)\, s_{l-1}^W(w)\\ & + m_{\boldsymbol{V}}(\boldsymbol{v})\\ \\ =\; & \beta_0 + \boldsymbol{s}_Z^\top(z)\boldsymbol{\beta}^Z + \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}^W + \big(\boldsymbol{s}_W^\top(w) \otimes \boldsymbol{s}_Z^\top(z)\big)\boldsymbol{\beta}^{ZW} + m_{\boldsymbol{V}}(v),\\ & \big(z, w, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}.\end{aligned}$$

Interpretation of such a model is then straightforward generalization of the situation described in Section 8.5.1. Namely, the regression function can be written to see that the W covariate is

a modifier of the effect of the Z covariate:

$$m(z, w, \boldsymbol{v}) = \underbrace{\left\{\beta_0 + \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}^W + m_{\boldsymbol{V}}(\boldsymbol{v})\right\}}_{=: \gamma_0^Z(w, \boldsymbol{v})} + \underbrace{(\beta_1^Z + \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}_{1\bullet}^{ZW})}_{=: \gamma_1^Z(w)} s_1^Z(z) + \cdots + \underbrace{(\beta_{k-1}^Z + \boldsymbol{s}_W^\top(w)\beta_{k-1\bullet}^{ZW})}_{=: \gamma_{k-1}^Z(w)} s_{k-1}^Z(z),$$
$$\left(z, w, \boldsymbol{v}^\top\right)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V},$$

where the effect of Z on the response expectation is given by the function $\boldsymbol{s}_Z$, and $\boldsymbol{\beta}_{j\bullet}^{ZW} = \left(\beta_{j,1}^{ZW}, \ldots, \beta_{j,l-1}^{ZW}\right)^\top$, $j = 1, \ldots, k-1$.

Analogously, the regression function can also be written to see that the Z covariate is a modifier of the effect of the W covariate:

$$m(z, w, \boldsymbol{v}) = \underbrace{\left\{\beta_0 + \boldsymbol{s}_Z^\top(z)\boldsymbol{\beta}^Z + m_{\boldsymbol{V}}(\boldsymbol{v})\right\}}_{=: \gamma_0^W(z, \boldsymbol{v})} + \underbrace{(\beta_1^W + \boldsymbol{s}_Z^\top(z)\boldsymbol{\beta}_{\bullet 1}^{ZW})}_{=: \gamma_1^W(z)} s_1^W(w) + \cdots + \underbrace{(\beta_{l-1}^W + \boldsymbol{s}_Z^\top(z)\beta_{\bullet l-1}^{ZW})}_{=: \gamma_{l-1}^W(z)} s_{l-1}^W(w),$$
$$\left(z, w, \boldsymbol{v}^\top\right)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V},$$

where the effect of W on the response expectation is given by the function $\boldsymbol{s}_W$, and $\boldsymbol{\beta}_{\bullet j}^{ZW} = \left(\beta_{1,j}^{ZW}, \ldots, \beta_{k-1,j}^{ZW}\right)^\top$, $j = 1, \ldots, l-1$.

8.5.3 Linear effect modification of a regression spline

Let us now again assume that the covariate Z is parameterized using a simple *identity* transformation and the reparameterizing matrix $\mathbb{S}^Z$ is given as in (8.25). Nevertheless, for the covariate W, let us assume its parameterization using the regression splines

$$\boldsymbol{B}_W = \left(B_1^W, \ldots, B_l^W\right)^\top$$

with the related model matrix

$$\mathbb{B}^W = \begin{pmatrix} \boldsymbol{B}_W^\top(W_1) \\ \vdots \\ \boldsymbol{B}_W^\top(W_n) \end{pmatrix} = \begin{pmatrix} B_1^W(W_1) & \ldots & B_l^W(W_1) \\ \vdots & \vdots & \vdots \\ B_1^W(W_n) & \ldots & B_l^W(W_n) \end{pmatrix}.$$

Analogously to previous usage of a matrix $\mathbb{S}^{ZW}$, let a matrix $\mathbb{B}^{ZW}$ be defined as

$$\mathbb{B}^{ZW} = \mathbb{S}^Z : \mathbb{B}^W = \begin{pmatrix} Z_1 B_1^W(W_1) & \ldots & Z_1 B_k^W(W_1) \\ \vdots & \vdots & \vdots \\ Z_n B_1^W(W_n) & \ldots & Z_n B_k^W(W_n) \end{pmatrix}.$$

Remember that for any $w \in \mathcal{W}$, $\sum_{j=1}^l B_j^W(w) = 1$ from which it follows that

(i) $\mathbf{1}_n \in \mathbb{B}^W$;

(ii) $\mathcal{M}\big(\mathbb{S}^Z\big) \subseteq \mathcal{M}\big(\mathbb{B}^{ZW}\big)$.

That is,

$$\mathcal{M}\Big(\big(\mathbf{1}_n,\, \mathbb{S}^Z,\, \mathbb{B}^W,\, \mathbb{B}^{ZW}\big)\Big) \;=\; \mathcal{M}\Big(\big(\mathbb{B}^W,\, \mathbb{B}^{ZW}\big)\Big). \tag{8.29}$$

Hence if a full-rank linear model is to be obtained, where interaction between a covariate parameterized using the regression splines and a covariate parameterized using the reparameterizing matrix $\mathbb{S}^Z = \big(Z_1, \ldots, Z_n\big)^\top$ is included, the model matrix $\mathbb{X}^{ZW}$ must be of the form

$$\mathbb{X}^{ZW} = \big(\mathbb{B}^W, \mathbb{B}^{ZW}\big) = \Bigg(\underbrace{\begin{array}{ccc} B_1^W(W_1) & \ldots & B_l^W(W_1) \\ \vdots & \vdots & \vdots \\ B_1^W(W_n) & \ldots & B_l^W(W_n) \end{array}}_{\mathbb{B}^W} \quad \underbrace{\begin{array}{ccc} Z_1\, B_1^W(W_1) & \ldots & Z_1\, B_l^W(W_1) \\ \vdots & \vdots & \vdots \\ Z_n\, B_1^W(W_n) & \ldots & Z_n\, B_l^W(W_n) \end{array}}_{\mathbb{B}^{ZW}} \Bigg).$$

If we denote the related regression coefficients as

$$\boldsymbol{\beta} = \big(\underbrace{\beta_1^W, \ldots, \beta_l^W}_{=:\,\boldsymbol{\beta}^W}, \underbrace{\beta_1^{ZW}, \ldots, \beta_l^{ZW}}_{=:\,\boldsymbol{\beta}^{ZW}}\big)^\top,$$

the regression function (8.19) becomes

$$\begin{aligned} m(z, w, \boldsymbol{v}) &= \beta_1^W B_1^W(w) + \cdots + \beta_l^W B_l^W(w) \\ &\qquad + \beta_1^{ZW} z\, B_1^W(w) + \cdots + \beta_l^{ZW} z\, B_l^W(w) + m_{\boldsymbol{V}}(\boldsymbol{v}) && (8.30) \\ &= \underbrace{\big\{\boldsymbol{B}_W^\top(w)\boldsymbol{\beta}^W + m_{\boldsymbol{V}}(\boldsymbol{v})\big\}}_{=:\,\gamma_0^Z(w,\,\boldsymbol{v})} + \underbrace{\boldsymbol{B}_W^\top(w)\boldsymbol{\beta}^{ZW}}_{=:\,\gamma_1^Z(w)} z && (8.31) \\ &= \underbrace{(\beta_1^W + \beta_1^{ZW} z)}_{=:\,\gamma_1^W(z)} B_1^W(w) + \cdots + \underbrace{(\beta_l^W + \beta_l^{ZW} z)}_{=:\,\gamma_l^W(z)} B_l^W(w) + m_{\boldsymbol{V}}(\boldsymbol{v}), && (8.32) \end{aligned}$$

$$\big(z, w, \boldsymbol{v}\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}.$$

The regression function (8.30) can again be interpreted twofold.

(i) Expression (8.31) shows that for any fixed $w \in \mathcal{W}$, the covariates Z and $\boldsymbol{V}$ act additively and the effect of W on the response expectation is expressed by a line. Nevertheless, both intercept γ_0^Z and the slope γ_1^Z of this line depend on w and this dependence is described by the regression splines $\boldsymbol{B}_W$. The intercept is further additively modified by the factor $m_{\boldsymbol{V}}(\boldsymbol{v})$.

(ii) Analogously, expression (8.32) shows that for any fixed $z \in \mathcal{Z}$, the covariates W and $\boldsymbol{V}$ act additively and the effect of W on the response expectation is expressed by the regression splines $\boldsymbol{B}_W$. Nevertheless, related spline coefficients $(\gamma_1^W, \ldots, \gamma_l^W)$ depend in a *linear* way on z. With respect to interpretation, this shows that the main effects $\boldsymbol{\beta}^W$ has an interpretation of the coefficients of the influence of the W covariate on the response if $Z = 0$.

8.5.4 More complex effect modification of a regression spline

Also with regression splines, a more complex reparameterizing matrix $\mathbb{S}^Z$ based on a transformation $\boldsymbol{s}_Z = \left(s_1, \ldots, s_{k-1}\right)^\top$ can be chosen. The property (8.29) still holds and the matrix $\mathbb{X}^{ZW}$ can still be chosen as $\left(\mathbb{B}^W, \mathbb{B}^{ZW}\right)$, $\mathbb{B}^{ZW} = \mathbb{S}^Z : \mathbb{B}^W$. Interpretation of the model is again a straightforward generalization of the situation of Section 8.5.3.

8.6 Interaction of a categorical and a numeric covariate

Consider now a situation when Z is a categorical covariate with $\mathcal{Z} = \{1, \ldots, G\}$ where $Z = g$, $g = 1, \ldots, G$, is repeated n_g-times in the data and W is a numeric covariate with $\mathcal{W} \subseteq \mathbb{R}$. As in the previous sections, there are possibly other covariates available, given by the vector $\boldsymbol{V} \in \mathcal{V} \subseteq \mathbb{R}^{p-2}$, $p \geq 2$. As before, the regression function is assumed to be

$$\mathbb{E}\big(Y \,\big|\, Z = z, W = w, \boldsymbol{V} = \boldsymbol{v}\big) = m(z, w, \boldsymbol{v}) = m_{ZW}(z, w) + m_{\boldsymbol{V}}(\boldsymbol{v}), \qquad (8.33)$$
$$\big(z, w, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V},$$

leading to the model matrix

$$\mathbb{X} = \big(\mathbb{X}^{ZW}, \mathbb{X}^{\boldsymbol{V}}\big),$$

where $\mathbb{X}^{ZW} = \boldsymbol{t}^{ZW}(\mathbb{Z}, \mathbb{W})$ corresponds to the regression function m_{ZW} and results from a certain transformation $\boldsymbol{t}^{ZW}$ of the Z and W covariates, and $\mathbb{X}^{\boldsymbol{V}} = \boldsymbol{t}^{\boldsymbol{V}}(\mathbb{V})$ corresponds to the regression function $m_{\boldsymbol{V}}$ and results from a certain transformation $\boldsymbol{t}^{\boldsymbol{V}}$ of the V covariates. The response vector expectation is

$$\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}, \mathbb{W}, \mathbb{V}\big) = \mathbb{X}^{ZW}\boldsymbol{\beta} + \mathbb{X}^{\boldsymbol{V}}\boldsymbol{\gamma},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown regression parameters.

In the following, we will assume (without loss of generality) that data are sorted according to the values of the categorical covariate Z. For the clarity of notation, we will analogously to Section 7.4 use also a double subscript to number the individual observations where the first subscript will indicate the value of the covariate Z. That is, we will use

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_{n_1} \\ - - - - \\ \vdots \\ - - - - \\ Z_{n-n_G+1} \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} Z_{1,1} \\ \vdots \\ Z_{1,n_1} \\ - - - \\ \vdots \\ - - - \\ Z_{G,1} \\ \vdots \\ Z_{G,n_G} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ - - \\ \vdots \\ - - \\ G \\ \vdots \\ G \end{pmatrix}, \qquad \begin{pmatrix} W_1 \\ \vdots \\ W_{n_1} \\ - - - - \\ \vdots \\ - - - - \\ W_{n-n_G+1} \\ \vdots \\ W_n \end{pmatrix} = \begin{pmatrix} W_{1,1} \\ \vdots \\ W_{1,n_1} \\ - - - \\ \vdots \\ - - - \\ W_{G,1} \\ \vdots \\ W_{G,n_G} \end{pmatrix}.$$

If the categorical covariate Z can be interpreted as a label that indicates pertinence to one of the G groups, the regression function (8.33) in which the value of z is fixed at $z = g$, $g = 1, \ldots, G$, can be viewed as a regression function that parameterizes dependence of the response expectation on the numeric covariate W and possibly other covariates $\boldsymbol{V}$ in group g. We have for $w \in \mathcal{W}$, $\boldsymbol{v} \in \mathcal{V}$:

$$\begin{aligned} m(1, w, \boldsymbol{v}) &= \mathbb{E}\big(Y \,\big|\, Z = 1, W = w, \boldsymbol{V} = \boldsymbol{v}\big) &&=: m_1(w, \boldsymbol{v}), \\ &\vdots && \vdots \\ m(G, w, \boldsymbol{v}) &= \mathbb{E}\big(Y \,\big|\, Z = G, W = w, \boldsymbol{V} = \boldsymbol{v}\big) &&=: m_G(w, \boldsymbol{v}). \end{aligned} \qquad (8.34)$$

Functions $m_1, \ldots, m_G$ are then *conditional* (given a value of Z) regression functions describing dependence of the response expectation on the covariates W and $\boldsymbol{V}$.

Alternatively, for fixed $w \in \mathcal{W}$ and $\boldsymbol{v} \in \mathcal{V}$, a vector

$$\boldsymbol{m}(w, \boldsymbol{v}) = \big(m_1(w, \boldsymbol{v}), \ldots, m_G(w, \boldsymbol{v})\big)^\top$$

can be interpreted as a vector of *conditional* (given W and $\boldsymbol{V}$) group means.

In the following assume that the categorical covariate Z is parameterized by the mean of a chosen (pseudo)contrast matrix

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix}, \qquad \begin{array}{l} \boldsymbol{c}_1 = \big(c_{1,1}, \ldots, c_{1,G-1}\big)^\top, \\ \qquad \vdots \\ \boldsymbol{c}_G = \big(c_{G,1}, \ldots, c_{G,G-1}\big)^\top, \end{array} \qquad \text{that is,} \quad \mathbb{S}^Z = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \mathbf{1}_{n_G} \otimes \boldsymbol{c}_G^\top \end{pmatrix}.$$

8.6.1 Categorical effect modification

First suppose that the numeric covariate W is parameterized using a parameterization

$$\boldsymbol{s}_W = \big(s_1^W, \ldots, s_{l-1}^W\big)^\top : \mathcal{W} \longrightarrow \mathbb{R}^{l-1},$$

$\mathbb{S}^W$ is the corresponding $n \times (l-1)$ reparameterizing matrix. Let $\mathbb{S}_1^W, \ldots, \mathbb{S}_G^W$ be the blocks of the reparameterizing matrix $\mathbb{S}^W$ that correspond to datapoints with $Z = 1, \ldots, Z = G$. That is, for $g = 1, \ldots, G$, matrix $\mathbb{S}_g^W$ is an $n_g \times (l-1)$ matrix,

$$\mathbb{S}_g^W = \begin{pmatrix} \boldsymbol{s}_W^\top(W_{g,1}) \\ \vdots \\ \boldsymbol{s}_W^\top(W_{g,n_g}) \end{pmatrix} = \begin{pmatrix} s_1^W(W_{g,1}) & \ldots & s_{l-1}^W(W_{g,1}) \\ \vdots & \vdots & \vdots \\ s_1^W(W_{g,n_g}) & \ldots & s_{l-1}^W(W_{g,n_g}) \end{pmatrix} \qquad \text{and} \qquad \mathbb{S}^W = \begin{pmatrix} \mathbb{S}_1^W \\ \vdots \\ \mathbb{S}_G^W \end{pmatrix}.$$

When using the interactions between the Z and W covariates to express their mutual effect modification, the matrix $\mathbb{X}^{ZW}$ that parameterizes the term $m_{ZW}(z, w)$ in the regression function (8.33) is again

$$\mathbb{X}^{ZW} = \big(\mathbf{1}_n, \mathbb{S}^Z, \mathbb{S}^W, \mathbb{S}^{ZW}\big),$$

where the interaction matrix $\mathbb{S}^{ZW} = \mathbb{S}^Z : \mathbb{S}^W$ is an $n \times (G-1)(l-1)$ matrix:

$$\mathbb{S}^{ZW} = \left(\begin{array}{ccc|c|ccc} c_{1,1}\, s_1^W(W_{1,1}) & \ldots & c_{1,G-1}\, s_1^W(W_{1,1}) & \ldots & c_{1,1}\, s_{l-1}^W(W_{1,1}) & \ldots & c_{1,G-1}\, s_{l-1}^W(W_{1,1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{1,1}\, s_1^W(W_{1,n_1}) & \ldots & c_{1,G-1}\, s_1^W(W_{1,n_1}) & \ldots & c_{1,1}\, s_{l-1}^W(W_{1,n_1}) & \ldots & c_{1,G-1}\, s_{l-1}^W(W_{1,n_1}) \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline c_{G,1}\, s_1^W(W_{G,1}) & \ldots & c_{G,G-1}\, s_1^W(W_{G,1}) & \ldots & c_{G,1}\, s_{l-1}^W(W_{G,1}) & \ldots & c_{G,G-1}\, s_{l-1}^W(W_{G,1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{G,1}\, s_1^W(W_{G,n_G}) & \ldots & c_{G,G-1}\, s_1^W(W_{G,n_G}) & \ldots & c_{G,1}\, s_{l-1}^W(W_{G,n_G}) & \ldots & c_{G,G-1}\, s_{l-1}^W(W_{G,n_1}) \end{array}\right)$$

$$= \begin{pmatrix} s_1^W(W_{1,1})\,\boldsymbol{c}_1^\top & \dots & s_{l-1}^W(W_{1,1})\,\boldsymbol{c}_1^\top \\ \vdots & \vdots & \vdots \\ s_1^W(W_{1,n_1})\,\boldsymbol{c}_1^\top & \dots & s_{l-1}^W(W_{1,n_1})\,\boldsymbol{c}_1^\top \\ - - - - - - - - - - - - - - - - - - \\ \vdots & \vdots & \vdots \\ - - - - - - - - - - - - - - - - - - \\ s_1^W(W_{G,1})\,\boldsymbol{c}_G^\top & \dots & s_{l-1}^W(W_{G,1})\,\boldsymbol{c}_G^\top \\ \vdots & \vdots & \vdots \\ s_1^W(W_{G,n_G})\,\boldsymbol{c}_G^\top & \dots & s_{l-1}^W(W_{G,n_G})\,\boldsymbol{c}_G^\top \end{pmatrix} = \begin{pmatrix} \mathbb{S}_1^W \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \mathbb{S}_G^W \otimes \boldsymbol{c}_G^\top \end{pmatrix}.$$

The model matrix $\mathbb{X}^{ZW}$ which parameterizes the m_{ZW} factor in the regression function (8.33) is then

$$\mathbb{X}^{ZW} = \left(\mathbf{1}_n,\, \mathbb{S}^Z,\, \mathbb{S}^W,\, \mathbb{S}^{ZW}\right) = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} \otimes \boldsymbol{c}_1^\top & \mathbb{S}_1^W & \mathbb{S}_1^W \otimes \boldsymbol{c}_1^\top \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1}_{n_G} & \mathbf{1}_{n_G} \otimes \boldsymbol{c}_G^\top & \mathbb{S}_G^W & \mathbb{S}_G^W \otimes \boldsymbol{c}_G^\top \end{pmatrix}$$

with the related regression coefficients being

$$\boldsymbol{\beta} = \big(\beta_0,\, \underbrace{\beta_1^Z, \dots, \beta_{G-1}^Z}_{=:\,\boldsymbol{\beta}^Z},\, \underbrace{\beta_1^W, \dots, \beta_{l-1}^W}_{=:\,\boldsymbol{\beta}^W},\, \underbrace{\beta_{1,1}^{ZW}, \dots, \beta_{G-1,1}^{ZW}, \dots, \beta_{1,l-1}^{ZW}, \dots, \beta_{G-1,l-1}^{ZW}}_{=:\,\boldsymbol{\beta}^{ZW}}\big)^\top.$$

For following considerations, it will be useful to denote subvectors of the interaction effects $\boldsymbol{\beta}^{ZW}$ as

$$\boldsymbol{\beta}^{ZW} = \big(\underbrace{\beta_{1,1}^{ZW}, \dots, \beta_{G-1,1}^{ZW}}_{=:\,\boldsymbol{\beta}_{\bullet 1}^{ZW}}, \dots, \underbrace{\beta_{1,l-1}^{ZW}, \dots, \beta_{G-1,l-1}^{ZW}}_{=:\,\boldsymbol{\beta}_{\bullet l-1}^{ZW}}\big)^\top.$$

The regression function (8.33) is then given by

$$m(z,\, w,\, \boldsymbol{v}) \;=\; \beta_0 \;+\; \boldsymbol{c}_z^\top \boldsymbol{\beta}^Z \;+\; \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}_W \;+\; \big(\boldsymbol{s}_W^\top(w) \otimes \boldsymbol{c}_z^\top\big)\boldsymbol{\beta}^{ZW} \;+\; m_{\boldsymbol{V}}(\boldsymbol{v}),$$
$$\big(z,\, w,\, \boldsymbol{v}^\top\big)^\top \in \mathcal{Z} \times \mathcal{W} \times \mathcal{V}. \quad (8.35)$$

For $z = g$, $g = 1, \dots, G$, $w \in \mathcal{W}$ and $\boldsymbol{v} \in \boldsymbol{V}$, we can write the regression function (8.35) also as

$$m(g,\, w,\, \boldsymbol{v}) = m_g(w,\, \boldsymbol{v}) = \beta_0 \;+\; \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z \;+\; \beta_1^W\, s_1^W(w) + \dots + \beta_{l-1}^W\, s_{l-1}^W(w)$$
$$+\; s_1^W(w)\, \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet 1}^{ZW} + \dots + s_{l-1}^W(w)\, \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet l-1}^{ZW} \;+\; m_{\boldsymbol{V}}(\boldsymbol{v}). \quad (8.36)$$

A useful interpretation of the regression function (8.36) is obtained if we view $m_g(w,\, \boldsymbol{v})$ as a conditional (given $Z = g$) regression function that describes dependence of the response expectation on the covariates W and $\boldsymbol{V}$ in group g and write it as

$$m_g(w,\, \boldsymbol{v}) = \underbrace{\big\{\beta_0 \;+\; \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z \;+\; m_{\boldsymbol{V}}(\boldsymbol{v})\big\}}_{=:\,\gamma_{g,0}^W(\boldsymbol{v})}$$
$$+\; \underbrace{\big(\beta_1^W + \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet 1}^{ZW}\big)}_{=:\,\gamma_{g,1}^W}\, s_1^W(w) \;+\; \dots \;+\; \underbrace{\big(\beta_{l-1}^W + \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet l-1}^{ZW}\big)}_{=:\,\gamma_{g,l-1}^W}\, s_{l-1}^W(w). \quad (8.37)$$

Expression (8.37) shows that a linear model with an interaction between a numeric covariate parameterized using a parameterization $\boldsymbol{s}_W$ and a categorical covariate can be interpreted such that for any fixed $Z = g$, the covariates W and $\boldsymbol{V}$ act additively and the effect of W on the response expectation is expressed by its parameterization $\boldsymbol{s}_W$. Nevertheless, the related coefficients depend on a value of a categorical covariate Z. The intercept term is further additively modified by a factor $m_{\boldsymbol{V}}(\boldsymbol{v})$.

In other words, if the categorical covariate Z expresses pertinence of a subject/experimental unit into one of G groups (internally labeled by numbers $1, \ldots, G$), the regression function (8.37) of the interaction model parameterizes a situation when, given the remaining covariates $\boldsymbol{V}$, dependence of the response expectation on the numeric covariate W can be in each of the G groups expressed by the same linear model (parameterized by the parameterization $\boldsymbol{s}_W$), nevertheless, the regression coefficients of the G linear models may differ. It follows from (8.37) that given $Z = g$ (and given $\boldsymbol{V} = \boldsymbol{v}$), the regression coefficients for the dependence of the response on the numeric covariate W expressed by the parameterization $\boldsymbol{s}_W$ are

$$\gamma^W_{g,0}(\boldsymbol{v}) \;=\; \beta_0 \;+\; \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z \;+\; m_{\boldsymbol{V}}(\boldsymbol{v}), \tag{8.38}$$

$$\gamma^W_{g,j} \;=\; \beta^W_j \;+\; \boldsymbol{c}_g^\top \boldsymbol{\beta}^{ZW}_{\bullet j}, \qquad\qquad j = 1, \ldots, l-1. \tag{8.39}$$

Chosen (pseudo)contrasts that parameterize a categorical covariate Z then determine interpretation of the intercept β_0, both sets of main effect $\boldsymbol{\beta}^Z$ and $\boldsymbol{\beta}^W$ and also the interaction effects $\boldsymbol{\beta}^{ZW}$. This interpretation is now a straightforward generalization of derivations shown earlier in Sections 7.4.4 and 8.3.

- Interpretation of the intercept term β_0 and the main effects $\boldsymbol{\beta}^Z$ of the categorical covariate Z is obtained by noting correspondence between the expression of the group specific intercepts $\gamma^W_{1,0}(\boldsymbol{v}), \ldots, \gamma^W_{G,0}(\boldsymbol{v})$ given by (8.38) and the conditional group means (8.8) in Section 8.3.
- Analogously, interpretation of the main effects $\boldsymbol{\beta}^W$ and the interaction effects $\boldsymbol{\beta}^{ZW}$ is obtained by noting that for each $j = 1, \ldots, l-1$, the group specific "slopes" $\gamma^W_{1,j}, \ldots, \gamma^W_{G,j}$ given by (8.39) play a role of the group specific means (7.19) in Section 7.4.4.

Example 8.3 (Reference group pseudocontrasts).

Suppose that $\mathbb{C}$ is the reference group pseudocontrast matrix (7.22). While viewing the group specific intercepts (8.38) as the conditional (given $\boldsymbol{V} = \boldsymbol{v}$) group means (8.8), we obtain, analogously to Example 8.1, the following interpretation of the intercept term β_0 and the main effects $\boldsymbol{\beta}^Z = \left(\beta^Z_1, \ldots, \beta^Z_{G-1}\right)^\top$ of the categorical covariate:

$$\begin{aligned}
\beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v}) &= \gamma^W_{1,0}(\boldsymbol{v}),\\
\beta^Z_1 &= \gamma^W_{2,0}(\boldsymbol{v}) - \gamma^W_{1,0}(\boldsymbol{v}),\\
&\vdots\\
\beta^Z_{G-1} &= \gamma^W_{G-1,0}(\boldsymbol{v}) - \gamma^W_{1,0}(\boldsymbol{v}).
\end{aligned}$$

If for given $j = 1, \ldots, l-1$, the group specific "slopes" $\gamma^W_{1,j}, \ldots, \gamma^W_{G,j}$ given by (8.39) are viewed as the group specific means (7.19) in Section 7.4.4, interpretation of the jth main effect β^W_j of the numeric covariate and the jth set of the interaction effects $\boldsymbol{\beta}^{ZW}_{\bullet j} = \left(\beta^{ZW}_{1,j}, \ldots, \beta^{ZW}_{G-1,j}\right)^\top$ is analogous to

expressions (7.23):

$$\begin{aligned}\beta_j^W &= \gamma_{1,j}^W,\\ \beta_{1,j}^{ZW} &= \gamma_{2,j}^W - \gamma_{1,j}^W,\\ &\vdots\\ \beta_{G-1,j}^{ZW} &= \gamma_{G-1,j}^W - \gamma_{1,j}^W.\end{aligned}$$

End of Lecture #13 (10/11/2016)
Partial start of Lecture #14 (16/11/2016)

Example 8.4 (Sum contrasts).

Suppose now that $\mathbb{C}$ is the sum contrast matrix (7.24). Again, while viewing the group specific intercepts (8.38) as the conditional (given $\boldsymbol{V} = \boldsymbol{v}$) group means (8.8), we obtain, now analogously to Example 8.2, the following interpretation of the intercept term β_0 and the main effects $\boldsymbol{\beta}^Z = \left(\beta_1^Z, \ldots, \beta_{G-1}^Z\right)^\top$ of the categorical covariate:

$$\begin{aligned}\beta_0 + m_{\boldsymbol{V}}(\boldsymbol{v}) &= \overline{\gamma_0}^W(\boldsymbol{v}),\\ \beta_1^Z &= \gamma_{1,0}^W(\boldsymbol{v}) - \overline{\gamma_0}^W(\boldsymbol{v}),\\ &\vdots\\ \beta_{G-1}^Z &= \gamma_{G-1,0}^W(\boldsymbol{v}) - \overline{\gamma_0}^W(\boldsymbol{v}),\end{aligned}$$

where

$$\overline{\gamma_0}^W(\boldsymbol{v}) = \frac{1}{G}\sum_{g=1}^G \gamma_{g,0}^W(\boldsymbol{v}), \qquad \boldsymbol{v} \in \mathcal{V}.$$

If for given $j = 1, \ldots, l-1$, the group specific "slopes" $\gamma_{1,j}^W, \ldots, \gamma_{G,j}^W$ given by (8.39) are viewed as the group specific means (7.19) in Section 7.4.4, interpretation of the jth main effect β_j^W of the numeric covariate and the jth set of the interaction effects $\boldsymbol{\beta}_{\bullet j}^{ZW} = \left(\beta_{1,j}^{ZW}, \ldots, \beta_{G-1,1}^{ZW}\right)^\top$ is analogous to expression (7.26):

$$\begin{aligned}\beta_j^W &= \overline{\gamma_j}^W,\\ \beta_{1,j}^{ZW} &= \gamma_{1,j}^W - \overline{\gamma_j}^W,\\ &\vdots\\ \beta_{G-1,j}^{ZW} &= \gamma_{G-1,j}^W - \overline{\gamma_j}^W,\end{aligned}$$

where

$$\overline{\gamma_j}^W = \frac{1}{G}\sum_{g=1}^G \gamma_{g,j}^W.$$

Alternative, interpretation of the regression function (8.36) is obtained if for a fixed $w \in \mathcal{W}$ and $\boldsymbol{v} \in \boldsymbol{V}$, the values of $m_g(w, \boldsymbol{v})$, $g = 1, \ldots, G$, are viewed as conditional (given W and $\boldsymbol{V}$) group means. Expression (8.36) can then be rewritten as

$$m_g(w, \boldsymbol{v}) = \underbrace{\left\{\beta_0 + \boldsymbol{s}_W^\top(w)\boldsymbol{\beta}^W + m_{\boldsymbol{V}}(\boldsymbol{v})\right\}}_{=:\, \gamma_0^Z(w,\, \boldsymbol{v})} + \boldsymbol{c}_g^\top \underbrace{\left\{\boldsymbol{\beta}^Z + s_1^W(w)\boldsymbol{\beta}_{\bullet 1}^{ZW} + \cdots + s_{l-1}^W(w)\boldsymbol{\beta}_{\bullet l-1}^{ZW}\right\}}_{\boldsymbol{\gamma}^{Z\star}(w)}. \tag{8.40}$$

That is, the vector $\boldsymbol{m}(w, \boldsymbol{v})$ is parameterized as

$$\boldsymbol{m}(w, \boldsymbol{v}) = \gamma_0^Z(w, \boldsymbol{v})\mathbf{1}_G + \mathbb{C}\boldsymbol{\gamma}^{Z\star}(w).$$

And the related coefficients $\gamma_0^Z(w, \boldsymbol{v})$, $\boldsymbol{\gamma}^{Z\star}(w)$ depend on w by a linear model parameterized by the parameterization $\boldsymbol{s}_W$, the intercept term is further additively modified by $m_{\boldsymbol{V}}(\boldsymbol{v})$. Expression (8.40) perhaps provide a way for the interpretation of the intercept term β_0 and the main effects $\boldsymbol{\beta}^Z$. Nevertheless, attempts to use (8.40) for interpretation of the main effects $\boldsymbol{\beta}^W$ and the interaction effects $\boldsymbol{\beta}^{ZW}$ are usually quite awkward.

8.6.2 Categorical effect modification with regression splines

Suppose now that the numeric covariate W is parameterized using the regression splines

$$\boldsymbol{B}_W = \left(B_1^W, \ldots, B_l^W\right)^\top$$

with the related model matrix $\mathbb{B}^W$ that we again factorize into blocks $\mathbb{B}_1^W, \ldots, \mathbb{B}_G^W$ that correspond to datapoints with $Z = 1, \ldots, Z = G$. That is, for $g = 1, \ldots, G$, matrix $\mathbb{B}_g^W$ is an $n_g \times l$ matrix,

$$\mathbb{B}_g^W = \begin{pmatrix} \boldsymbol{B}_W^\top(W_{g,1}) \\ \vdots \\ \boldsymbol{B}_W^\top(W_{g,n_g}) \end{pmatrix} = \begin{pmatrix} B_1^W(W_{g,1}) & \ldots & B_l^W(W_{g,1}) \\ \vdots & \vdots & \vdots \\ B_1^W(W_{g,n_g}) & \ldots & B_l^W(W_{g,n_g}) \end{pmatrix} \quad \text{and} \quad \mathbb{B}^W = \begin{pmatrix} \mathbb{B}_1^W \\ \vdots \\ \mathbb{B}_G^W \end{pmatrix}.$$

Let $\mathbb{B}^{ZW} = \mathbb{S}^Z : \mathbb{B}^W$ which is an $n \times (G-1)l$ matrix:

$$\mathbb{B}^{ZW} = \left(\begin{array}{ccc|c|ccc} c_{1,1}\,B_1^W(W_{1,1}) & \ldots & c_{1,G-1}\,B_1^W(W_{1,1}) & \ldots & c_{1,1}\,B_l^W(W_{1,1}) & \ldots & c_{1,G-1}\,B_l^W(W_{1,1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{1,1}\,B_1^W(W_{1,n_1}) & \ldots & c_{1,G-1}\,B_1^W(W_{1,n_1}) & \ldots & c_{1,1}\,B_l^W(W_{1,n_1}) & \ldots & c_{1,G-1}\,B_l^W(W_{1,n_1}) \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline c_{G,1}\,B_1^W(W_{G,1}) & \ldots & c_{G,G-1}\,B_1^W(W_{G,1}) & \ldots & c_{G,1}\,B_l^W(W_{G,1}) & \ldots & c_{G,G-1}\,B_l^W(W_{G,1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{G,1}\,B_1^W(W_{G,n_G}) & \ldots & c_{G,G-1}\,B_1^W(W_{G,n_G}) & \ldots & c_{G,1}\,B_l^W(W_{G,n_G}) & \ldots & c_{G,G-1}\,B_l^W(W_{G,n_1}) \end{array}\right)$$

$$= \begin{pmatrix} B_1^W(W_{1,1})\,\boldsymbol{c}_1^\top & \ldots & B_l^W(W_{1,1})\,\boldsymbol{c}_1^\top \\ \vdots & \vdots & \vdots \\ B_1^W(W_{1,n_1})\,\boldsymbol{c}_1^\top & \ldots & B_l^W(W_{1,n_1})\,\boldsymbol{c}_1^\top \\ \hline \vdots & \vdots & \vdots \\ \hline B_1^W(W_{G,1})\,\boldsymbol{c}_G^\top & \ldots & B_l^W(W_{G,1})\,\boldsymbol{c}_G^\top \\ \vdots & \vdots & \vdots \\ B_1^W(W_{G,n_G})\,\boldsymbol{c}_G^\top & \ldots & B_l^W(W_{G,n_G})\,\boldsymbol{c}_G^\top \end{pmatrix} = \begin{pmatrix} \mathbb{B}_1^W \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \mathbb{B}_G^W \otimes \boldsymbol{c}_G^\top \end{pmatrix}.$$

As in Section 8.5.3, remember that for any $w \in \mathcal{W}$, $\sum_{j=1}^l B_j^W(w) = 1$ from which it follows that

(i) $\mathbf{1}_n \in \mathbb{B}^W$;

(ii) $\mathcal{M}(\mathbb{S}^Z) \subseteq \mathcal{M}(\mathbb{B}^{ZW})$.

Hence

$$\mathcal{M}\Big((\mathbf{1}_n,\, \mathbb{S}^Z,\, \mathbb{B}^W,\, \mathbb{B}^{ZW})\Big) \;=\; \mathcal{M}\Big((\mathbb{B}^W,\, \mathbb{B}^{ZW})\Big)$$

and if a full-rank linear model is to be obtained that includes an interaction between a numeric covariate parameterized using the regression splines and a categorical covariate parameterized by the reparameterizing matrix $\mathbb{S}^Z$ derived from a (pseudo)contrast matrix $\mathbb{C}$, the model matrix $\mathbb{X}^{ZW}$ that parameterizes the term $m_{ZW}(z,\, w)$ in the regression function (8.33) is

$$\mathbb{X}^{ZW} = \big(\mathbb{B}^W,\, \mathbb{B}^{ZW}\big) = \begin{pmatrix} \mathbb{B}_1^W, & \mathbb{B}_1^W \otimes \boldsymbol{c}_1^\top \\ \vdots & \vdots \\ \mathbb{B}_G^W, & \mathbb{B}_G^W \otimes \boldsymbol{c}_G^\top \end{pmatrix}.$$

The regression coefficients related to the model matrix $\mathbb{X}^{ZW}$ are

$$\boldsymbol{\beta} = \big(\underbrace{\beta_1^W, \ldots, \beta_l^W}_{=:\, \boldsymbol{\beta}^W},\, \underbrace{\underbrace{\beta_{1,1}^{ZW}, \ldots, \beta_{G-1,1}^{ZW}}_{=:\, \boldsymbol{\beta}_{\bullet 1}^{ZW}}, \ldots, \underbrace{\beta_{1,k}^{ZW}, \ldots, \beta_{G-1,k}^{ZW}}_{=:\, \boldsymbol{\beta}_{\bullet k}^{ZW}}}_{=:\, \boldsymbol{\beta}^{ZW}}\big)^\top.$$

The value of the regression function (8.19) for $z = g$, $g = 1, \ldots, G$, $w \in \mathcal{W}$, and $\boldsymbol{v} \in \boldsymbol{V}$, i.e., the values of the conditional regression functions (8.34) can then be written as

$$\begin{aligned} m_g(w,\, \boldsymbol{v}) = {} & \beta_1^W\, B_1^W(w) + \cdots + \beta_l^W\, B_l^W(w) \\ & +\; B_1^W(w)\, \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet 1}^{ZW} + \cdots + B_l^W(w)\, \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet l}^{ZW} \;+\; m_{\boldsymbol{V}}(\boldsymbol{v}). \end{aligned}$$

Its useful interpretation is obtained if we write it as

$$m_g(w,\, \boldsymbol{v}) = \underbrace{\big(\beta_1^W + \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet 1}^{ZW}\big)}_{=:\, \gamma_{g,1}^W} B_1^W(w) \;+\; \cdots \;+\; \underbrace{\big(\beta_l^W + \boldsymbol{c}_g^\top \boldsymbol{\beta}_{\bullet l}^{ZW}\big)}_{=:\, \gamma_{g,k}^W} B_l^W(w) \;+\; m_{\boldsymbol{V}}(\boldsymbol{v}),$$

which shows that the underlying linear model assumes that given $Z = g$, the covariates W and $\boldsymbol{V}$ act additively and the effect of the numeric covariate W on the response expectation is described by the regression spline whose coefficients $\gamma_{g,1}^W, \ldots, \gamma_{g,k}^W$, however, depend on the value of the categorical covariate Z. Analogously to Section 8.6.1, interpretation of the regression coefficients $\boldsymbol{\beta}^W$ and $\boldsymbol{\beta}^{ZW}$ depends on chosen (pseudo)contrasts used to parameterize the categorical covariate Z.

8.7 Interaction of two categorical covariates

Finally, we could consider a situation when both Z and W are categorical covariates with

$$\mathcal{Z} = \{1, \ldots, G\}, \qquad \mathcal{W} = \{1, \ldots, H\}.$$

Analogously to Section 8.6, each of the two categorical covariates can classically be parameterized by the means of (pseudo)contrasts, not necessarily of the same type for the two covariates at hand. The interaction part of the model matrix is then created in the same way as before. Nevertheless, we postpone more detailed discussion of the meaning of the related interaction terms into Chapter 9, and in particular into its Section 9.2 which deals with so called two-way classification.

End of Lecture #14 (16/11/2016)

8.8 Hierarchically well-formulated models, ANOVA tables

8.8.1 Model terms

Start of Lecture #15 (16/11/2016)

In majority of applications of a linear model, a particular covariate $Z \in \mathcal{Z} \subseteq \mathbb{R}$ enters the regression function using one of the parameterizations described in Sections 7.3 and 7.4 or inside an interaction (see Defition 8.2) or inside a so called *higher order* interaction (will be defined in a while). As a summary, depending on whether the covariate is numeric or categorical, several parameterizations $\boldsymbol{s}$ were introduced in Sections 7.3 and 7.4 that with the covariate values $Z_1, \ldots, Z_n$ in the data lead to a reparameterizing matrix

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{s}^\top(Z_1) \\ \vdots \\ \boldsymbol{s}^\top(Z_n) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix},$$

where $\boldsymbol{X}_1 = \boldsymbol{s}(Z_1), \ldots, \boldsymbol{X}_n = \boldsymbol{s}(Z_n)$ are the regressors used in the linear model. The considered parameterizations were the following.

Numeric covariate

(i) **Simple transformation**: $\boldsymbol{s} = s : \mathcal{Z} \longrightarrow \mathbb{R}$ with

$$\mathbb{S} = \begin{pmatrix} s(Z_1) \\ \vdots \\ s(Z_n) \end{pmatrix} = (\boldsymbol{S}), \qquad \begin{aligned} \boldsymbol{X}_1 = X_1 &= s(Z_1), \\ &\vdots \\ \boldsymbol{X}_n = X_n &= s(Z_n). \end{aligned} \tag{8.41}$$

(ii) **Polynomial**: $\boldsymbol{s} = \left(s_1, \ldots, s_{k-1}\right)^\top$ such that $s_j(z) = P^j(z)$ is polynomial in z of degree j, $j = 1, \ldots, k-1$. This leads to

$$\mathbb{S} = \begin{pmatrix} P^1(Z_1) & \ldots & P^{k-1}(Z_1) \\ \vdots & \vdots & \vdots \\ P^1(Z_n) & \ldots & P^{k-1}(Z_n) \end{pmatrix} = \left(\boldsymbol{P}^1, \ \ldots, \ \boldsymbol{P}^{k-1}\right), \tag{8.42}$$

$$\begin{aligned} \boldsymbol{X}_1 &= \left(P^1(Z_1), \ldots, P^{k-1}(Z_1)\right)^\top, \\ &\vdots \\ \boldsymbol{X}_n &= \left(P^1(Z_n), \ldots, P^{k-1}(Z_n)\right)^\top. \end{aligned}$$

For a particular form of the basis polynomials $P^1, \ldots, P^{k-1}$, raw or orthonormal polynomials have been suggested in Sections 7.3.2 and 7.3.3. Other choices are possible as well.

(iii) **Regression spline**: $\boldsymbol{s} = \left(s_1, \ldots, s_k\right)^\top$ such that $s_j(z) = B_j(z)$, $j = 1, \ldots, k$, where $B_1, \ldots, B_k$ is the spline basis of chosen degree $d \in \mathbb{N}_0$ composed of basis B-splines built above a set of chosen knots $\boldsymbol{\lambda} = \left(\lambda_1, \ldots, \lambda_{k-d+1}\right)^\top$. This leads to

$$\mathbb{S} = \mathbb{B} = \begin{pmatrix} B_1(Z_1) & \ldots & B_k(Z_1) \\ \vdots & \vdots & \vdots \\ B_1(Z_n) & \ldots & B_k(Z_n) \end{pmatrix} = \left(\boldsymbol{B}^1, \ \ldots, \ \boldsymbol{B}^k\right), \tag{8.43}$$

$$\begin{aligned} \boldsymbol{X}_1 &= \big(B_1(Z_1), \ldots, B_k(Z_1)\big)^\top, \\ &\vdots \\ \boldsymbol{X}_n &= \big(B_1(Z_n), \ldots, B^k(Z_n)\big)^\top. \end{aligned}$$

Categorical covariate with $\mathcal{Z} = \{1, \ldots, G\}$. The parameterization $\boldsymbol{s}$ is $\boldsymbol{s}(z) = \boldsymbol{c}_z$, $z \in \mathcal{Z}$, where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_G \in \mathbb{R}^{G-1}$ are the rows of a chosen (pseudo)contrast matrix $\mathbb{C}_{G\times G-1}$. This leads to

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{c}_{Z_1}^\top \\ \vdots \\ \boldsymbol{c}_{Z_n}^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{C}^1, & \ldots, & \boldsymbol{C}^{G-1} \end{pmatrix}, \qquad \begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{c}_{Z_1}, \\ &\vdots \\ \boldsymbol{X}_n &= \boldsymbol{c}_{Z_n}. \end{aligned} \tag{8.44}$$

Main effect model terms

In the following, we restrict ourselves only into situations when the considered covariates are parameterized by one of above mentioned ways. The following definitions define sets of the columns of a possible model matrix which will be called the *model terms* and which are useful to be always considered "together" when proposing a linear model for a problem at hand.

Definition 8.3 The main effect model term.

Depending on a chosen parameterization, the main effect model term[6] *(of order one) of a given covariate Z is defined as a matrix $\mathbb{T}$ with columns:*

Numeric covariate

(i) ***Simple transformation****: (the only) column $\boldsymbol{S}$ of the reparameterizing matrix $\mathbb{S}$ given by (8.41), i.e.,*

$$\mathbb{T} = \big(\boldsymbol{S}\big).$$

(ii) ***Polynomial****: the first column $\boldsymbol{P}^1$ of the reparameterizing matrix $\mathbb{S}$ (given by Eq. 8.42) that corresponds to the linear transformation of the covariate Z, i.e.,*

$$\mathbb{T} = \big(\boldsymbol{P}^1\big).$$

(iii) ***Regression spline****: (all) columns $\boldsymbol{B}^1, \ldots, \boldsymbol{B}^k$ of the reparameterizing matrix $\mathbb{S} = \mathbb{B}$ given by (8.43), i.e.,*

$$\mathbb{T} = \big(\boldsymbol{B}^1, \ldots, \boldsymbol{B}^k\big).$$

Categorical covariate: *(all) columns $\boldsymbol{C}^1, \ldots, \boldsymbol{C}^{G-1}$ of the reparameterizing matrix $\mathbb{S}$ given by (8.44), i.e.,*

$$\mathbb{T} = \big(\boldsymbol{C}^1, \ldots, \boldsymbol{C}^{G-1}\big).$$

Definition 8.4 The main effect model term of order j.

If a numeric *covariate Z is parameterized using the polynomial of degree $k-1$ then* the main effect model term of order j, $j = 2, \ldots, k-1$, *means a matrix $\mathbb{T}^j$ whose the only column is the jth*

[6] *hlavní efekt*

column $\boldsymbol{P}^j$ of the reparameterizing matrix $\mathbb{S}$ (given by Eq. 8.42) that corresponds to the polynomial of degree j, i.e.,

$$\mathbb{T}^j = \left(\boldsymbol{P}^j\right).$$

Note. The terms $\mathbb{T}$, ..., $\mathbb{T}^{j-1}$ are called as *lower order* terms included in the term $\mathbb{T}^j$.

Two-way interaction model terms

In the following, consider two covariates Z and W and their main effect model terms $\mathbb{T}_Z$ and $\mathbb{T}_W$.

Definition 8.5 The two-way interaction model term.
The two-way interaction[7] *model term means a matrix $\mathbb{T}^{ZW}$, where*

$$\mathbb{T}^{ZW} := \mathbb{T}_Z : \mathbb{T}_W.$$

Notes.

- The main effect model term $\mathbb{T}_Z$ and/or the main effect model term $\mathbb{T}_W$ that enter the two-way interaction may also be of a degree $j > 1$.
- Both the main effect model terms $\mathbb{T}_Z$ and $\mathbb{T}_W$ are called as *lower order* terms included in the two-way interaction term $\mathbb{T}_Z : \mathbb{T}_W$.

Higher order interaction model terms

In the following, consider three covariates Z, W and V and their main effect model terms $\mathbb{T}_Z$, $\mathbb{T}_W$, $\mathbb{T}_V$.

Definition 8.6 The three-way interaction model term.
The three-way interaction[8] *model term means a matrix $\mathbb{T}^{ZWV}$, where*

$$\mathbb{T}^{ZWV} := \left(\mathbb{T}_Z : \mathbb{T}_W\right) : \mathbb{T}_V.$$

Notes.

- Any of the main effect model terms $\mathbb{T}_Z$, $\mathbb{T}_W$, $\mathbb{T}_V$ that enter the three-way interaction may also be of a degree $j > 1$.
- All main effect terms $\mathbb{T}_Z$, $\mathbb{T}_W$ and $\mathbb{T}_V$ and also all two-way interaction terms $\mathbb{T}_Z : \mathbb{T}_W$, $\mathbb{T}_Z : \mathbb{T}_V$ and $\mathbb{T}_W : \mathbb{T}_V$ are called as *lower order* terms included in the three-way interaction term $\mathbb{T}^{ZWV}$.
- By induction, we could define also four-way, five-way, ..., i.e., *higher order* interaction model terms and a notion of corresponding lower order nested terms.

---

[7] *dvojná interakce* [8] *trojná interakce*

8.8.2 Model formula

To write concisely linear models based on several covariates, the *model formula* is used. The following symbols in the model formula have the following meaning:

- **1**:
  intercept term in the model if this is the only term in the model (i.e., intercept only model).
- **Letter** or **abbreviation**:
  main effect of order one of a particular covariate (which is identified by the letter or abbreviation). It is assumed that chosen parameterization is either known from context or is indicated in some way (e.g., by the used abbreviation). Letters or abbreviations will also be used to indicate a response variable.
- **Power** of j, $j > 1$ (above a letter or abbreviation):
  main effect of order j of a particular covariate.
- **Colon (:)** between two or more letters or abbreviations:
  interaction term based on particular covariates.
- **Plus sign (+)**:
  a delimiter of the model terms.
- **Tilde ($\sim$)**:
  a delimiter between the response and description of the regression function.

Further, when using a model formula, it is assumed that the intercept term is explicitely included in the regression function. If the explicit intercept should not be included, this will be indicated by writing $-\mathbf{1}$ among the model terms.

8.8.3 Hierarchically well formulated model

Definition 8.7 Hierarchically well formulated model.

Hierarchically well formulated (HWF) model[9] *is such a model that contains an intercept term (possibly implicitely) and with each model term also all lower order terms that are nested in this term.*

Notes.

- Unless there is some well-defined specific reason, models used in practice should be *hierarchically well formulated.*
- Reason for use of the HWF models is the fact that the regression space of such models is invariant towards linear (location-scale) transformations of the regressors where invariance is meant with respect to possibility to obtain the equivalent linear models.

Example 8.5.

Consider a quadratic regression function

$$m_x(x) = \beta_0 + \beta_1\, x + \beta_2\, x^2$$

and perform a linear transformation of the regressor:

$$x = \delta\,(t - \varphi), \qquad t = \varphi + \frac{x}{\delta}, \tag{8.45}$$

[9] *hierarchicky dobře formulovaný model*

where $\delta \neq 0$ and $\varphi \neq 0$ are pre-specified constants and t is a new regressor. The regression function in t is

$$m_t(t) = \gamma_0 + \gamma_1\, t + \gamma_2\, t^2,$$

where $\gamma_0 = \beta_0 - \beta_1\delta\varphi + \beta_2\delta^2\varphi^2$,
$\gamma_1 = \beta_1\delta - 2\beta_2\delta^2\varphi$,
$\gamma_2 = \beta_2\delta^2$.

With at least three different x values in the data, both regression functions lead to two equivalent linear models of rank 3.

Suppose now that the initial regression function m_x did not include a linear term, i.e., it was

$$m_x(x) = \beta_0 + \beta_2\, x^2$$

which leads to a linear model of rank 2 (with at least three or even two different covariate values in data). Upon performing the linear transformation (8.45) of the regressor x, the regression function becomes

$$m_t(t) = \gamma_0 + \gamma_1\, t + \gamma_2\, t^2$$

with $\gamma_0 = \beta_0 + \beta_2\delta^2\varphi^2$,
$\gamma_1 = -2\beta_2\delta^2\varphi$,
$\gamma_2 = \beta_2\delta^2$.

With at least three different covariate values in data, this leads to the linear model of rank 3.

To use a non-HWF model in practice, there should always be a (physical, ...) reason for that. For example,

- No intercept in the model $\equiv$ it can be assumed that the response expectation is zero if all regressors in a chosen parameterization take zero values.
- No linear term in a model with a quadratic regression function $m(x) = \beta_0 + \beta_2\, x^2 \equiv$ it can be assumed that the regression function is a parabola with the vertex in a point $(0,\, \beta_0)$ with respect to the x parameterization.
- No main effect of one covariate in an interaction model with two numeric covariates and a regression function $m(x,\, z) = \beta_0 + \beta_1\, z + \beta_2\, x\, z \equiv$ it can be assumed that with $z = 0$, the response expectation does not depend on a value of x, i.e., $\mathbb{E}\big(Y \,\big|\, X = x,\, Z = 0\big) = \beta_0$ (a constant).

8.8.4 ANOVA tables

For a particular linear model, so called ANOVA tables are often produced to help the analyst to decide which model terms are important with respect to its influence on the response expectation. Similarly to well known one-way ANOVA table (see any of introductory statistical courses and also Section 9.1), ANOVA tables produced in a context of linear models provide on each row input of a certain F-statistic, now that based on Theorem 5.2. The last row of the table (labeled often as *Residual*, *Error* or *Within*) provides

(i) residual degrees of freedom ν_e of the considered model;

(ii) residual sum of squares SS_e of the considered model;

(iii) residual mean square $\mathsf{MS}_e = \mathsf{SS}_e/\nu_e$ of the considered model.

Each of the remaining rows of the ANOVA table provides input for the numerator of the F-statistic that corresponds to comparison of certain two models $\mathsf{M}_1 \subset \mathsf{M}_2$ which are both submodels of the considered model (or M_2 is the considered model itself) and which have ν_1 and ν_2 degrees of freedom, respectively. The following quantities are provided on each of the remaining rows of the ANOVA table:

(i) degrees of freedom for the numerator of the F-statistic (*effect degrees of freedom* $\nu_E = \nu_1 - \nu_2$);

(ii) difference in the residual sum of squares of the two models (*effect sum of squares* $\mathsf{SS}_E = \mathsf{SS}\big(\mathsf{M}_2 \,\big|\, \mathsf{M}_1\big)$);

(iii) ratio of the above two values which is the numerator of the F-statistic (*effect mean square* $\mathsf{MS}_E = \mathsf{SS}_E/\nu_E$);

(iv) value of the F-statistic $F_E = \mathsf{MS}_E/\mathsf{MS}_e$;

(v) a p-value based on the F-statistic F_E and the $\mathcal{F}_{\nu_E,\,\nu_e}$ distribution.

Several types of the ANOVA tables are distinguished which differ by definition of a pair of the two models M_1 and M_2 that are being compared on a particular row. Consequently, interpretation of results provided by the ANOVA tables of different type differs. Further, it is important to know that in all ANOVA tables, the lower order terms always appear on earlier rows in the table than the higher order terms that include them. Finally, for some ANOVA tables, different interpretation of the results is obtained for different ordering of the rows with the terms of the same hierarchical level, e.g., for different ordering of the main effect terms. We introduce ANOVA tables of three types which are labeled by the R software (and by many others as well) as tables of type I, II or III (arabic numbers can be used as well). Nevertheless, note that there exist software packages and literature that use different typology. In the reminder of this section we assume that intercept term is included in the considered model.

In the following, we illustrate each type of the ANOVA table on a linear model based on two covariates whose main effect terms will be denoted as A and B. Next to the main effects, the model will include also an interaction term $\mathsf{A}:\mathsf{B}$. That is, the model formula of the considered model, denoted as M_{AB} is $\sim \mathsf{A} + \mathsf{B} + \mathsf{A}:\mathsf{B}$. In total, the following (sub)models of this model will appear in the ANOVA tables:

$$\begin{aligned}
&\mathsf{M}_0\text{:} && \sim 1,\\
&\mathsf{M}_A\text{:} && \sim \mathsf{A},\\
&\mathsf{M}_B\text{:} && \sim \mathsf{B},\\
&\mathsf{M}_{A+B}\text{:} && \sim \mathsf{A} + \mathsf{B},\\
&\mathsf{M}_{AB}\text{:} && \sim \mathsf{A} + \mathsf{B} + \mathsf{A}:\mathsf{B}.
\end{aligned}$$

The symbol $\mathsf{SS}\big(F_2 \,\big|\, F_1\big)$ will denote a difference in the residual sum of squares of the models with model formulas F_1 and F_2.

Type I (sequential) ANOVA table

Example 8.6 (Type I ANOVA table for model $\mathsf{M}_{AB} :\sim \mathsf{A} + \mathsf{B} + \mathsf{A} : \mathsf{B}$).

In the type I ANOVA table, the presented results depend on the ordering of the rows with the terms of the same hierarchical level. In this example, those are the rows that correspond to the main effect terms A *and* B.

Order A + B + A:B

Effect (Term)	*Degrees of freedom*	*Effect sum of squares*	*Effect mean square*	*F-stat.*	*P-value*
A	$\star$	$\mathsf{SS}(\mathsf{A} \mid 1)$	$\star$	$\star$	$\star$
B	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{A})$	$\star$	$\star$	$\star$
$\mathsf{A}:\mathsf{B}$	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B} \mid \mathsf{A}+\mathsf{B})$	$\star$	$\star$	$\star$
Residual	ν_e	SS_e	MS_e		

Order B + A + A:B

Effect (Term)	*Degrees of freedom*	*Effect sum of squares*	*Effect mean square*	*F-stat.*	*P-value*
B	$\star$	$\mathsf{SS}(\mathsf{B} \mid 1)$	$\star$	$\star$	$\star$
A	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{B})$	$\star$	$\star$	$\star$
$\mathsf{A}:\mathsf{B}$	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B} \mid \mathsf{A}+\mathsf{B})$	$\star$	$\star$	$\star$
Residual	ν_e	SS_e	MS_e		

The row of the effect (term) E in the type I ANOVA table has in general the following interpretation and properties.

- It compares two models $\mathsf{M}_1 \subset \mathsf{M}_2$, where
  - M_1 contains all terms included in the rows that precede the row of the term E.
  - M_2 contains the terms of model M_1 and additionally the term E.
- The sum of squares shows increase of the explained variability of the response due to the term E on top of the terms shown on the preceding rows.
- The p-value provides a significance of the influence of the term E on the response while controlling (adjusting) for all terms shown on the preceding rows.
  - Interpretation of the F-tests is different for rows labeled equally A in the two tables in Example 8.6. Similarly, interpretation of the F-tests is different for rows labeled equally B in the two tables in Example 8.6.
- The sum of all sums of squares shown in the type I ANOVA table gives the total sum of squares SS_T of the considered model. This follows from the construction of the table where the terms are added *sequentially* one-by-one and from a sequential use of Theorem 5.8 (Breakdown of the total sum of squares in a linear model with intercept).

Type II ANOVA table

Example 8.7 (Type II ANOVA table for model $\mathsf{M}_{AB} :\sim \mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B}$).

In the type II ANOVA table, the presented results do not depend on the ordering of the rows with the terms of the same hierarchical level as should become clear from subsequent explanation.

Effect (Term)	*Degrees of freedom*	*Effect sum of squares*	*Effect mean square*	*F-stat.*	*P-value*
A	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{B})$	$\star$	$\star$	$\star$
B	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{A})$	$\star$	$\star$	$\star$
$\mathsf{A}:\mathsf{B}$	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B} \mid \mathsf{A}+\mathsf{B})$	$\star$	$\star$	$\star$
Residual	ν_e	SS_e	MS_e		

The row of the effect (term) E in the type II ANOVA table has in general the following interpretation and properties.

- It compares two models $\mathsf{M}_1 \subset \mathsf{M}_2$, where
  - M_1 is the considered (full) model without the term E and also all *higher order* terms than E that include E.
  - M_2 contains the terms of model M_1 and additionally the term E (this is the same as in type I ANOVA table).
- The sum of squares shows increase of the explained variability of the response due to the term E on top of all other terms that do not include the term E.
- The p-value provides a significance of the influence of the term E on the response while controlling (adjusting) for all other terms that do not include E.
- For practical purposes, this is probably the most useful ANOVA table.

Type III ANOVA table

Example 8.8 (Type III ANOVA table for model $\mathsf{M}_{AB} :\sim \mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B}$).
Also in the type III ANOVA table, the presented results do not depend on the ordering of the rows with the terms of the same hierarchical level as should become clear from subsequent explanation.

Effect (Term)	*Degrees of freedom*	*Effect sum of squares*	*Effect mean square*	*F-stat.*	*P-value*
A	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B} \mid \mathsf{B}+\mathsf{A}:\mathsf{B})$	$\star$	$\star$	$\star$
B	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B} \mid \mathsf{A}+\mathsf{A}:\mathsf{B})$	$\star$	$\star$	$\star$
$\mathsf{A}:\mathsf{B}$	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B}+\mathsf{A}:\mathsf{B} \mid \mathsf{A}+\mathsf{B})$	$\star$	$\star$	$\star$
Residual	ν_e	SS_e	MS_e		

The row of the effect (term) E in the type III ANOVA table has in general the following interpretation and properties.

- It compares two models $\mathsf{M}_1 \subset \mathsf{M}_2$, where
  - M_1 is the considered (full) model without the term E.
  - M_2 contains the terms of model M_1 and additionally the term E (this is the same as in type I and type II ANOVA table). Due to the construction of M_1, the model M_2 is always equal to the considered (full) model.

- The submodel M_1 is not necessarily hierarchically well formulated. If M_1 is not HWF, interpretation of its comparison to model M_2 depends on a parameterization of the term E. Consequently, also the interpretation of the F-test depends on the used parameterization.
- For general practical purposes, most rows of the type III ANOVA table are often useless.

Chapter 9

Analysis of Variance

In this chapter, we examine several specific issues of linear models where all covariates are *categorical*. That is, the covariate vector $\boldsymbol{Z}$ is $\boldsymbol{Z} = \left(Z_1, \ldots, Z_p\right)^\top$, $Z_j \in \mathcal{Z}_j$, $j = 1, \ldots, p$, and each $\mathcal{Z}_j$ is a finite set (with usually a "low" cardinality). The corresponding linear models are traditionally used in the area of designed (industrial, agricultural, ...) experiments or controlled clinical studies. The elements of the covariate vector $\boldsymbol{Z}$ then correspond to p factors whose influence on the response Y is of interest. The values of those factors for experimental units/subjects are typically within the control of an experimenter in which case the covariates are fixed rather than being random. Nevertheless, since the whole theory presented in this chapter is based on statements on the conditional distribution of the response given the covariate values, everything applies for both fixed and random covariates.

9.1 One-way classification

One-way classification corresponds to situation of one categorical covariate $Z \in \mathcal{Z} = \{1, \ldots, G\}$, see also Section 7.4. A linear model is then used to parameterize a set of G (conditional) response expectations $\mathbb{E}(Y \mid Z = 1)$, …, $\mathbb{E}(Y \mid Z = G)$ that we call as *one-way classified group means*:

$$m(g) = \mathbb{E}(Y \mid Z = g) =: m_g, \qquad g = 1, \ldots, G.$$

Without loss of generality, we can assume that the response random variables $Y_1, \ldots, Y_n$ are sorted such that

$$\begin{array}{lcccccl} Z_1 & = & \cdots & = & Z_{n_1} & = & 1, \\ Z_{n_1+1} & = & \cdots & = & Z_{n_1+n_2} & = & 2, \\ & & & & & \vdots & \\ Z_{n_1+\cdots+n_{G-1}+1} & = & \cdots & = & Z_n & = & G. \end{array}$$

For notational clarity in theoretical derivations, it is useful to use a double subscript to index the individual observations and to merge responses with a common covariate value $Z = g$, $g = 1, \ldots, G$, into response subvectors $\boldsymbol{Y}_g$:

$$\begin{array}{lll} Z = 1: & \boldsymbol{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})^\top & = (Y_1, \ldots, Y_{n_1})^\top, \\ \vdots & \vdots & \vdots \\ Z = G: & \boldsymbol{Y}_G = (Y_{G,1}, \ldots, Y_{G,n_G})^\top & = (Y_{n_1+\cdots+n_{G-1}+1}, \ldots, Y_n)^\top. \end{array}$$

The full response vector is $\boldsymbol{Y}$ and its (conditional, given $\mathbb{Z} = (Z_1, \ldots, Z_n)^\top$) mean are

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_G \end{pmatrix}, \qquad \mathbb{E}(\boldsymbol{Y} \mid \mathbb{Z}) = \begin{pmatrix} m_1 \mathbf{1}_{n_1} \\ \vdots \\ m_G \mathbf{1}_{n_G} \end{pmatrix} =: \boldsymbol{\mu}. \tag{9.1}$$

A standard linear model then additionally assumes

$$\mathsf{var}(\boldsymbol{Y} \mid \mathbb{Z}) = \sigma^2 \mathbf{I}_n. \tag{9.2}$$

With the i.i.d. data $(Y_i, Z_i)^\top \overset{\text{i.i.d.}}{\sim} (Y, Z)^\top$ for which (9.1) and (9.2) are assumed, the random variables $Y_{g,1}, \ldots, Y_{g,n_G}$ (elements of the vector $\boldsymbol{Y}_g$) are i.i.d. from a distribution of $Y \mid Z = g$ whose mean is m_g and the variance is σ^2. That is, the response random variables form G independent i.i.d. samples:

$$\begin{array}{lll} \text{Sample } 1: & \boldsymbol{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})^\top, & Y_{1,j} \overset{\text{i.i.d.}}{\sim} (m_1, \sigma^2), \quad j = 1, \ldots, n_1, \\ \vdots & \vdots & \vdots \\ \text{Sample } G: & \boldsymbol{Y}_G = (Y_{G,1}, \ldots, Y_{G,n_G})^\top, & Y_{G,j} \overset{\text{i.i.d.}}{\sim} (m_G, \sigma^2), \quad j = 1, \ldots, n_G. \end{array}$$

We arrive at the same situation even if the covariate values are fixed rather than random. A conceptual difference between the situation of random and fixed covariates in this setting is that with random covariates, the group sample sizes $n_1, \ldots, n_G$ are random as well, whereas with the fixed covariates, also those sample sizes are fixed. As in Section 7.4, we keep assuming that $n_1 > 0$, …, $n_G > 0$ (almost surely in case of random covariates). A linear model with the inference being conditioned by the covariate values can now be used to infere on the group means $m_1, \ldots, m_G$ or on their linear combinations.

9.1.1 Parameters of interest

Differences between the group means

The principal inferential interest with one-way classification lies in estimation of and tests on parameters

$$\theta_{g,h} = m_g - m_h, \qquad g,\, h = 1, \ldots, G,\ g \neq h,$$

which are the differences between the group means. Since each $\theta_{g,h}$ is a linear combination of the elements of the mean vector $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z})$, it is trivially an estimable parameter of the underlying linear model irrespective of its parameterization. The LSE of each $\theta_{g,h}$ is then a difference between the corresponding fitted values.

The principal null hypothesis being tested in context of the one-way classification is the null hypothesis on equality of the group means, i.e., the null hypothesis

$$\mathrm{H}_0\colon\ m_1 = \cdots = m_G,$$

which written in terms of the differences between the group means is

$$\mathrm{H}_0\colon\ \theta_{g,h} = 0, \quad g,\, h = 1, \ldots, G,\ g \neq h.$$

Factor effects

One-way classification often corresponds to a designed experiment which aims in evaluating the effect of a certain factor on the response. In that case, the following quantities, called as *factor effects*, are usually of primary interest.

Definition 9.1 Factor effects in a one-way classification.

By factor effects *in case of a one-way classification we understand the quantities* $\eta_1, \ldots, \eta_G$ *defined as*

$$\eta_g = m_g - \overline{m}, \qquad g = 1, \ldots, G,$$

where $\overline{m} = \dfrac{1}{G} \displaystyle\sum_{h=1}^{G} m_h$ *is the mean of the group means.*

Notes.

- The factor effects are again linear combinations of the elements of the mean vector $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z})$ and hence all are estimable parameters of the underlying linear model with the LSE being equal to the appropriate linear combination of the fitted values.
- Each factor effect shows how the mean of a particular group differ from the mean of all the group means.
- The null hypothesis

  $$\mathrm{H}_0\colon \eta_g = 0, \quad g = 1, \ldots, G,$$

  is equivalent to the null hypothesis $\mathrm{H}_0\colon m_1 = \cdots = m_G$ on the equality of the group means.

End of Lecture #15 (16/11/2016)

9.1.2 One-way ANOVA model

Start of Lecture #16 (23/11/2016)

As a reminder from Section 7.4.2, the regression space of the one-way classification is

$$\left\{ \begin{pmatrix} m_1 \, \mathbf{1}_{n_1} \\ \vdots \\ m_G \, \mathbf{1}_{n_G} \end{pmatrix} : \; m_1, \, \ldots, \, m_G \in \mathbb{R} \right\} \subseteq \mathbb{R}^n.$$

While assuming $n_g > 0$, $g = 1, \ldots, G$, $n > G$, its vector dimension is G. In Sections 7.4.3 and 7.4.4, we introduced two classical classes of parameterizations of this regression space and of the response mean vector $\boldsymbol{\mu}$ as $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$, $\boldsymbol{\beta} \in \mathbb{R}^k$.

ANOVA (less-than-full rank) parameterization

$$m_g \;\; = \;\; \alpha_0 + \alpha_g, \qquad g = 1, \ldots, G$$

with $k = G + 1$, $\boldsymbol{\beta} =: \boldsymbol{\alpha} = \big(\alpha_0, \; \underbrace{\boldsymbol{\alpha}^Z}_{(\alpha_1, \, \ldots, \, \alpha_G)^\top}\big)^\top$.

Full-rank parameterization

$$m_g \;\; = \;\; \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, \qquad g = 1, \ldots, G$$

with $k = G$, $\boldsymbol{\beta} = \big(\beta_0, \; \underbrace{\boldsymbol{\beta}^Z}_{(\beta_1, \, \ldots, \, \beta_{G-1})^\top}\big)^\top$,

where $\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix}$ is a chosen $G \times (G-1)$ (pseudo)contrast matrix.

Note. If the parameters in the ANOVA parameterization are identified by the sum constraint $\sum_{g=1}^G \alpha_g = 0$, we get

$$\begin{aligned} \alpha_0 \;\; &= \frac{1}{G} \sum_{g=1}^{G} m_g \qquad\qquad &&= \overline{m}, \\ \alpha_g \;\; &= \eta_g = m_g - \frac{1}{H} \sum_{h=1}^{H} m_h \qquad &&= m_g - \overline{m}, \end{aligned}$$

that is, parameters $\alpha_1, \, \ldots, \, \alpha_G$ are then equal to the factor effects.

Terminology. Related linear model is referred to as *one-way ANOVA model*[1]

[1] *model analýzy rozptylu jednoduchého třídění*

Notes.

- Depending on chosen parameterization (ANOVA or full-rank) the differences between the group means, parameters $\theta_{g,h}$, are expressed as

$$\theta_{g,h} = \alpha_g - \alpha_h = \big(\boldsymbol{c}_g - \boldsymbol{c}_h\big)^\top \boldsymbol{\beta}^Z, \qquad g \neq h.$$

  The null hypothesis $\mathrm{H}_0 \colon m_1 = \cdots m_G$ on equality of the group means is the expressed as

  (a) $\mathrm{H}_0 \colon \alpha_1 = \cdots = \alpha_G$.

  (b) $\mathrm{H}_0 \colon \beta_1 = 0 \;\&\; \ldots \;\&\; \beta_{G-1} = 0$, i.e., $\mathrm{H}_0 \colon \boldsymbol{\beta}^Z = \mathbf{0}_{G-1}$.

- If a normal linear model is assumed, test on a value of the estimable vector parameter or a submodel test which compares the one-way ANOVA model with the intercept-only model can be used to test the above null hypotheses. The corresponding F-test is indeed a well known one-way ANOVA F-test.

9.1.3 Least squares estimation

In case of a one-way ANOVA linear model, explicit formulas for the LSE related quantities can easily be derived.

Theorem 9.1 Least squares estimation in one-way ANOVA linear model.

The fitted values and the LSE of the group means in a one-way ANOVA linear model are equal to the group sample means:

$$\widehat{m}_g = \widehat{Y}_{g,j} = \frac{1}{n_g} \sum_{l=1}^{n_g} Y_{g,l} =: \overline{Y}_{g\bullet}, \qquad g = 1, \ldots, G,\ j = 1, \ldots, n_g.$$

That is,

$$\widehat{\boldsymbol{m}} := \begin{pmatrix} \widehat{m}_1 \\ \vdots \\ \widehat{m}_G \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{G\bullet} \end{pmatrix}, \qquad \widehat{\boldsymbol{Y}} = \begin{pmatrix} \overline{Y}_{1\bullet} \mathbf{1}_{n_1} \\ \vdots \\ \overline{Y}_{G\bullet} \mathbf{1}_{n_G} \end{pmatrix}.$$

If additionally normality is assumed, i.e., $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\big(\boldsymbol{\mu},\, \sigma^2\, \mathbf{I}_n\big)$*, where* $\boldsymbol{\mu} = \big(m_1\, \mathbf{1}_{n_1}^\top,\, \ldots,\, m_G\, \mathbf{1}_{n_G}^\top\big)^\top$*, then* $\widehat{\boldsymbol{m}} \,|\, \mathbb{Z} \sim \mathcal{N}_G\big(\boldsymbol{m},\, \sigma^2\, \mathbb{V}\big)$*, where*

$$\mathbb{V} = \begin{pmatrix} \frac{1}{n_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{n_G} \end{pmatrix}.$$

Proof. Use a full-rank parameterization $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$ with

$$\mathbb{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \ldots & \mathbf{0}_{n_1} \\ \vdots & \vdots & \vdots \\ \mathbf{0}_{n_G} & \vdots & \mathbf{1}_{n_G} \end{pmatrix}, \qquad \boldsymbol{\beta} = \big(m_1, \ldots, m_G\big)^\top.$$

We have

$$\mathbb{X}^\top\mathbb{X} = \begin{pmatrix} n_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & n_G \end{pmatrix}, \quad \mathbb{X}^\top\boldsymbol{Y} = \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1,j} \\ \vdots \\ \sum_{j=1}^{n_G} Y_{G,j} \end{pmatrix}, \quad \left(\mathbb{X}^\top\mathbb{X}\right)^{-1} = \begin{pmatrix} \frac{1}{n_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{n_G} \end{pmatrix},$$

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{m}} = \left(\widehat{m}_1, \dots, \widehat{m}_G\right)^\top = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y} = \left(\overline{Y}_{1\bullet}, \dots, \overline{Y}_{G\bullet}\right)^\top.$$

Finally,

$$\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{m}_1\mathbf{1}_{n_1} \\ \vdots \\ \widehat{m}_G\mathbf{1}_{n_G} \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1\bullet}\mathbf{1}_{n_1} \\ \vdots \\ \overline{Y}_{G\bullet}\mathbf{1}_{n_G} \end{pmatrix}.$$

Normality and the form of the covariance matrix of $\widehat{\boldsymbol{m}}$ follows from a general LSE theory.

❑

LSE of regression coefficients and estimable parameters

With a full-rank parameterization, a vector $\boldsymbol{m}$ is linked to the regression coefficients $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}^Z\right)^\top$, $\boldsymbol{\beta}^Z = \left(\beta_1, \dots, \beta_{G-1}\right)^\top$, by the relationship

$$\boldsymbol{m} = \beta_0\mathbf{1}_G + \mathbb{C}\boldsymbol{\beta}^Z.$$

Due to the fact that $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}}$, where $\mathbb{X}$ is a model matrix derived from the (pseudo)contrast matrix $\mathbb{C}$, the LSE $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^Z\right)^\top$ of the regression coefficients in a full-rank parameterization satisfy

$$\widehat{\boldsymbol{m}} = \widehat{\beta}_0\mathbf{1}_G + \mathbb{C}\widehat{\boldsymbol{\beta}}^Z,$$

which is a regular linear system with the solution

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}^Z \end{pmatrix} = \left(\mathbf{1}_G, \mathbb{C}\right)^{-1} \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{G\bullet} \end{pmatrix}.$$

That is, the LSE of the regression coefficients is always a linear combination of the group sample means. The same then holds for any estimable parameter. For example, the LSE of the differences between the group means $\theta_{g,h} = m_g - m_h$, $g, h = 1, \dots, G$, are

$$\widehat{\theta}_{g,h} = \overline{Y}_{g\bullet} - \overline{Y}_{h\bullet}, \qquad g, h = 1, \dots, G.$$

Analogously, the LSE of the factor effects $\eta_g = m_g - \frac{1}{G}\sum_{h=1}^{G} m_h$, $g = 1, \dots, G$, are

$$\widehat{\eta}_g = \overline{Y}_{g\bullet} - \frac{1}{G}\sum_{h=1}^{G}\overline{Y}_{h\bullet}, \qquad g = 1, \dots, G.$$

9.1.4 Within and between groups sums of squares, ANOVA F-test

Sums of squares

Let as usual, $\overline{Y}$ denote a sample mean based on the response vector $\boldsymbol{Y}$, i.e.,

$$\overline{Y} = \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} Y_{g,j} = \frac{1}{n}\sum_{g=1}^{G} n_g \overline{Y}_{g\bullet}.$$

In a one-way ANOVA linear model, the residual and the regression sums of squares and corresponding degrees of freedom are

$$\mathsf{SS}_e = \left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\left(Y_{g,j} - \widehat{Y}_{g,j}\right)^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\left(Y_{g,j} - \overline{Y}_{g\bullet}\right)^2,$$
$$\nu_e = n - G,$$

$$\mathsf{SS}_R = \left\|\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}_n\right\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\left(\widehat{Y}_{g,j} - \overline{Y}\right)^2 = \sum_{g=1}^{G} n_g\left(\overline{Y}_{g\bullet} - \overline{Y}\right)^2,$$
$$\nu_R = G - 1.$$

In this context, the residual sum of squares SS_e is also called *the within groups* sum of squares[2], the regression sum of squares SS_R is called *the between groups* sum of squares[3].

One-way ANOVA F-test

Let us assume normality of the response and consider a submodel $\boldsymbol{Y} \mid \mathbb{Z} \sim \mathcal{N}_n\left(\mathbf{1}_n\beta_0,\, \sigma^2\mathbf{I}_n\right)$ of the one-way ANOVA model. A residual sum of squares of the submodel is

$$\mathsf{SS}_e^0 = \mathsf{SS}_T = \left\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\right\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\left(Y_{g,j} - \overline{Y}\right)^2.$$

Breakdown of the total sum of squares (Theorem 5.8) gives $\mathsf{SS}_R = \mathsf{SS}_T - \mathsf{SS}_e = \mathsf{SS}_e^0 - \mathsf{SS}_e$ and hence the statistic of the F-test on a submodel is

$$F = \frac{\frac{\mathsf{SS}_R}{G-1}}{\frac{\mathsf{SS}_e}{n-G}} = \frac{\mathsf{MS}_R}{\mathsf{MS}_e}, \tag{9.3}$$

where

$$\mathsf{MS}_R = \frac{\mathsf{SS}_R}{G-1}, \qquad \mathsf{MS}_e = \frac{\mathsf{SS}_e}{n-G}.$$

The F-statistic (9.3) is indeed a classical one-way ANOVA F-statistics which under the null hypothesis of validity of a submodel, i.e., under the null hypothesis of equality of the group means, follows an $\mathcal{F}_{G-1,\,n-G}$ distribution. Above quantities, together with the P-value derived from the $\mathcal{F}_{G-1,\,n-G}$ distribution are often recorded in a form of the ANOVA table:

[2] *vnitroskupinový součet čtverců* [3] *meziskupinový součet čtverců*

Effect (Term)	Degrees of freedom	Effect sum of squares	Effect mean square	F-stat.	P-value
Factor	$G-1$	SS_R	MS_R	F	p
Residual	$n-G$	SS_e	MS_e		

Consider a terminology introduced in Section 8.8, and denote as Z main effect terms that correspond to the covariate Z. We have $\mathsf{SS}_R = \mathsf{SS}\big(\mathsf{Z} \,\big|\, 1\big)$ and the above ANOVA table is now type I as well as type II ANOVA table. If intercept is explicitely included in the model matrix then it is also the type III ANOVA table.

9.2 Two-way classification

Suppose now that there are two categorical covariates Z and W available with

$$Z \in \mathcal{Z} = \{1, \ldots, G\}, \qquad W \in \mathcal{W} = \{1, \ldots, H\}.$$

This can be viewed as if the two covariates correspond to division of population of interest into $G \cdot H$ subpopulations/groups. Each group is then identified by a combination of values of two factors (Z and W) and hence the situation is commonly referred to as *two-way classification*[4]. A linear model can now be used to parameterize a set of $G \cdot H$ (conditional) response expectations $\mathbb{E}\big(Y \,\big|\, Z = g, W = h\big)$, $g = 1, \ldots, G$, $h = 1, \ldots, H$ (group specific response expectations). Those will be called, in this context, as *two-way classified group means*:

$$m(g, h) = \mathbb{E}\big(Y \,\big|\, Z = g, W = h\big) =: m_{g,h}, \qquad g = 1, \ldots, G, \; h = 1, \ldots, H.$$

Suppose that a combination $\big(Z, W\big)^\top = \big(g, h\big)^\top$ is repeated $n_{g,h}$-times in the data, $g = 1, \ldots, G$, $h = 1, \ldots, H$. That is,

$$n = \sum_{g=1}^{G} \sum_{h=1}^{H} n_{g,h}.$$

Analogously to Section 9.1, it will overally be assumed that $n_{g,h} > 0$ (almost surely) for each g and h. That is, it is assumed that each group identified by $\big(Z, W\big)^\top = \big(g, h\big)^\top$ is (almost surely) represented in the data.

For the clarity of notation, we will now use also a triple subscript to index the individual observations. The first subscript will indicate a value of the covariate Z, the second subscript will indicate a value of the covariate W and the third subscript will consecutively number the observations with the same $\big(Z, W\big)^\top$ combination. Finally, without loss of generality, we will assume that data are sorted primarily with respect to the value of the covariate W and secondarily with respect to the value of the covariate Z. That is, the covariate matrix and the response vector take a form as shown in Table 9.1.

As usually, let $\mathbb{Z} = \big(Z_{1,1,1}, \ldots, Z_{G,H,n_{G,H}}\big)^\top$ denote the $n \times 1$ matrix with all values of the Z covariate in the data and similarly, let $\mathbb{W} = \big(W_{1,1,1}, \ldots, W_{G,H,n_{G,H}}\big)^\top$ denote the $n \times 1$ matrix with all values of the W covariate.

Still in the same spirit of Section 9.1, we merge response random variables with a common value of the two covariates into response subvectors $\boldsymbol{Y}_{g,h} = \big(Y_{g,h,1}, \ldots, Y_{g,h,n_{g,h}}\big)^\top$, $g = 1, \ldots, G$, $h = 1, \ldots, H$. The overall response vector $\boldsymbol{Y}$ is then

$$\boldsymbol{Y} = \big(\boldsymbol{Y}_{1,1}^\top, \ldots, \boldsymbol{Y}_{G,1}^\top, \;\; \ldots, \;\; \boldsymbol{Y}_{1,H}^\top, \ldots, \boldsymbol{Y}_{G,H}^\top\big)^\top.$$

Similarly, a vector $\boldsymbol{m}$ will now be a vector of the two-way classified group means. That is,

$$\boldsymbol{m} = \big(m_{1,1}, \ldots, \; m_{G,1}, \ldots \ldots, m_{1,H}, \ldots, m_{G,H}\big)^\top.$$

Further, let

$$n_{g\bullet} = \sum_{h=1}^{H} n_{g,h}, \qquad g = 1, \ldots, G$$

denote the number of datapoints with $Z = g$ and similarly, let

$$n_{\bullet h} = \sum_{g=1}^{G} n_{g,h}, \qquad h = 1, \ldots, H$$

[4] *dvojné třídění*

Table 9.1: Two-way classification: Covariate matrix and overall response vector.

$$\begin{pmatrix} Z_1 & W_1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ Z_n & W_n \end{pmatrix} = \left(\begin{array}{cc} Z_{1,1,1} & W_{1,1,1} \\ \vdots & \vdots \\ Z_{1,1,n_{1,1}} & W_{1,1,n_{1,1}} \\ \hline \vdots & \vdots \\ \hline Z_{G,1,1} & W_{G,1,1} \\ \vdots & \vdots \\ Z_{G,1,n_{G,1}} & W_{G,1,n_{G,1}} \\ \hline \vdots & \vdots \\ \vdots & \vdots \\ \hline Z_{1,H,1} & W_{1,H,1} \\ \vdots & \vdots \\ Z_{1,H,n_{1,H}} & W_{1,H,n_{1,H}} \\ \hline \vdots & \vdots \\ \hline Z_{G,H,1} & W_{G,H,1} \\ \vdots & \vdots \\ Z_{G,H,n_{G,H}} & W_{G,H,n_{G,H}} \end{array}\right) = \left(\begin{array}{cc} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ \hline \vdots & \vdots \\ \hline G & 1 \\ \vdots & \vdots \\ G & 1 \\ \hline \vdots & \vdots \\ \vdots & \vdots \\ \hline 1 & H \\ \vdots & \vdots \\ 1 & H \\ \hline \vdots & \vdots \\ \hline G & H \\ \vdots & \vdots \\ G & H \end{array}\right), \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{pmatrix} = \left(\begin{array}{c} Y_{1,1,1} \\ \vdots \\ Y_{1,1,n_{1,1}} \\ \hline \vdots \\ \hline Y_{G,1,1} \\ \vdots \\ Y_{G,1,n_{G,1}} \\ \hline \vdots \\ \vdots \\ \hline Y_{1,H,1} \\ \vdots \\ Y_{1,H,n_{1,H}} \\ \hline \vdots \\ \hline Y_{G,H,1} \\ \vdots \\ Y_{G,H,n_{G,H}} \end{array}\right).$$

denote the number of datapoints with $W = h$. Finally, we will denote various means of the group means as follows.

$$\begin{aligned} \overline{m} &:= \frac{1}{G \cdot H} \sum_{g=1}^{G} \sum_{h=1}^{H} m_{g,h}, \\ \overline{m}_{g\bullet} &:= \frac{1}{H} \sum_{h=1}^{H} m_{g,h}, \qquad g = 1, \ldots, G, \\ \overline{m}_{\bullet h} &:= \frac{1}{G} \sum_{g=1}^{G} m_{g,h}, \qquad h = 1, \ldots, H. \end{aligned}$$

For following considerations, it is useful to view data as if each subpopulation/group corresponds to a cell in an $G \times H$ table whose rows are indexed by the values of the Z and W covariates as shown in Table 9.2.

Notes.

- The above defined quantities $\overline{m}_{g\bullet}$, $\overline{m}_{\bullet h}$, $\overline{m}$ are the means of the group means which are *not* weighted by the corresponding sample sizes (which are moreover random if the covariates are random). As such, all above defined means are always real constants and never random variables (irrespective of whether the covariates are considered as being fixed or random).

Table 9.2: Two-way classification: Response variables, group means, sample sizes in a tabular display.

Response variables

	W		
Z	1	$\dots$	H
1	$\boldsymbol{Y}_{1,1} = \big(Y_{1,1,1}, \dots, Y_{1,1,n_{1,1}}\big)^\top$	$\vdots$	$\boldsymbol{Y}_{1,H} = \big(Y_{1,H,1}, \dots, Y_{1,H,n_{1,H}}\big)^\top$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$\boldsymbol{Y}_{G,1} = \big(Y_{G,1,1}, \dots, Y_{G,1,n_{G,1}}\big)^\top$	$\vdots$	$\boldsymbol{Y}_{G,H} = \big(Y_{G,H,1}, \dots, Y_{G,H,n_{G,H}}\big)^\top$

Group means

	W			
Z	1	$\dots$	H	$\bullet$
1	$m_{1,1}$	$\vdots$	$m_{1,H}$	$\overline{m}_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$m_{G,1}$	$\vdots$	$m_{G,H}$	$\overline{m}_{G\bullet}$
$\bullet$	$\overline{m}_{\bullet 1}$	$\dots$	$\overline{m}_{\bullet H}$	$\overline{m}$

Sample sizes

	W			
Z	1	$\dots$	H	$\bullet$
1	$n_{1,1}$	$\vdots$	$n_{1,H}$	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$n_{G,1}$	$\vdots$	$n_{G,H}$	$n_{G\bullet}$
$\bullet$	$n_{\bullet 1}$	$\dots$	$n_{\bullet H}$	n

- When interpreting the means of the group means, it must be taken into account that in general, it is not necessarily true that $\overline{m}_{g\bullet} = \mathbb{E}\big(Y \,\big|\, Z = g\big)$ $(g = 1, \dots, G)$, $\overline{m}_{\bullet h} = \mathbb{E}\big(Y \,\big|\, W = h\big)$ $(h = 1, \dots, H)$, or $\overline{m} = \mathbb{E}\big(Y\big)$.
- Data in the overall response vector $\boldsymbol{Y}$ are sorted as if we put columns of the response matrix from Table 9.2 one after each other.

The full response vector is $\boldsymbol{Y}$ and its (conditional, given $\mathbb{Z}$ and $\mathbb{W}$) mean are

$$\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}, \mathbb{W}\big) = \begin{pmatrix} m_{1,1}\, \mathbf{1}_{n_{1,1}} \\ \vdots \\ m_{G,H}\, \mathbf{1}_{n_{G,H}} \end{pmatrix} =: \boldsymbol{\mu}. \tag{9.4}$$

A standard linear model then additionally assumes

$$\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}, \mathbb{W}\big) = \sigma^2\, \mathbf{I}_n. \tag{9.5}$$

With the i.i.d. data $\big(Y_i, Z_i, W_i\big)^\top \overset{\text{i.i.d.}}{\sim} \big(Y, Z, W\big)^\top$ for which (9.4) and (9.5) are assumed, the random variables $Y_{g,h,1}, \dots, Y_{g,h,n_{g,h}}$ (elements of the vector $\boldsymbol{Y}_{g,h}$) are i.i.d. from a distribution of $Y \,|\, Z = g, W = h$ whose mean is $m_{g,h}$ and the variance is σ^2. That is, the response random

variables form $G \cdot H$ independent i.i.d. samples:

$$\text{Sample } (1,\,1): \quad \boldsymbol{Y}_{1,1} = \big(Y_{1,1,1},\,\ldots,\,Y_{1,1,n_{1,1}}\big)^\top,$$
$$Y_{1,1,j} \overset{\text{i.i.d.}}{\sim} (m_{1,1},\,\sigma^2), \qquad j = 1,\ldots,n_{1,1},$$
$$\vdots \qquad\qquad \vdots$$
$$\text{Sample } (G,\,H): \quad \boldsymbol{Y}_{G,H} = \big(Y_{G,H,1},\,\ldots,\,Y_{G,H,n_{G,H}}\big)^\top,$$
$$Y_{G,H,j} \overset{\text{i.i.d.}}{\sim} (m_{G,H},\,\sigma^2), \qquad j = 1,\ldots,n_{G,H}.$$

As in Section 9.1, we arrive at the same situation even if the covariate values are fixed rather than random. A conceptual difference between the situation of random and fixed covariates in this setting is that with random covariates, the group sample sizes $n_{1,1},\,\ldots,\,n_{G,H}$ are random as well, whereas with the fixed covariates, also those sample sizes are fixed. Analogously to Section 9.1, we keep assuming that $n_{1,1} > 0,\,\ldots,\,n_{G,H} > 0$ (almost surely in case of random covariates). A linear model with the inference being conditioned by the covariate values can now be used to infere on the group means $m_{1,1},\,\ldots,\,m_{G,H}$ or on their linear combinations.

9.2.1 Parameters of interest

Various quantities, all being linear combinations of the two-way classified group means, i.e., all being estimable in any parameterization of the two-way classification, are clasically of interest, especially in the area of *designed experiments* used often in industrial statistics. Here, the levels of the two covariates Z and W correspond to certain experimental (machine) settings of two factors that may influence the output Y of interest (e.g., production of the machine). The group mean $m_{g,h}$ is then the mean outcome if the Z factor is set to level g and the W factor to level h. Next to the group means themselves, additional quantities of interest clasically include

(i) The mean of the group means $\overline{m}$.

- For designed experiment, this is the mean outcome value if we perform the experiment with all combinations of the input factors Z and W (each combination equally replicated).
- If Y represents some industrial production then $\overline{m}$ provides the mean production as if all combinations of inputs are equally often used in the production process.

(ii) The means of the means by the first or the second factor, i.e., parameters

$$\overline{m}_{1\bullet},\,\ldots,\,\overline{m}_{G\bullet}, \qquad \text{and} \qquad \overline{m}_{\bullet 1},\,\ldots,\,\overline{m}_{\bullet H}.$$

- For designed experiment, the value of $\overline{m}_{g\bullet}$ $(g = 1,\ldots,G)$ is the mean outcome value if we fix the factor Z on its level g and perform the experiment while setting the factor W to all possible levels (again, each equally replicated).
- If Y represents some industrial production then $\overline{m}_{g\bullet}$ provides the mean production as if the Z input is set to g but all possible values of the second input W are equally often used in the production process.
- Interpretation of $\overline{m}_{\bullet 1}$ $(h = 1,\ldots,H)$ just mirrors interpretation of $\overline{m}_{g\bullet}$.

(iii) Differences between the means of the means by the first or the second factor, i.e., parameters

$$\theta_{g_1,g_2\bullet} := \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet}, \qquad g_1,\,g_2 = 1,\ldots,G,\;\; g_1 \neq g_2,$$
$$\theta_{\bullet h_1,h_2} := \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2}, \qquad h_1,\,h_2 = 1,\ldots,H,\;\; h_1 \neq h_2.$$

Those, in a certain sense quantify the mean effect of the first or the second factor on the response.

- For designed experiment, the value of $\theta_{g_1,g_2\bullet}$ ($g_1 \neq g_2$) is the mean difference between the outcome values if we fix the factor Z to its levels g_1 and g_2, repectively and perform the experiment while setting the factor W to all possible levels (again, each equally replicated).
- If Y represents some industrial production then $\theta_{g_1,g_2\bullet}$ ($g_1 \neq g_2$) provides difference between the mean productions with Z set to g_1 and g_2, respectively while using all possible values of the second input W equally often in the production process.
- Interpretation of $\theta_{\bullet h_1,h_2}$ ($h_1 \neq h_2$) just mirrors interpretation of θ_{g_1,g_2}.

9.2.2 Additivity and interactions

By now, there is practically no difference compared to the one-way classification examined in Section 9.1 except the fact that the group means are indexed by a combination of the two covariate values.

Next to estimation of specific parameters of interest, specific questions related to the structure of the vector $\boldsymbol{m}$ of the two-way classified group means are being examined. To proceed, note that we can write the group means as follows

$$m_{g,h} = \overline{m} + (\overline{m}_{g\bullet} - \overline{m}) + (\overline{m}_{\bullet h} - \overline{m}) + (m_{g,h} - \overline{m}_{g\bullet} - \overline{m}_{\bullet h} + \overline{m}),$$
$$g = 1, \ldots, G,\ h = 1, \ldots, H,$$

which motivates the following definition.

Main and interaction effects

Definition 9.2 Main and interaction effects in two-way classification.

Consider a two-way classification based on factors Z and W. By main effects *of the factor Z, we understand quantities $\eta_1^Z, \ldots, \eta_G^Z$ defined as*

$$\eta_g^Z := \overline{m}_{g\bullet} - \overline{m}, \qquad g = 1, \ldots, G.$$

By main effects *of the factor W, we understand quantities $\eta_1^W, \ldots, \eta_H^W$ defined as*

$$\eta_h^W := \overline{m}_{\bullet h} - \overline{m}, \qquad h = 1, \ldots, H.$$

By interaction effects*, we understand quantities $\eta_{1,1}^{ZW}, \ldots, \eta_{G,H}^{ZW}$ defined as*

$$\eta_{g,h}^{ZW} := m_{g,h} - \overline{m}_{g\bullet} - \overline{m}_{\bullet h} + \overline{m}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H.$$

That is, the two-way classified group means are given as

$$m_{g,h} = \overline{m} + \eta_g^Z + \eta_h^W + \eta_{g,h}^{ZW}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H. \tag{9.6}$$

Having defined the main effects, we can note that their differences provide also differences between the means of the means by the corresponding factor. That is,

$$\begin{aligned} \theta_{g_1,g_2\bullet} &= \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet} = \eta_{g_1}^Z - \eta_{g_2}^Z, && g_1, g_2 = 1, \ldots, G,\ g_1 \neq g_2, \\ \theta_{\bullet h_1,h_2} &= \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2} = \eta_{h_1}^W - \eta_{h_2}^W, && h_1, h_2 = 1, \ldots, H,\ h_1 \neq h_2. \end{aligned} \tag{9.7}$$

Additivity

Suppose now that the factors Z and W act additively on the response expectation. That is, effect of change of one covariate (let say Z) does not depend on a value of the other covariate (let say W). That is, for any $g_1,\ g_2 = 1, \ldots, G$

$$\mathbb{E}\big(Y \,\big|\, Z = g_1,\, W = h\big) \;-\; \mathbb{E}\big(Y \,\big|\, Z = g_2,\, W = h\big) \;=\; m_{g_1,h} \;-\; m_{g_2,h}$$

does not depend on a value of $h = 1, \ldots, H$. Consequently, for any $g_1,\ g_2$ and any h

$$m_{g_1,h} \;-\; m_{g_2,h} \;=\; \overline{m}_{g_1\bullet} \;-\; \overline{m}_{g_2\bullet}. \tag{9.8}$$

This implies

$$m_{g_1,h} \;-\; \overline{m}_{g_1\bullet} \;=\; m_{g_2,h} \;-\; \overline{m}_{g_2\bullet}, \qquad g_1,\, g_2 = 1, \ldots, G,\ h = 1, \ldots, H.$$

In other words, additivity implies that for any $h = 1, \ldots, H$, the differences

$$\Delta(g, h) \;=\; m_{g,h} \;-\; \overline{m}_{g\bullet}$$

do not depend on a value of $g = 1, \ldots, G$. Then (for any $g = 1, \ldots, G$ and $h = 1, \ldots, H$)

$$\begin{aligned} m_{g,h} \;-\; \overline{m}_{g\bullet} \;&=\; \Delta(g,\, h) \\ &=\; \frac{1}{G}\sum_{g^\star=1}^{G} \Delta(g^\star, h) \\ &=\; \frac{1}{G}\sum_{g^\star=1}^{G} (m_{g^\star,h} \;-\; \overline{m}_{g^\star\bullet}) \\ &=\; \overline{m}_{\bullet h} \;-\; \overline{m}. \end{aligned}$$

Clearly, we would arrive at the same conclusion if we started oppositely from assuming that for any $h_1,\ h_2 = 1, \ldots, H$

$$\mathbb{E}\big(Y \,\big|\, Z = g,\, W = h_1\big) \;-\; \mathbb{E}\big(Y \,\big|\, Z = g,\, W = h_2\big) \;=\; m_{g,h_1} \;-\; m_{g,h_2}$$

does not depend on a value of $g = 1, \ldots, G$.

In summary, additivity implies

$$\underbrace{m_{g,h} \;-\; \overline{m}_{g\bullet} \;-\; \overline{m}_{\bullet h} \;+\; \overline{m}}_{\eta^{ZW}_{g,h}} \;=\; 0, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H.$$

Easily, we see that this is also a sufficient condition for additivity. That is, hypothesis of additivity of the effect of the two covariates on the response expectation is given as

$$\mathrm{H}_0\colon\ \eta^{ZW}_{g,h} = 0, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H.$$

Main effects under additivity, partial effects

Under additivity, the two-way classified group means can be written as

$$m_{g,h} \;=\; \overline{m} \;+\; \eta^{Z}_{g} \;+\; \eta^{W}_{h}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H.$$

In that case, in agreement with (9.8), we have

$$\begin{aligned}
\theta_{g_1,g_2\bullet} &= \eta^Z_{g_1} - \eta^Z_{g_2} = m_{g_1,h} - m_{g_2,h}, && g_1, g_2 = 1, \ldots, G,\ g_1 \neq g_2,\\
& && h = 1, \ldots, H,\\
\theta_{\bullet h_1,h_2} &= \eta^W_{h_1} - \eta^W_{h_2} = m_{g,h_1} - m_{g,h_2}, && h_1, h_2 = 1, \ldots, H,\ h_1 \neq h_2,\\
& && g = 1, \ldots, G.
\end{aligned} \tag{9.9}$$

That is, the differences between the main effects not only provide differences between the means of the means by corresponding factor (as indicated by Eq. 9.7) but also differences between the two group means if we change the value of one factor and keep the value of the second factor which can be arbitrary.

In Section 8.3.1 we introduced a notion of a partial effects of a certain categorical covariate (let say $Z \in \mathcal{Z} = \{1, \ldots, G\}$) on the response expectation if it with other covariates (or a covariate, let say W) acts additivively. The partial effects of the covariate Z, given the other covariate W were introduced as the model parameters that determine quantities $\mathbb{E}(Y \,|\, Z = g_1, W = w) - \mathbb{E}(Y \,|\, Z = g_2, W = w)$, $g_1, g_2 = 1, \ldots, G$. If the other covariate, W, is categorical as well (with $W \in \mathcal{W} = \{1, \ldots, H\}$) then the partial effects of the Z covariate are related to the quantities

$$\begin{aligned}
\mathbb{E}(Y \,|\, Z = g_1, W = h) - \mathbb{E}(Y \,|\, Z = g_2, W = h) &= m_{g_1,h} - m_{g_2,h},\\
& g_1, g_2 = 1, \ldots, G,\ h = 1, \ldots, H,
\end{aligned}$$

which are then given by (9.9), i.e., as differences between related main effects.

9.2.3 Linear model parameterization of two-way classified group means

Estimation of all parameters of interest may proceed by considering suitably parameterized linear model. When doing so, remember that we assume data being sorted primarily by the value of the covariate W and secondarily by the value of the Z covariate as indicated in Table 9.1. The full response expectation is then given by (9.4) which is, if written in more detail

$$\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z}, \mathbb{W}) = \boldsymbol{\mu} = \begin{pmatrix}
m_{1,1}\,\mathbf{1}_{n_{1,1}}\\
\vdots\\
m_{G,1}\,\mathbf{1}_{n_{G,1}}\\
- - - - -\\
\vdots\\
\vdots\\
- - - - -\\
m_{1,H}\,\mathbf{1}_{n_{1,H}}\\
\vdots\\
m_{G,H}\,\mathbf{1}_{n_{G,H}}
\end{pmatrix}.$$

It is our aim to parameterize the vector $\boldsymbol{\mu}$ as $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$, where $\mathbb{X}$ is the $n \times k$ model matrix and $\boldsymbol{\beta} \in \mathbb{R}^k$ a vector of regression coefficients. The situation is basically the same as in case of a single

categorical covariate in Section 7.4 if we view each of the $G \cdot H$ combinations of the Z and W covariates as one of the values of a new categorical covariate with $G \cdot H$ levels labeled by double indeces $(1,1), \ldots, (G,H)$. The following facts then directly follow from Section 7.4 (given our assumption that $n_{g,h} > 0$ for all $(g,\, h)$).

- Matrix $\mathbb{X}$ must have a rank of $G \cdot H$, i.e., at least $k = G \cdot H$ columns and its choice simplifies into selecting an $(G \cdot H) \times k$ matrix $\widetilde{\mathbb{X}}$ such that

$$\widetilde{\mathbb{X}} = \begin{pmatrix} \boldsymbol{x}_{1,1}^\top \\ \vdots \\ \boldsymbol{x}_{G,1}^\top \\ - - - \\ \vdots \\ \vdots \\ - - - \\ \boldsymbol{x}_{1,H}^\top \\ \vdots \\ \boldsymbol{x}_{G,H}^\top \end{pmatrix}, \quad \text{leading to } \mathbb{X} = \begin{pmatrix} \mathbf{1}_{n_{1,1}} \otimes \boldsymbol{x}_{1,1}^\top \\ \vdots \\ \mathbf{1}_{n_{G,1}} \otimes \boldsymbol{x}_{G,1}^\top \\ - - - - - - \\ \vdots \\ \vdots \\ - - - - - - \\ \mathbf{1}_{n_{1,H}} \otimes \boldsymbol{x}_{1,H}^\top \\ \vdots \\ \mathbf{1}_{n_{G,H}} \otimes \boldsymbol{x}_{G,H}^\top \end{pmatrix}.$$

- $\mathsf{rank}\big(\mathbb{X}\big) = \mathsf{rank}\big(\widetilde{\mathbb{X}}\big)$.
- Matrix $\widetilde{\mathbb{X}}$ parameterizes the two-way classified group means as

$$\begin{aligned} m_{g,h} &= \boldsymbol{x}_{g,h}^\top \boldsymbol{\beta}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H, \\ \boldsymbol{m} &= \widetilde{\mathbb{X}} \boldsymbol{\beta}. \end{aligned}$$

If purely parameterization of a vector $\boldsymbol{m}$ of the two-way classified group means is of interest, matrix $\widetilde{\mathbb{X}}$ can be chosen using the methods discussed in Section 7.4 applied to a combined categorical covariate with $G \cdot H$ levels. Nevertheless, it is usually of interest to use such parameterization where

(i) (at least some of) the regression coefficients have meaning of primary parameters of interest (also other parameters of interest than those proposed in Section 9.2.1 can be considered);

(ii) hypothesis of additivity corresponds to setting a subvector of the regression coefficients vector to the zero vector. In that case, a model expressing the additivity is a submodel of the full two-way classification model obtained by omitting some columns fromm the model matrix.

Both requirements will be fulfilled, as we shall show, if we parameterize the two-way classified group means by the interaction model based on the covariates Z and W using common guidelines introduced in Section 8.4.

9.2.4 ANOVA parameterization of two-way classified group means

End of Lecture #16 (23/11/2016)
Start of Lecture #17 (24/11/2016)

As shown by (9.6), the two-way classified group means can be written using the main and interaction effects as $m_{g,h} = \overline{m} + \eta_g^Z + \eta_h^W + \eta_{g,h}^{ZW}$, $g = 1, \ldots, G,\ h = 1, \ldots, H$. This motivates so called ANOVA parameterization of the two-way classified group means being given as

$$m_{g,h} = \alpha_0 + \alpha_g^Z + \alpha_h^W + \alpha_{g,h}^{ZW}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H, \tag{9.10}$$

where a vector of regression coefficients $\boldsymbol{\alpha} = \big(\alpha_0,\, \boldsymbol{\alpha}^{Z\top},\, \boldsymbol{\alpha}^{W\top},\, \boldsymbol{\alpha}^{ZW\top}\big)^\top$ is composed of

- the intercept term α_0;
- coefficients $\boldsymbol{\alpha}^Z = \left(\alpha_1^Z, \ldots, \alpha_G^Z\right)^\top$ that are, in a certain sense, related to the main effects of the Z covariate;
- coefficients $\boldsymbol{\alpha}^W = \left(\alpha_1^W, \ldots, \alpha_H^W\right)^\top$ that are, in a certain sense, related to the main effects of the W covariate;
- coefficients $\boldsymbol{\alpha}^{ZW} = \left(\alpha_{1,1}^{ZW}, \ldots, \alpha_{G,1}^{ZW}, \ldots\ldots, \alpha_{1,H}^{ZW}, \ldots, \alpha_{G,H}^{ZW}\right)^\top$ that are, in a certain sense, related to the interaction effects.

Notes.

- The intercept term α_0 is not necessarily equal to $\overline{m}$;
- The coefficients $\alpha_1^Z, \ldots, \alpha_G^Z$ are not necessarily equal to the main effects $\eta_1^Z, \ldots, \eta_G^Z$;
- The coefficients $\alpha_1^W, \ldots, \alpha_H^W$ are not necessarily equal to the main effects $\eta_1^W, \ldots, \eta_H^W$;
- The coefficients $\alpha_{1,1}^{ZW}, \ldots, \alpha_{G,H}^{ZW}$ are not necessarily equal to the interaction effects $\eta_{1,1}^{ZW}, \ldots, \eta_{G,H}^{ZW}$;

as the parameterization (9.10) does not lead to the full-rank linear model as will be immediately shown.

Let $\boldsymbol{m}_{\bullet h} = \left(m_{1,h}, \ldots, m_{G,h}\right)^\top$, $h = 1, \ldots, H$ be subvectors of $\boldsymbol{m}$. In a matrix form, parameterization (9.10) is

$$\boldsymbol{m} = \begin{pmatrix} \boldsymbol{m}_{\bullet 1} \\ \hline \vdots \\ \hline \boldsymbol{m}_{\bullet H} \end{pmatrix} = \underbrace{\left(\begin{array}{c|c|ccccc|ccccc} \mathbf{1}_G & \mathbf{I}_G & \mathbf{1}_G & \ldots & \ldots & \mathbf{0}_G & \mathbf{I}_G & \ldots & \ldots & \mathbf{0}_{G\times G} \\ \hline \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots \\ \hline \mathbf{1}_G & \mathbf{I}_G & \mathbf{0}_G & \ldots & \ldots & \mathbf{1}_G & \mathbf{0}_{G\times G} & \ldots & \ldots & \mathbf{I}_G \end{array}\right)}_{\widetilde{\mathbb{X}}_\alpha} \begin{pmatrix} \alpha_0 \\ \boldsymbol{\alpha}^Z \\ \boldsymbol{\alpha}^W \\ \boldsymbol{\alpha}^{ZW} \end{pmatrix},$$

where matrix $\widetilde{\mathbb{X}}_\alpha$ is an $(G\cdot H)\times(1+G+H+G\cdot H)$ matrix and its rank is hence at most $G\cdot H$ (it is indeed precisely equal to $G\cdot H$). That is, matrix $\widetilde{\mathbb{X}}_\alpha$ provides less-than-full rank parameterization of the two-way classified group means. Note that matrix $\widetilde{\mathbb{X}}_\alpha$ can concisely be written as

$$\widetilde{\mathbb{X}}_\alpha = \left(\begin{array}{c|c|c|c} \mathbf{1}_H \otimes \mathbf{1}_G & \mathbf{1}_H \otimes \mathbf{I}_G & \mathbf{I}_H \otimes \mathbf{1}_G & \mathbf{I}_H \otimes \mathbf{I}_G \end{array}\right) \tag{9.11}$$

$$= \left(\begin{array}{c|c|c|c} \underbrace{\mathbf{1}_H \otimes \mathbf{1}_G}_{\mathbf{1}_{G\cdot H}} & \mathbf{1}_H \otimes \widetilde{\mathbb{C}} & \widetilde{\mathbb{D}} \otimes \mathbf{1}_G & \widetilde{\mathbb{D}} \otimes \widetilde{\mathbb{C}} \end{array}\right), \tag{9.12}$$

where

$$\widetilde{\mathbb{C}} = \mathbf{I}_G, \qquad \widetilde{\mathbb{D}} = \mathbf{I}_H.$$

That is, we have

$$\boldsymbol{m} \;=\; \alpha_0 \mathbf{1}_{G\cdot H} \;+\; \left(\mathbf{1}_H \otimes \widetilde{\mathbb{C}}\right)\boldsymbol{\alpha}^Z \;+\; \left(\widetilde{\mathbb{D}} \otimes \mathbf{1}_G\right)\boldsymbol{\alpha}^W \;+\; \left(\widetilde{\mathbb{D}} \otimes \widetilde{\mathbb{C}}\right)\boldsymbol{\alpha}^{ZW}. \tag{9.13}$$

Lemma 9.2 Column rank of a matrix that parameterizes two-way classified group means.

Matrix $\widetilde{\mathbb{X}}$ *being divided into blocks as*

$$\widetilde{\mathbb{X}} = \left(\mathbf{1}_H \otimes \mathbf{1}_G \,\middle|\, \mathbf{1}_H \otimes \widetilde{\mathbb{C}} \,\middle|\, \widetilde{\mathbb{D}} \otimes \mathbf{1}_G \,\middle|\, \widetilde{\mathbb{D}} \otimes \widetilde{\mathbb{C}} \right)$$

has the column rank given by a product of column ranks of matrices $\left(\mathbf{1}_G,\ \widetilde{\mathbb{C}}\right)$ *and* $\left(\mathbf{1}_H,\ \widetilde{\mathbb{D}}\right)$*. That is,*

$$\text{col-rank}\left(\widetilde{\mathbb{X}}\right) \ = \ \text{col-rank}\Big(\left(\mathbf{1}_G,\ \widetilde{\mathbb{C}}\right)\Big) \ \cdot \ \text{col-rank}\Big(\left(\mathbf{1}_H,\ \widetilde{\mathbb{D}}\right)\Big).$$

Proof. **Proof/calculations below are shown only for those who are interested.**

By point (x) of Theorem A.3, matrix $\widetilde{\mathbb{X}}$ is upon suitable reordering of columns (which does not have any influence on the rank of the matrix) equal to a matrix

$$\widetilde{\mathbb{X}}_{reord} = \Big(\mathbf{1}_H \otimes \left(\mathbf{1}_G,\ \widetilde{\mathbb{C}}\right) \ \Big|\ \widetilde{\mathbb{D}} \otimes \left(\mathbf{1}_G,\ \widetilde{\mathbb{C}}\right)\Big).$$

Further, by point (ix) of Theorem A.3:

$$\widetilde{\mathbb{X}}_{reord} = \left(\mathbf{1}_H,\ \widetilde{\mathbb{D}}\right) \otimes \left(\mathbf{1}_G,\ \widetilde{\mathbb{C}}\right).$$

Finally, by point (xi) of Theorem A.3:

$$\text{col-rank}\left(\widetilde{\mathbb{X}}\right) = \text{col-rank}\left(\widetilde{\mathbb{X}}_{reord}\right) = \text{col-rank}\Big(\left(\mathbf{1}_G,\ \widetilde{\mathbb{C}}\right)\Big) \ \cdot \ \text{col-rank}\Big(\left(\mathbf{1}_H,\ \widetilde{\mathbb{D}}\right)\Big).$$

❑

Lemma 9.2 can now be used to get easily that the rank of the matrix $\widetilde{\mathbb{X}}_\alpha$ given by (9.11) is indeed $G \cdot H$ and hence it can be used to parameterize $G \cdot H$ two-way classified group means.

Sum constraints identification

Deficiency in the rank of the matrix $\widetilde{\mathbb{X}}_\alpha$ is $1 + G + H$. By Scheffé's theorem on identification in a linear model (Theorem 7.1), $(1+G+H)$ (or more) linear constraints on the regression coefficients $\boldsymbol{\alpha}$ are needed to identify the vector $\boldsymbol{\alpha}$ in the related linear model. In practice, the following set of $(2 + G + H)$ constraints is often used:

$$\begin{aligned} &\sum_{g=1}^{G} \alpha_g^Z = 0, && \sum_{h=1}^{H} \alpha_h^W = 0, \\ &\sum_{h=1}^{H} \alpha_{g,h}^{ZW} = 0, \quad g = 1, \ldots, G, && \sum_{g=1}^{G} \alpha_{g,h}^{ZW} = 0, \quad h = 1, \ldots, H, \end{aligned} \tag{9.14}$$

which in matrix notation is written as

$$\mathbb{A}\boldsymbol{\alpha} = \mathbf{0}_{2+G+H}, \qquad \mathbb{A} = \left(\begin{array}{ccccccc} 0 & \mathbf{1}_G^\top & \mathbf{0}_H^\top & \mathbf{0}_G^\top & \dots & \mathbf{0}_G^\top \\ \hline 0 & \mathbf{0}_G^\top & \mathbf{1}_H^\top & \mathbf{0}_G^\top & \dots & \mathbf{0}_G^\top \\ \hline \mathbf{0}_G & \mathbf{0}_{G\times G} & \mathbf{0}_{G\times H} & \mathbf{I}_G & \dots & \mathbf{I}_G \\ \hline 0 & \mathbf{0}_G^\top & \mathbf{0}_H^\top & \mathbf{1}_G^\top & \dots & \mathbf{0}_G^\top \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mathbf{0}_G^\top & \mathbf{0}_H^\top & \mathbf{0}_G^\top & \dots & \mathbf{1}_G^\top \end{array}\right).$$

We leave it as an exercise in linear algebra to verify that an $(2+G+H)\times(1+G+H+G\cdot H)$ matrix $\mathbb{A}$ satisfies conditions of Scheffé's theorem, i.e.,

$$\mathsf{rank}\big(\mathbb{A}\big) = 1+G+H, \qquad \mathcal{M}\big(\mathbb{A}^\top\big) \cap \mathcal{M}\Big(\big(\widetilde{\mathbb{X}}_\alpha\big)^\top\Big) = \{\mathbf{0}\}.$$

Interpretation of the regression coefficients under the sum constraints

The coefficients $\boldsymbol{\alpha}$ identified by a set of constraints (9.14) have the following (easy to see using simple algebra with expressions (9.10) while taking into account the constraints) useful interpretation.

$$\begin{array}{rcllll} \alpha_0 & = & \overline{m}, & & & \\ \alpha_g^Z & = & \overline{m}_{g\bullet} - \overline{m} & = & \eta_g^Z, & g = 1, \dots, G, \\ \alpha_h^W & = & \overline{m}_{\bullet h} - \overline{m} & = & \eta_h^W, & h = 1, \dots, H, \\ \alpha_{g,h}^{ZW} & = & m_{g,h} - \overline{m}_{g\bullet} - \overline{m}_{\bullet h} + \overline{m} & = & \eta_{g,h}^{ZW}, & g = 1, \dots, G, \\ & & & & & h = 1, \dots, H. \end{array}$$

That is, with the regression coefficients $\boldsymbol{\alpha}$ being identified by the sum constraints, the intercept is equal to the mean of all group means, the subvector $\boldsymbol{\alpha}^Z$ has a meaning of the main effects of the Z covariate, the subvector $\boldsymbol{\alpha}^W$ has a meaning of the main effects of the W covariate, and the subvector $\boldsymbol{\alpha}^{ZW}$ has a meaning of the interaction effects according to Definition 9.2. The most importantly, we have,

$$\begin{array}{rcrcrcll} \alpha_{g_1}^Z - \alpha_{g_2}^Z & = & \eta_{g_1}^Z - \eta_{g_2}^Z & = & \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet} & = & \theta_{g_1,g_2\bullet}, & g_1, g_2 = 1, \dots, G, \\ \alpha_{h_1}^W - \alpha_{h_2}^W & = & \eta_{h_1}^W - \eta_{h_2}^W & = & \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2} & = & \theta_{\bullet h_1,h_2}, & h_1, h_2 = 1, \dots, H. \end{array}$$

9.2.5 Full-rank parameterization of two-way classified group means

With the ANOVA parameterization, the model matrix $\mathbb{X}$ was derived from a matrix $\widetilde{\mathbb{X}}_\alpha$ given by (9.12). Let us start from this expression while using (pseudo)contrast matrices that we used to parameterize categorical covariates on place of $\widetilde{\mathbb{C}}$ and $\widetilde{\mathbb{D}}$.

Let

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix}, \qquad \begin{array}{l} \boldsymbol{c}_1 = \big(c_{1,1}, \dots, c_{1,G-1}\big)^\top, \\ \qquad \vdots \\ \boldsymbol{c}_G = \big(c_{G,1}, \dots, c_{G,G-1}\big)^\top \end{array}$$

be a $G \times (G-1)$ (pseudo)contrast matrix that could be used to parameterize the Z covariate (i.e., $\mathbf{1}_G \notin \mathcal{M}(\mathbb{C})$, $\mathsf{rank}(\mathbb{C}) = G-1$). Similarly, let

$$\mathbb{D} = \begin{pmatrix} \boldsymbol{d}_1^\top \\ \vdots \\ \boldsymbol{d}_H^\top \end{pmatrix}, \qquad \begin{aligned} \boldsymbol{d}_1 &= \big(d_{1,1}, \ldots, d_{1,H-1}\big)^\top, \\ &\vdots \\ \boldsymbol{d}_H &= \big(d_{H,1}, \ldots, d_{H,H-1}\big)^\top \end{aligned}$$

be an $H \times (H-1)$ (pseudo)contrast matrix that could be used to parameterize the W covariate (i.e., $\mathbf{1}_H \notin \mathcal{M}(\mathbb{D})$, $\mathsf{rank}(\mathbb{D}) = H-1$).

Note that we do not require that matrices $\mathbb{C}$ and $\mathbb{D}$ are based on (pseudo)contrasts of the same type. Let

$$\widetilde{\mathbb{X}}_\beta = \Big(\; \mathbf{1}_H \otimes \mathbf{1}_G \;\Big|\; \mathbf{1}_H \otimes \mathbb{C} \;\Big|\; \mathbb{D} \otimes \mathbf{1}_G \;\Big|\; \mathbb{D} \otimes \mathbb{C} \; \Big) \tag{9.15}$$

$$= \left(\begin{array}{c|c|c|c} 1 & \boldsymbol{c}_1^\top & \boldsymbol{d}_1^\top & \boldsymbol{d}_1^\top \otimes \boldsymbol{c}_1^\top \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \boldsymbol{c}_G^\top & \boldsymbol{d}_1^\top & \boldsymbol{d}_1^\top \otimes \boldsymbol{c}_G^\top \\ \hline \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \hline 1 & \boldsymbol{c}_1^\top & \boldsymbol{d}_H^\top & \boldsymbol{d}_G^\top \otimes \boldsymbol{c}_1^\top \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \boldsymbol{c}_G^\top & \boldsymbol{d}_H^\top & \boldsymbol{d}_G^\top \otimes \boldsymbol{c}_G^\top \end{array} \right),$$

which is a matrix with $G \cdot H$ rows and $1 + (G-1) + (H-1) + (G-1)(H-1) = G \cdot H$ columns and its structure is the same as a structure of the matrix (9.12). Using Lemma 9.2 and properties of (pseudo)contrast matrices, we have

$$\mathsf{col\text{-}rank}\big(\widetilde{\mathbb{X}}_\beta\big) = \mathsf{col\text{-}rank}\Big(\big(\mathbf{1}_G, \; \mathbb{C}\big)\Big) \; \cdot \; \mathsf{col\text{-}rank}\Big(\big(\mathbf{1}_H, \; \mathbb{D}\big)\Big) = G \cdot H.$$

That is, the matrix $\widetilde{\mathbb{X}}_\beta$ is of full-rank $G \cdot H$ and hence can be used to parameterize the two-way classified group means as

$$\boldsymbol{m} = \widetilde{\mathbb{X}}_\beta \boldsymbol{\beta}, \qquad \boldsymbol{\beta} = \big(\beta_0, \; \boldsymbol{\beta}^{Z\top}, \; \boldsymbol{\beta}^{W\top}, \; \boldsymbol{\beta}^{ZW\top}\big)^\top,$$

where

$$\boldsymbol{\beta}^Z = \big(\beta_1^Z, \cdots, \beta_{G-1}^Z\big)^\top, \qquad \boldsymbol{\beta}^W = \big(\beta_1^W, \cdots, \beta_{H-1}^W\big)^\top,$$
$$\boldsymbol{\beta}^{ZW} = \big(\beta_{1,1}^{ZW}, \ldots, \beta_{G-1,1}^{ZW}, \cdots\cdots, \beta_{1,H-1}^{ZW}, \ldots, \beta_{G-1,H-1}^{ZW}\big)^\top.$$

We can also write

$$\boldsymbol{m} \;=\; \beta_0 \mathbf{1}_{G\cdot H} \;+\; \big(\mathbf{1}_H \otimes \mathbb{C}\big)\boldsymbol{\beta}^Z \;+\; \big(\mathbb{D} \otimes \mathbf{1}_G\big)\boldsymbol{\beta}^W \;+\; \big(\mathbb{D} \otimes \mathbb{C}\big)\boldsymbol{\beta}^{ZW},$$

$$m_{g,h} \;=\; \beta_0 \;+\; \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z \;+\; \boldsymbol{d}_h^\top \boldsymbol{\beta}^W \;+\; \big(\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top\big)\boldsymbol{\beta}^{ZW}, \qquad g = 1, \ldots, G, \; h = 1, \ldots, H. \tag{9.16}$$

Different choices of the (pseudo)contrast matrices $\mathbb{C}$ and $\mathbb{D}$ lead to different interpretations of the regression coefficients $\boldsymbol{\beta}$.

Link to general interaction model

If we take expression (9.15) of the matrix $\widetilde{\mathbb{X}}_\beta$, it is directly seen that it can also be written as

$$\widetilde{\mathbb{X}}_\beta = \big(\mathbf{1}_{G\cdot H},\ \widetilde{\mathbb{S}}^Z,\ \widetilde{\mathbb{S}}^W,\ \widetilde{\mathbb{S}}^{ZW}\big),$$

where

$$\widetilde{\mathbb{S}}^Z = \mathbf{1}_H \otimes \mathbb{C}, \qquad \widetilde{\mathbb{S}}^W = \mathbb{D} \otimes \mathbf{1}_G, \qquad \widetilde{\mathbb{S}}^{ZW} = \widetilde{\mathbb{S}}^Z : \widetilde{\mathbb{S}}^W.$$

Similarly, the related model matrix $\mathbb{X}_\beta$ which parameterizes a vector $\boldsymbol{\mu}$ (in which a value $m_{g,h}$ is repeated $n_{g,h}$-times) as $\boldsymbol{\mu} = \mathbb{X}_\beta\boldsymbol{\beta}$ is factorized as

$$\mathbb{X}_\beta = \big(\mathbf{1}_n,\ \mathbb{S}^Z,\ \mathbb{S}^W,\ \mathbb{S}^{ZW}\big), \tag{9.17}$$

where $\mathbb{S}^Z$ and $\mathbb{S}^W$ is obtained from matrices $\widetilde{\mathbb{S}}^Z$ and $\widetilde{\mathbb{S}}^W$, respectively, by appropriately repeating their rows and $\mathbb{S}^{ZW} = \mathbb{S}^Z : \mathbb{S}^W$. That is, the model matrix (9.17) is precisely of the form given in (8.18) that we used to parameterize a linear model with interactions. In context of this chapter, the (pseudo)contrast matrices $\mathbb{C}$ and $\mathbb{D}$, respectively, play the role of parameterizations $\boldsymbol{s}_Z$ in (8.15) and $\boldsymbol{s}_W$ in (8.16), respectively and a linear model with the model matrix $\mathbb{X}_\beta$ is a linear model with interactions between two categorical covariates.

9.2.6 Relationship between the full-rank and ANOVA parameterizations

With the full-rank parameterization of the two-way classified group means, expression (9.16) shows that we can also write

$$m_{g,h} \ = \ \alpha_0 \ + \ \alpha_g^Z \ + \ \alpha_h^W \ + \ \alpha_{g,h}^{ZW}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H, \tag{9.18}$$

where

$$\begin{aligned}
\alpha_0 &:= \beta_0, \\
\alpha_g^Z &:= \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, && g = 1, \ldots, G, \\
\alpha_h^W &:= \boldsymbol{d}_h^\top \boldsymbol{\beta}^W, && h = 1, \ldots, H, \\
\alpha_{g,h}^{ZW} &:= \big(\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top\big)\boldsymbol{\beta}^{ZW}, && g = 1, \ldots, G,\ h = 1, \ldots, H.
\end{aligned} \tag{9.19}$$

That is, chosen full-rank parameterization of the two-way classified group means corresponds to the ANOVA parameterization (9.18) in which $(1 + G + H + G\cdot H)$ regression coefficients $\boldsymbol{\alpha}$ are *uniquely* obtained from $G\cdot H$ coefficients $\boldsymbol{\beta}$ of the full-rank parameterization using the relationships (9.19). In other words, expressions (9.19) correspond to identifying constraints on $\boldsymbol{\alpha}$ in the less-than-full-rank parameterization. Note also, that in matrix notation, (9.19) can be written as

$$\begin{aligned}
\alpha_0 &:= \beta_0, \\
\boldsymbol{\alpha}^Z &:= \mathbb{C}\,\boldsymbol{\beta}^Z, \\
\boldsymbol{\alpha}^W &:= \mathbb{D}\,\boldsymbol{\beta}^W, \\
\boldsymbol{\alpha}^{ZW} &:= (\mathbb{D} \otimes \mathbb{C})\,\boldsymbol{\beta}^{ZW}.
\end{aligned} \tag{9.20}$$

9.2.7 Additivity in the linear model parameterization

In Section 9.2.2, we have shown that necessary and sufficient condition for additivity of the effect of the two covariates on the response expectation is that the interaction terms $\eta_{g,h}^{ZW} = m_{g,h} - \overline{m}_{g\bullet} - \overline{m}_{\bullet h} + \overline{m}$, $g = 1, \ldots, G$, $h = 1, \ldots, H$, are all equal to zero. It is easily seen that in general ANOVA parameterization

$$m_{g,h} = \alpha_0 + \alpha_g^Z + \alpha_h^W + \alpha_{g,h}^{ZW}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H,$$

(where the vector $\boldsymbol{\alpha}$ of the regression coefficients is not unique with respect to the vector of $\boldsymbol{m}$ of the two-way classified group means) a sufficient condition for additivity is given as

$$\mathrm{H}_0: \quad \alpha_{1,1}^{ZW} = \cdots = \alpha_{G,H}^{ZW},$$

or written differently as

$$\mathrm{H}_0: \quad \boldsymbol{\alpha}^{ZW} = a\,\mathbf{1}_{G\cdot H} \qquad \text{for some } a \in \mathbb{R}.$$

By using similar calculations as in Section 9.2.2, we can find that this condition is also necessary condition of additivity.

If we take into account (9.20), which links the ANOVA parameterization and the full-rank parameterization, we have that $\boldsymbol{\alpha}^{ZW} = a\,\mathbf{1}$ for some $a \in \mathbb{R}$, if and only if $(\mathbb{D} \otimes \mathbb{C})\,\boldsymbol{\beta}^{ZW} = a\,\mathbf{1}$. Due to the fact that $\mathbf{1} \notin \mathcal{M}(\mathbb{D} \otimes \mathbb{C})$ (if both $\mathbb{C}$ and $\mathbb{D}$ are (pseudo)contrast matrices), this is only possible with $a = 0$ and $\boldsymbol{\beta}^{ZW} = \mathbf{0}$. Hence with the full-rank parameterization of the two-way classified group means, the hypothesis of additivity is (as expected) expressed by

$$\mathrm{H}_0: \boldsymbol{\beta}^{ZW} = \mathbf{0}_{(G-1)(H-1)}.$$

In the ANOVA parameterization, additivity ($\boldsymbol{\alpha}^{ZW} = a\,\mathbf{1}_{G\cdot H}$ for some $a \in \mathbb{R}$) corresponds to simplification of the interaction block $\mathbf{I}_H \otimes \mathbf{I}_G$ of the model matrix $\widetilde{\mathbb{X}}_\alpha$ in (9.11) into an intercept column $\mathbf{1}_{G\cdot H}$. In the full-rank parameterization, additivity ($\boldsymbol{\beta}^{ZW} = \mathbf{0}_{(G-1)(H-1)}$) corresponds to omitting the interaction block $\mathbb{D}\otimes\mathbb{C}$ from the model matrix $\widetilde{\mathbb{X}}_\beta$ given by (9.15). The model matrices in the two parameterizations become

$$\begin{aligned}
\widetilde{\mathbb{X}}_\alpha^{Z+W} &= \left(\mathbf{1}_H \otimes \mathbf{1}_G \,\middle|\, \mathbf{1}_H \otimes \mathbf{I}_G \,\middle|\, \mathbf{I}_H \otimes \mathbf{1}_G \right), \\
\widetilde{\mathbb{X}}_\beta^{Z+W} &= \left(\mathbf{1}_H \otimes \mathbf{1}_G \,\middle|\, \mathbf{1}_H \otimes \mathbb{C} \,\middle|\, \mathbb{D} \otimes \mathbf{1}_G \right),
\end{aligned}$$

where $\widetilde{\mathbb{X}}_\alpha^{Z+W}$ has $1 + G + H$ columns and $\widetilde{\mathbb{X}}_\beta^{Z+W}$ has $1 + (G-1) + (H-1) = G + H - 1$ columns. Both matrices are of the same rank

$$\mathsf{rank}\big(\widetilde{\mathbb{X}}_\alpha^{Z+W}\big) = \mathsf{rank}\big(\widetilde{\mathbb{X}}_\beta^{Z+W}\big) = G + H - 1.$$

That is, the matrix $\widetilde{\mathbb{X}}_\alpha^{Z+W}$ has a deficiency of 2 in its rank, the matrix $\widetilde{\mathbb{X}}_\beta^{Z+W}$ is of full rank.

The vector of the two-way classified group means is parameterized as

$$\begin{aligned}
\boldsymbol{m} &= \widetilde{\mathbb{X}}_\alpha^{Z+W}\boldsymbol{\alpha} &&= \alpha_0\mathbf{1}_{G\cdot H} + (\mathbf{1}_H \otimes \mathbf{I}_G)\boldsymbol{\alpha}^Z + (\mathbf{I}_H \otimes \mathbf{1}_G)\boldsymbol{\alpha}^W \\
&= \widetilde{\mathbb{X}}_\beta^{Z+W}\boldsymbol{\beta} &&= \beta_0\mathbf{1}_{G\cdot H} + (\mathbf{1}_H \otimes \mathbb{C})\boldsymbol{\beta}^Z + (\mathbb{D} \otimes \mathbf{1}_G)\boldsymbol{\beta}^W, \\
m_{g,h} &= && \alpha_0 + \alpha_g^Z + \alpha_h^W \\
&= && \beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta}^Z + \boldsymbol{d}_h^\top\boldsymbol{\beta}^W, \\
& && \qquad g = 1, \ldots, G,\ h = 1, \ldots, H,
\end{aligned} \tag{9.21}$$

where the related vectors of regression coefficients are

$$\begin{aligned}\boldsymbol{\alpha} &= \big(\alpha_0, \underbrace{\alpha_1^Z, \ldots, \alpha_G^Z}_{\boldsymbol{\alpha}^Z}, \underbrace{\alpha_1^W, \ldots, \alpha_H^W}_{\boldsymbol{\alpha}^W}\big)^\top,\\ \boldsymbol{\beta} &= \big(\beta_0, \underbrace{\beta_1^Z, \ldots, \beta_{G-1}^Z}_{\boldsymbol{\beta}^Z}, \underbrace{\beta_1^W, \ldots, \beta_{H-1}^W}_{\boldsymbol{\beta}^W}\big)^\top.\end{aligned}$$

Two $(1 + G + H - (G + H - 1))$ constraints are needed to identify the coefficients of the ANOVA parameterization. This can be achieved, e.g., by using the identifying constraints $\mathbb{A}\boldsymbol{\alpha} = \mathbf{0}$ with

$$\mathbb{A} = \begin{pmatrix} 0 & | & 1 & \ldots & 1 & | & 0 & \ldots & 0 \\ 0 & | & 0 & \ldots & 0 & | & 1 & \ldots & 1 \end{pmatrix},$$

i.e., by using two sum constraints

$$\sum_{g=1}^{G} \alpha_g^Z = 0, \qquad \sum_{h=1}^{H} \alpha_h^W = 0. \tag{9.22}$$

It can be easily checked that having considered the sum constraints (9.22), coefficients $\boldsymbol{\alpha}^Z$ and $\boldsymbol{\alpha}^W$ lead to the corresponding main effects, i.e.,

$$\begin{aligned}\alpha_g^Z &= \overline{m}_{g\bullet} - \overline{m} = \eta_g^Z \qquad g = 1, \ldots, G,\\ \alpha_h^Z &= \overline{m}_{\bullet h} - \overline{m} = \eta_h^W \qquad h = 1, \ldots, H.\end{aligned}$$

Partial effects

As explained in Section 9.2.2, we have under additivity that for any $g_1, g_2 = 1, \ldots, G$

$$\theta_{g_1,g_2\bullet} = \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet} = m_{g_1,h} - m_{g_2,h}$$

does not depend on the value of $h = 1, \ldots, H$ and hence determines the partial effects of the covariate Z on the response expectation given the covariate W. With the two considered parameterizations, the related quatities are calculated as

$$\theta_{g_1,g_2\bullet} = \alpha_{g_1}^Z - \alpha_{g_2}^Z = \big(\boldsymbol{c}_{g_1} - \boldsymbol{c}_{g_2}\big)^\top \boldsymbol{\beta}^Z. \tag{9.23}$$

Note that (9.23) holds for the $\boldsymbol{\alpha}$ vector being arbitrarily identified (i.e., not necessarily by the sum constraints).

Analogously for the partial effects of the W covariate given the Z covariate, for any $h_1, h_2 = 1, \ldots, H$

$$\theta_{\bullet h_1,h_2} = \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2} = m_{g,h_1} - m_{g,h_2}$$

does not depend on the value of $g = 1, \ldots, G$ and

$$\theta_{\bullet h_1,h_2} = \alpha_{h_1}^W - \alpha_{h_2}^W = \big(\boldsymbol{d}_{h_1} - \boldsymbol{d}_{h_2}\big)^\top \boldsymbol{\beta}^W.$$

9.2.8 Interpretation of model parameters for selected choices of (pseudo)contrasts

Let us again consider the full-rank parameterizations of the two-way classified group means, i.e.,

$$m_{g,h} = \beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta}^Z + \boldsymbol{d}_h^\top\boldsymbol{\beta}^W + (\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top)\boldsymbol{\beta}^{ZW}, \quad g = 1, \ldots, G,\ h = 1, \ldots, H, \tag{9.24}$$

$$\boldsymbol{m} = \beta_0\mathbf{1}_{G\cdot H} + (\mathbf{1}_H \otimes \mathbb{C})\boldsymbol{\beta}^Z + (\mathbb{D} \otimes \mathbf{1}_G)\boldsymbol{\beta}^W + (\mathbb{D} \otimes \mathbb{C})\boldsymbol{\beta}^{ZW},$$

where $\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_G^\top$ are rows of the $G \times (G-1)$ (pseudo)contrast matrix $\mathbb{C}$ and $\boldsymbol{d}_1^\top, \ldots, \boldsymbol{d}_H^\top$ are rows of the $H \times (H-1)$ (pseudo)contrast matrix $\mathbb{D}$, and $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}^{Z\top}, \boldsymbol{\beta}^{W\top}, \boldsymbol{\beta}^{ZW\top}\right)^\top$ are related regression coefficients.

Chosen (pseudo)contrast matrices $\mathbb{C}$ and $\mathbb{D}$ determine interpretation of the coefficients $\boldsymbol{\beta}$ of the full-rank parameterization (9.24) of the two-way classified group means and also of the coefficients $\boldsymbol{\alpha}$ given by (9.19) and determine an ANOVA parameterization with certain identification. For interpretation, it is useful to view the two-way classified group means as entries in the $G \times H$ table as shown in Table 9.3. Corresponding sample sizes $n_{g,h}$, $g = 1, \ldots, G$, $h = 1, \ldots, H$, form a $G \times H$ contingency table based on the values of the two categorical covariates Z and W (see also Table 9.2).

With the ANOVA parameterization

$$m_{g,h} = \alpha_0 + \alpha_g^Z + \alpha_h^W + \alpha_{g,h}^{ZW}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H,$$

$$\boldsymbol{m} = \alpha_0\mathbf{1}_{G\cdot H} + (\mathbf{1}_H \otimes \mathbf{I}_G)\boldsymbol{\alpha}^Z + (\mathbf{I}_H \otimes \mathbf{1}_G)\boldsymbol{\alpha}^W + (\mathbf{I}_H \otimes \mathbf{I}_G)\boldsymbol{\alpha}^{ZW},$$

the coefficients $\boldsymbol{\alpha}^Z = \mathbb{C}\boldsymbol{\beta}^Z$ and $\boldsymbol{\alpha}^W = \mathbb{D}\boldsymbol{\beta}^W$ can be interpreted as the row and column effects,

Table 9.3: Decomposition of the two-way classified group means into row, column and cell effects.

Group means

	W		
Z	1	…	H
1	$m_{1,1}$	…	$m_{1,H}$
⋮	⋮	…	⋮
G	$m_{G,1}$	…	$m_{G,H}$

≡

ANOVA parameterization

	W			
Z	1	…	H	
1	$\alpha_{1,1}^{ZW}$	…	$\alpha_{1,H}^{ZW}$	$+\,\alpha_1^Z$
⋮	⋮	…	⋮	⋮
G	$\alpha_{G,1}^{ZW}$	…	$\alpha_{G,H}^{ZW}$	$+\,\alpha_G^Z$
	$+\,\alpha_1^W$	…	$+\,\alpha_H^W$	$+\,\alpha_0$

≡

Full-rank parameterization

	W			
Z	1	…	H	
1	$(\boldsymbol{d}_1^\top \otimes \boldsymbol{c}_1^\top)\boldsymbol{\beta}^{ZW}$	…	$(\boldsymbol{d}_H^\top \otimes \boldsymbol{c}_1^\top)\boldsymbol{\beta}^{ZW}$	$+\,\boldsymbol{c}_1^\top\boldsymbol{\beta}^Z$
⋮	⋮	⋮	⋮	⋮
G	$(\boldsymbol{d}_1^\top \otimes \boldsymbol{c}_G^\top)\boldsymbol{\beta}^{ZW}$	…	$(\boldsymbol{d}_H^\top \otimes \boldsymbol{c}_G^\top)\boldsymbol{\beta}^{ZW}$	$+\,\boldsymbol{c}_G^\top\boldsymbol{\beta}^Z$
	$+\,\boldsymbol{d}_1^\top\boldsymbol{\beta}^W$	…	$+\,\boldsymbol{d}_H^\top\boldsymbol{\beta}^W$	$+\,\beta_0$

respectively, in the group means table. The interaction effects $\boldsymbol{\alpha}^{ZW} = (\mathbb{D} \otimes \mathbb{C})\boldsymbol{\beta}^{ZW}$ can also be placed in a table (see Table 9.3) whose entries can be interpreted as cell effects in the group means table. In other words, each group mean is obtained as a sum of the intercept term $\alpha_0 = \beta_0$ and corresponding row, column and cell effects as depicted in Table 9.3.

As was already mentioned, the two (pseudo)contrast matrices $\mathbb{C}$ and $\mathbb{D}$ can be both of different type, e.g., matrix $\mathbb{C}$ being the reference group pseudocontrast matrix and matrix $\mathbb{D}$ being the sum contrast matrix. Nevertheless, in practice, both of them are mostly chosen as being of the same type. In the reminder of this section, we shall discuss interpretation of the model parameters for two most common choices of the (pseudo)contrasts which are (i) the reference group pseudocontrasts and (ii) the sum contrasts.

Reference group pseudocontrasts

Suppose that both $\mathbb{C}$ and $\mathbb{D}$ are reference group pseudocontrasts, i.e.,

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \boldsymbol{c}_2^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix} = \begin{pmatrix} 0 & \dots & 0 \\ 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{G-1}^\top \\ \mathbf{I}_{G-1} \end{pmatrix}, \quad \mathbb{D} = \begin{pmatrix} \boldsymbol{d}_1^\top \\ \boldsymbol{d}_2^\top \\ \vdots \\ \boldsymbol{d}_H^\top \end{pmatrix} = \begin{pmatrix} 0 & \dots & 0 \\ 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{H-1}^\top \\ \mathbf{I}_{H-1} \end{pmatrix}.$$

We have,

$$\begin{pmatrix} \alpha_1^Z \\ \alpha_2^Z \\ \vdots \\ \alpha_G^Z \end{pmatrix} = \boldsymbol{\alpha}^Z = \mathbb{C}\boldsymbol{\beta}^Z = \begin{pmatrix} 0 \\ \beta_1^Z \\ \vdots \\ \beta_{G-1}^Z \end{pmatrix}, \qquad \begin{pmatrix} \alpha_1^W \\ \alpha_2^W \\ \vdots \\ \alpha_H^W \end{pmatrix} = \boldsymbol{\alpha}^W = \mathbb{D}\boldsymbol{\beta}^W = \begin{pmatrix} 0 \\ \beta_1^W \\ \vdots \\ \beta_{H-1}^W \end{pmatrix}, \tag{9.25}$$

To get the link between the full-rank interaction terms $\boldsymbol{\beta}^{ZW}$ and their ANOVA counterparts $\boldsymbol{\alpha}^{ZW}$, we have to explore the form of vectors $\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top$, $g = 1, \dots, G$, $h = 1, \dots, H$. With the reference group pseudocontrasts, we easily see that

$$\begin{aligned} \boldsymbol{d}_h^\top \otimes \boldsymbol{c}_1^\top &= \mathbf{0}, && \text{for all } h = 1, \dots, H, \\ \boldsymbol{d}_1^\top \otimes \boldsymbol{c}_g^\top &= \mathbf{0}, && \text{for all } g = 1, \dots, G, \\ \boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top &= (0, \dots, 1, \dots, 0), && \text{if } g \neq 1 \ \& \ h \neq 1, \\ & && 1 \text{ on a place that in } (\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top)\boldsymbol{\beta}^{ZW} \text{ multiplies } \beta_{g-1,h-1}^{ZW}, \end{aligned}$$

which leads to

$$\begin{pmatrix} \alpha_{1,1}^{ZW} & \alpha_{1,2}^{ZW} & \dots & \alpha_{1,H}^{ZW} \\ \alpha_{2,1}^{ZW} & \alpha_{2,2}^{ZW} & \dots & \alpha_{2,H}^{ZW} \\ \vdots & \vdots & \dots & \vdots \\ \alpha_{G,1}^{ZW} & \alpha_{G,2}^{ZW} & \dots & \alpha_{G,H}^{ZW} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \beta_{1,1}^{ZW} & \dots & \beta_{1,H-1}^{ZW} \\ \vdots & \vdots & \dots & \vdots \\ 0 & \beta_{G-1,1}^{ZW} & \dots & \beta_{G-1,H-1}^{ZW} \end{pmatrix}. \tag{9.26}$$

That is, decomposition of the two-way classified group means becomes as shown in Table 9.4. This leads to the following interpretation of the regression coefficients in the ANOVA and full-rank

Table 9.4: Decomposition of the two-way classified group means with the reference group pseudo-contrasts.

Group means

	W			
Z	1	2	...	H
1	$m_{1,1}$	$m_{1,2}$	...	$m_{1,H}$
2	$m_{2,1}$	$m_{2,2}$	...	$m_{2,H}$
⋮	⋮	⋮	...	⋮
G	$m_{G,1}$	$m_{G,2}$	...	$m_{G,H}$

$\equiv$

ANOVA parameterization

	W				
Z	1	2	...	H	
1	$\alpha_{1,1}^{ZW}$	$\alpha_{1,2}^{ZW}$	...	$\alpha_{1,H}^{ZW}$	$+\alpha_1^Z$
2	$\alpha_{2,1}^{ZW}$	$\alpha_{2,2}^{ZW}$	...	$\alpha_{2,H}^{ZW}$	$+\alpha_2^Z$
⋮	⋮	⋮	...	⋮	⋮
G	$\alpha_{G,1}^{ZW}$	$\alpha_{G,2}^{ZW}$	...	$\alpha_{G,H}^{ZW}$	$+\alpha_G^Z$
	$+\alpha_1^W$	$+\alpha_2^W$	...	$+\alpha_H^W$	$+\alpha_0$

$\equiv$

Reference group pseudocontrasts

	W				
Z	1	2	...	H	
1	0	0	...	0	$+0$
2	0	$\beta_{1,1}^{ZW}$	...	$\beta_{1,H-1}^{ZW}$	$+\beta_1^Z$
⋮	⋮	⋮	...	⋮	⋮
G	0	$\beta_{G-1,1}^{ZW}$	...	$\beta_{G-1,H-1}^{ZW}$	$+\beta_{G-1}^Z$
	$+0$	$+\beta_1^W$	...	$+\beta_{H-1}^W$	$+\beta_0$

parameterizations:

$$
\begin{aligned}
\alpha_0 = \beta_0 \qquad &= \quad m_{1,1}, \\
\alpha_g^Z = \beta_{g-1}^Z \qquad &= \quad m_{g,1} - m_{1,1}, && g = 2, \ldots, G, \\
\alpha_h^W = \beta_{h-1}^W \qquad &= \quad m_{1,h} - m_{1,1}, && h = 2, \ldots, H, \\
\alpha_{g,h}^{ZW} = \beta_{g-1,h-1}^{ZW} \quad &= \quad m_{g,h} - m_{g,1} - m_{1,h} + m_{1,1}, && g = 2, \ldots, G, \quad h = 2, \ldots, H.
\end{aligned}
$$

It is also seen from Table 9.4 that the ANOVA coefficients are identified by a set of $3 + (H-1) + (G-1) = G + H + 1$ constraints

$$
\alpha_1^Z = 0, \qquad \alpha_1^W = 0,
$$

$$
\alpha_{1,1}^{ZW} = 0, \qquad \alpha_{1,h}^{ZW} = 0, \quad h = 2, \ldots, H, \qquad \alpha_{g,1}^{ZW} = 0, \quad g = 2, \ldots, G.
$$

The first two constraints come from (9.25), remaining ones correspond to zeros in the matrix (9.26).

***Note** (Reference group pseudocontrasts in the additive model).*

If the additive model is assumed where $m_{g,h} = \alpha_0 + \alpha_g^Z + \alpha_h^W$, $g = 1, \ldots, G$, $h = 1, \ldots, H$, and the reference group pseudocontrasts are used in the full-rank parameterization in (9.21), the ANOVA coefficients $\boldsymbol{\alpha} = \left(\alpha_0, \boldsymbol{\alpha}^{Z\top}, \boldsymbol{\alpha}^{W\top}\right)^\top$ are obtained from the full-rank coefficients $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}^{Z\top}, \boldsymbol{\beta}^{W\top}\right)^\top$ again by (9.25), that is, they are identified by two constraints

$$\alpha_1^Z = 0, \qquad \alpha_1^W = 0.$$

Their interpretation becomes

$$\begin{aligned}
\alpha_0 = \beta_0 &= m_{1,1}, \\
\alpha_g^Z = \beta_{g-1}^Z &= m_{g,h} - m_{1,h}, \qquad g = 2, \ldots, G, \quad \text{arbitrary } h \in \{1, \ldots, H\} \\
&= \overline{m}_{g\bullet} - \overline{m}_{1\bullet}, \\
\alpha_h^W = \beta_{h-1}^W &= m_{g,h} - m_{g,1}, \qquad h = 2, \ldots, H, \quad \text{arbitrary } g \in \{1, \ldots, G\} \\
&= \overline{m}_{\bullet h} - \overline{m}_{\bullet 1}.
\end{aligned}$$

Sum contrasts

Suppose now that both $\mathbb{C}$ and $\mathbb{D}$ are the sum contrasts, i.e.,

$$\mathbb{C} = \begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \\ -1 & \ldots & -1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_{G-1}^\top \\ \boldsymbol{c}_G^\top \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{G-1} \\ -\mathbf{1}_{G-1}^\top \end{pmatrix},$$

$$\mathbb{D} = \begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \\ -1 & \ldots & -1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{d}_1^\top \\ \vdots \\ \boldsymbol{d}_{H-1}^\top \\ \boldsymbol{d}_H^\top \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{H-1} \\ -\mathbf{1}_{H-1}^\top \end{pmatrix}.$$

We have,

$$\begin{aligned}
\begin{pmatrix} \alpha_1^Z \\ \vdots \\ \alpha_{G-1}^Z \\ \alpha_G^Z \end{pmatrix} = \boldsymbol{\alpha}^Z = \mathbb{C}\boldsymbol{\beta}^Z = \begin{pmatrix} \beta_1^Z \\ \vdots \\ \beta_{G-1}^Z \\ -\sum_{g=1}^{G-1} \beta_g^Z \end{pmatrix}, \\
\begin{pmatrix} \alpha_1^W \\ \vdots \\ \alpha_{H-1}^W \\ \alpha_H^W \end{pmatrix} = \boldsymbol{\alpha}^W = \mathbb{D}\boldsymbol{\beta}^W = \begin{pmatrix} \beta_1^W \\ \vdots \\ \beta_{H-1}^W \\ -\sum_{h=1}^{H-1} \beta_h^W \end{pmatrix},
\end{aligned} \tag{9.27}$$

The form of the vectors $\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top$, $g = 1, \ldots, G$, $h = 1, \ldots, H$, needed to calculate the interaction terms $\alpha_{g,h}^{ZW}$ is the following

$$\begin{aligned}
\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top &= (0, \ldots, 1, \ldots, 0), && \text{for } g = 1, \ldots, G-1, \\
& && \quad h = 1, \ldots, H-1, \\
& \text{1 on a place that in } (\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top)\boldsymbol{\beta}^{ZW} \text{ multiplies } \beta_{g,h}^{ZW}, \\
\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_G^\top &= (\mathbf{0}_{G-1}, \ldots, -\mathbf{1}_{G-1}, \ldots, \mathbf{0}_{G-1}), && \text{for all } h = 1, \ldots, H-1, \\
& -\mathbf{1}_{G-1} \text{ block on places that in } (\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_G^\top)\boldsymbol{\beta}^{ZW} \text{ multiply } \beta_{\bullet h}^{ZW}, \\
\boldsymbol{d}_H^\top \otimes \boldsymbol{c}_g^\top &= (0, \ldots, -1, \ldots, 0, \ldots \ldots, 0, \ldots, -1, \ldots, 0), && \text{for all } g = 1, \ldots, G-1, \\
& -1\text{'s on places that in } (\boldsymbol{d}_H^\top \otimes \boldsymbol{c}_g^\top)\boldsymbol{\beta}^{ZW} \text{ multiply } \beta_{g\bullet}^{ZW}, \\
\boldsymbol{d}_H^\top \otimes \boldsymbol{c}_G^\top &= (1, \ldots, 1) = \mathbf{1}_{(G-1)(H-1)}.
\end{aligned}$$

This leads to

$$\begin{pmatrix}
\alpha_{1,1}^{ZW} & \cdots & \alpha_{1,H-1}^{ZW} & \alpha_{1,H}^{ZW} \\
\vdots & \vdots & \vdots & \vdots \\
\alpha_{G-1,1}^{ZW} & \cdots & \alpha_{G-1,H-1}^{ZW} & \alpha_{G-1,H}^{ZW} \\
\alpha_{G,1}^{ZW} & \cdots & \alpha_{G,H-1}^{ZW} & \alpha_{G,H}^{ZW}
\end{pmatrix}$$

$$= \begin{pmatrix}
\beta_{1,1}^{ZW} & \cdots & \beta_{1,H-1}^{ZW} & -\sum_{h=1}^{H-1} \beta_{1,h}^{ZW} \\
\vdots & \cdots & \vdots & \vdots \\
\beta_{G-1,1}^{ZW} & \cdots & \beta_{G-1,H-1}^{ZW} & -\sum_{h=1}^{H-1} \beta_{G-1,h}^{ZW} \\
-\sum_{g=1}^{G-1} \beta_{g,1}^{ZW} & \cdots & -\sum_{g=1}^{G-1} \beta_{g,H-1}^{ZW} & \sum_{g=1}^{G-1} \sum_{h=1}^{H-1} \beta_{g,h}^{ZW}
\end{pmatrix}.$$

That is, decomposition of the two-way classified group means becomes as shown in Table 9.5.

Note that the entries in each row of the table with the cell effects and also in each column sum up to zero. Similarly, the row effects (coefficients $\boldsymbol{\alpha}^Z$) and also the column effects (coefficients $\boldsymbol{\alpha}^W$) sum up to zero. Identifying constraints for the ANOVA coefficients $\boldsymbol{\alpha}$ that correspond to considered sum contrast full-rank parameterization are hence

$$\begin{aligned}
&\sum_{g=1}^{G} \alpha_g^Z = 0, \qquad \sum_{h=1}^{H} \alpha_h^W = 0, \\
&\sum_{h=1}^{H} \alpha_{g,h}^{ZW} = 0, \quad \text{for each } g = 1, \ldots, G, \\
&\sum_{g=1}^{G} \alpha_{g,h}^{ZW} = 0, \quad \text{for each } h = 1, \ldots, H.
\end{aligned} \tag{9.28}$$

Note that in a set of $G + H$ constraints on the interaction terms $\alpha_{g,h}^{ZW}$, one constraint is redundant and the last two rows could also be replaced by a set of $(G-1) + (H-1) + 1 = G + H - 1$

Table 9.5: Decomposition of the two-way classified group means with the sum contrasts.

Group means

	W			
Z	1	$\dots$	$H-1$	H
1	$m_{1,1}$	$\dots$	$m_{1,H-1}$	$m_{1,H}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$G-1$	$m_{G-1,1}$	$\dots$	$m_{G-1,H-1}$	$m_{G-1,H}$
G	$m_{G,1}$	$\dots$	$m_{G,H-1}$	$m_{G,H}$

$\equiv$

ANOVA parameterization

	W				
Z	1	$\dots$	$H-1$	H	
1	$\alpha^{ZW}_{1,1}$	$\dots$	$\alpha^{ZW}_{1,H-1}$	$\alpha^{ZW}_{1,H}$	$+\alpha^{Z}_{1}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$
$G-1$	$\alpha^{ZW}_{G-1,1}$	$\dots$	$\alpha^{ZW}_{G-1,H-1}$	$\alpha^{ZW}_{G-1,H}$	$+\alpha^{Z}_{G-1}$
G	$\alpha^{ZW}_{G,1}$	$\dots$	$\alpha^{ZW}_{G,H-1}$	$\alpha^{ZW}_{G,H}$	$+\alpha^{Z}_{G}$
	$+\alpha^{W}_{1}$	$\dots$	$+\alpha^{W}_{H-1}$	$+\alpha^{W}_{H}$	$+\alpha_0$

$\equiv$

Sum contrasts

	W				
Z	1	$\dots$	$H-1$	H	
1	$\beta^{ZW}_{1,1}$	$\dots$	$\beta^{ZW}_{1,H-1}$	$-\sum_{h=1}^{H-1}\beta^{ZW}_{1,h}$	$+\beta^{Z}_{1}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$
$G-1$	$\beta^{ZW}_{G-1,1}$	$\dots$	$\beta^{ZW}_{G-1,H-1}$	$-\sum_{h=1}^{H-1}\beta^{ZW}_{G-1,h}$	$+\beta^{Z}_{G-1}$
G	$-\sum_{g=1}^{G-1}\beta^{ZW}_{g,1}$	$\dots$	$-\sum_{g=1}^{G-1}\beta^{ZW}_{g,H-1}$	$\sum_{g=1}^{G-1}\sum_{h=1}^{H-1}\beta^{ZW}_{g,h}$	$-\sum_{g=1}^{G-1}\beta^{Z}_{g}$
	$+\beta^{W}_{1}$	$\dots$	$+\beta^{W}_{H-1}$	$-\sum_{h=1}^{H-1}\beta^{W}_{h}$	$+\beta_0$

---

constraints:

$$\sum_{h=1}^{H}\alpha^{ZW}_{g,h}=0, \qquad \text{for each } g=1,\dots,G-1,$$

$$\sum_{g=1}^{G}\alpha^{ZW}_{g,h}=0, \qquad \text{for each } h=1,\dots,H-1,$$

$$\sum_{g=1}^{G-1}\sum_{h=1}^{H-1}\alpha^{ZW}_{g,h}=\alpha^{ZW}_{G,H}.$$

We see that the set of equations (9.28) exactly corresponds to identification by the sum constraints (9.14), see Section 9.2.4. Hence the interpretation of the regression coefficients is the same as derived there, namely,

$$\begin{aligned}
\alpha_0 &= \overline{m}, \\
\alpha^{Z}_{g} &= \overline{m}_{g\bullet}-\overline{m}, && g=1,\dots,G, \\
\alpha^{W}_{h} &= \overline{m}_{\bullet h}-\overline{m}, && h=1,\dots,H, \\
\alpha^{ZW}_{g,h} &= m_{g,h}-\overline{m}_{g\bullet}-\overline{m}_{\bullet h}+\overline{m}, && g=1,\dots,G, \quad h=1,\dots,H.
\end{aligned}$$

Additionally,

$$\begin{aligned} \alpha^Z_{g_1} - \alpha^Z_{g_2} &= \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet}, \qquad g_1,\, g_2 = 1,\, \ldots,\, G, \\ \alpha^W_{h_1} - \alpha^W_{h_2} &= \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2}, \qquad h_1,\, h_2 = 1,\, \ldots,\, H. \end{aligned}$$

***Note** (Sum contrasts in the additive model).*

If the additive model is assumed where $m_{g,h} = \alpha_0 + \alpha^Z_g + \alpha^W_h$, $g = 1, \ldots, G$, $h = 1, \ldots, H$ and the sum contrasts are used in the full-rank parameterization (9.21), the ANOVA coefficients $\boldsymbol{\alpha} = \left(\alpha_0,\, \boldsymbol{\alpha}^{Z\top},\, \boldsymbol{\alpha}^{W\top}\right)^\top$ are obtained from the full-rank coefficients $\boldsymbol{\beta} = \left(\beta_0,\, \boldsymbol{\beta}^{Z\top},\, \boldsymbol{\beta}^{W\top}\right)^\top$ again by (9.27), that is, they are identified by two constraints

$$\sum_{g=1}^{G} \alpha^Z_g = 0, \qquad \sum_{h=1}^{H} \alpha^W_h = 0.$$

Their interpretation becomes

$$\begin{aligned} \alpha_0 &= \overline{m}, \\ \alpha^Z_g &= m_{g,h} - \overline{m}_{\bullet h}, \qquad g = 1,\, \ldots,\, G, \quad \text{arbitrary } h \in \{1, \ldots, H\} \\ &= \overline{m}_{g\bullet} - \overline{m}, \\ \alpha^W_h &= m_{g,h} - \overline{m}_{g\bullet}, \qquad h = 1,\, \ldots,\, H, \quad \text{arbitrary } g \in \{1, \ldots, G\} \\ &= \overline{m}_{\bullet h} - \overline{m}. \end{aligned}$$

Additionally,

$$\begin{aligned} \alpha^Z_{g_1} - \alpha^Z_{g_2} &= m_{g_1,h} - m_{g_2,h}, \qquad g_1,\, g_2 = 1,\, \ldots,\, G, \quad \text{arbitrary } h \in \{1, \ldots, H\} \\ &= \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet}, \\ \alpha^W_{h_1} - \alpha^W_{h_2} &= m_{g,h_1} - m_{g,h_2}, \qquad h_1,\, h_2 = 1,\, \ldots,\, H, \quad \text{arbitrary } g \in \{1, \ldots, G\} \\ &= \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2}. \end{aligned}$$

9.2.9 Two-way ANOVA models

In this section, we explore in few more details properties of the linear models that can be considered in context of two-way classification. They are as follows and each of them corresponds to different structure for the two-way classified group means.

Interaction model

No structure is imposed on the group means that, in the two considered parameterizations, are written as

$$\begin{aligned} m_{g,h} &= \alpha_0 + \alpha^Z_g + \alpha^W_h + \alpha^{ZW}_{g,h}, \\ &= \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z + \boldsymbol{d}_h^\top \boldsymbol{\beta}^W + \left(\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top\right)\boldsymbol{\beta}^{ZW}, \qquad (9.29) \\ &\qquad\qquad g = 1, \ldots, H,\, h = 1, \ldots, H. \end{aligned}$$

where

$$\alpha_0, \quad \boldsymbol{\alpha}^Z = \left(\alpha_1^Z, \ldots, \alpha_G^Z\right)^\top, \quad \boldsymbol{\alpha}^W = \left(\alpha_1^W, \ldots, \alpha_H^W\right)^\top,$$
$$\boldsymbol{\alpha}^{ZW} = \left(\alpha_{1,1}^{ZW}, \ldots, \alpha_{G,1}^{ZW}, \ \ldots, \ \alpha_{1,H}^{ZW}, \ldots, \alpha_{G,H}^{ZW}\right)^\top$$

are the regression parameters in the (less-than-full rank) ANOVA parameterization and

$$\beta_0, \quad \boldsymbol{\beta}^Z = \left(\beta_1^Z, \ldots, \beta_{G-1}^Z\right)^\top, \quad \boldsymbol{\beta}^W = \left(\beta_1^W, \ldots, \beta_{H-1}^W\right)^\top,$$
$$\boldsymbol{\beta}^{ZW} = \left(\beta_{1,1}^{ZW}, \ldots, \beta_{G-1,1}^{ZW}, \ \ldots, \ \beta_{1,H-1}^{ZW}, \ldots, \beta_{G-1,H-1}^{ZW}\right)^\top$$

are the regression parameters in the full-rank parameterization given by the (pseudo)contrast matrices with rows $\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_G^\top$ and $\boldsymbol{d}_1^\top, \ldots, \boldsymbol{d}_H^\top$, respectively.

If (almost surely) $n_{g,h} > 0$ for all g, h, the rank of the related linear model is (almost surely) $G \cdot H$, see Section 9.2.3. This explains why the interaction model is also called as *the saturated model*[5]. The reason is that its regression space has maximal possible vector dimension equal to the number of the group means.

In the following, let symbols Z and W denote the terms in the model matrix that correspond to the coefficients $\boldsymbol{\alpha}^Z$ or $\boldsymbol{\beta}^Z$, and $\boldsymbol{\alpha}^W$ or $\boldsymbol{\beta}^W$, respectively. Let further Z : W denote the terms corresponding to the interaction coefficients $\boldsymbol{\alpha}^{ZW}$ or $\boldsymbol{\beta}^{ZW}$. The interaction model will then symbolically be written as

$$\mathsf{M}_{ZW}\colon \ \sim \mathsf{Z} + \mathsf{W} + \mathsf{Z}{:}\mathsf{W}.$$

Additive model

It is obtained as a submodel of the interaction model (9.29) where it is requested

$$\alpha_{1,1}^{ZW} = \cdots = \alpha_{G,H}^{ZW},$$

which in the full-rank parameterization corresponds to requesting

$$\boldsymbol{\beta}^{ZW} = \mathbf{0}_{(G-1)\cdot(H-1)}.$$

Hence the group means can be written as

$$\begin{aligned} m_{g,h} &= \alpha_0 + \alpha_g^Z + \alpha_h^W, \\ &= \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z + \boldsymbol{d}_h^\top \boldsymbol{\beta}^W, \qquad g = 1, \ldots, H, \ h = 1, \ldots, H. \end{aligned} \tag{9.30}$$

In Section 9.2.7, we have shown that if $n_{g,h} > 0$ (almost surely) for all g, h, the rank of the linear model with the two-way classified group means that satisfy (9.30), is $G + H - 1$ (almost surely). The additive model will symbolically be written as

$$\mathsf{M}_{Z+W}\colon \ \sim \mathsf{Z} + \mathsf{W}.$$

Note. It can easily be shown that $n_{g\bullet}$ for all $g = 1, \ldots, G$ and $n_{\bullet h}$ for all $h = 1, \ldots, H$ suffice to get a rank of the related linear model being still $G+H-1$. This guarantees, among other things, that all parameters that are estimable in the additive model with $n_{g,h} > 0$ for all g, h, are still estimable under a weaker requirement $n_{g\bullet}$ for all $g = 1, \ldots, G$ and $n_{\bullet h}$ for all $h = 1, \ldots, H$. That is, if the additive model can be assumed, it is not necessary to have observations for all possible combinations of the values of the two covariates (factors) and the same types of the statistical

[5] *saturovaný model*

inference are possible. This is often exploited in the area of designed experiments where it might be impractical or even impossible to get observations under all possible covariate combinations.

See Section 9.2.2 what the additive model implies for the two-way classified group means. Most importantly:

(i) for each $g_1 \neq g_2$, $g_1, g_2 \in \{1, \ldots, G\}$, the difference $m_{g_1,h} - m_{g_2,h}$ does not depend on a value of $h \in \{1, \ldots, H\}$ and is equal to the difference between the corresponding means of the means by the first factor, i.e.,

$$m_{g_1,h} - m_{g_2,h} \;=\; \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet} \;=\; \theta_{g_1,g_2\bullet},$$

which is expressed using the parameterizations (9.30) as

$$\theta_{g_1,g_2\bullet} \;=\; \alpha^Z_{g_1} - \alpha^Z_{g_2} \;=\; \left(\boldsymbol{c}_{g_1} - \boldsymbol{c}_{g_2}\right)^\top \boldsymbol{\beta}^Z;$$

(ii) for each $h_1 \neq h_2$, $h_1, h_2 \in \{1, \ldots, H\}$, the difference $m_{g,h_1} - m_{g,h_2}$ does not depend on a value of $g \in \{1, \ldots, G\}$ and is equal to the difference between the corresponding means of the means by the second factor, i.e.,

$$m_{g,h_1} - m_{g,h_2} \;=\; \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2} \;=\; \theta_{\bullet h_1,h_2},$$

which is expressed using the parameterizations (9.30) as

$$\theta_{\bullet h_1,h_2} \;=\; \alpha^W_{h_1} - \alpha^W_{h_2} \;=\; \left(\boldsymbol{d}_{h_1} - \boldsymbol{d}_{h_2}\right)^\top \boldsymbol{\beta}^W.$$

Model of effect of Z only

It is obtained as a submodel of the additive model (9.30) by requesting

$$\alpha^W_1 = \cdots = \alpha^W_H,$$

which in the full-rank parameterization corresponds to requesting

$$\boldsymbol{\beta}^W = \mathbf{0}_{H-1}.$$

Hence the group means can be written as

$$\begin{aligned} m_{g,h} \;&=\; \alpha_0 + \alpha^Z_g, \\ &=\; \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, \qquad g = 1, \ldots, H,\ h = 1, \ldots, H. \end{aligned} \tag{9.31}$$

This is in fact a linear model for the one-way classified (by the values of the covariate Z) group means whose rank is G as soon as $n_{g\bullet} > 0$ for all $g = 1, \ldots, G$. The model of effect of Z only will symbolically be written as

$$\mathsf{M}_Z\colon\ \sim \mathsf{Z}.$$

The two-way classified group means then satisfy

(i) For each $g = 1, \ldots, G$, $m_{g,1} = \cdots = m_{g,H} = \overline{m}_{g\bullet}$.

(ii) $\overline{m}_{\bullet 1} = \cdots = \overline{m}_{\bullet H}$.

Model of effect of W only

It is the same as the model of effect of Z only with exchaged meaning of Z and W. That is, the model of effect of W only is obtained as a submodel of the additive model (9.30) by requesting

$$\alpha_1^Z = \cdots = \alpha_G^Z,$$

which in the full-rank parameterization corresponds to requesting

$$\boldsymbol{\beta}^Z = \mathbf{0}_{G-1}.$$

Hence the group means can be written as

$$\begin{aligned} m_{g,h} &= \alpha_0 + \alpha_h^W, \\ &= \beta_0 + \boldsymbol{d}_h^\top \boldsymbol{\beta}^W, \qquad g = 1, \ldots, H,\ h = 1, \ldots, H. \end{aligned} \tag{9.32}$$

The model of effect of W only will symbolically be written as

$$\mathsf{M}_W\colon\ \sim \mathsf{W}.$$

Intercept only model

This is a submodel of either the model (9.31) of effect of Z only where it is requested

$$\alpha_1^Z = \cdots = \alpha_G^Z \quad \text{or} \quad \boldsymbol{\beta}^Z = \mathbf{0}_{G-1}, \quad \text{respectively}$$

or the model (9.32) of effect of W only where it is requested

$$\alpha_1^W = \cdots = \alpha_H^W \quad \text{or} \quad \boldsymbol{\beta}^W = \mathbf{0}_{H-1}, \quad \text{respectively.}$$

Hence the group means can be written as

$$\begin{aligned} m_{g,h} &= \alpha_0, \\ &= \beta_0, \qquad g = 1, \ldots, H,\ h = 1, \ldots, H. \end{aligned}$$

As usual, this model will symbolically be denoted as

$$\mathsf{M}_0\colon\ \sim 1.$$

Summary

The models that we consider for the two-way classification are summarized by Table 9.6. The considered models form two sequence of nested submodels:

(i) $\mathsf{M}_0 \subset \mathsf{M}_Z \subset \mathsf{M}_{Z+W} \subset \mathsf{M}_{ZW}$;

(ii) $\mathsf{M}_0 \subset \mathsf{M}_W \subset \mathsf{M}_{Z+W} \subset \mathsf{M}_{ZW}$.

Related submodel testing then corresponds to evaluating whether the two-way classified group means satisfy a particular structure invoked by the submodel at hand. If normality of the error terms is assumed, the testing can be performed by the methodology of Chapter 5 (F-tests on submodels).

End of Lecture #17 (24/11/2016)

Table 9.6: Two-way ANOVA models.

Model	Rank	Requirement for Rank
$\mathsf{M}_{ZW}\colon\ \sim \mathsf{Z}+\mathsf{W}+\mathsf{Z}{:}\mathsf{W}$	$G \cdot H$	$n_{g,h} > 0$ for all $g = 1, \ldots, G$, $h = 1, \ldots, H$
$\mathsf{M}_{Z+W}\colon\ \sim \mathsf{Z}+\mathsf{W}$	$G + H - 1$	$n_{g\bullet} > 0$ for all $g = 1, \ldots, G$, $n_{\bullet h} > 0$ for all $h = 1, \ldots, H$
$\mathsf{M}_{Z}\colon\ \sim \mathsf{Z}$	G	$n_{g\bullet} > 0$ for all $g = 1, \ldots, G$
$\mathsf{M}_{W}\colon\ \sim \mathsf{W}$	H	$n_{\bullet h} > 0$ for all $h = 1, \ldots, H$
$\mathsf{M}_{0}\colon\ \sim 1$	1	$n > 0$

9.2.10 Least squares estimation

Start of Lecture #18 (30/11/2016)

Also with the two-way classification, explicit formulas for some of the LSE related quantities can be derived and then certain properties of the least squares based inference drawn.

Notation *(Sample means in two-way classification).*

$$\overline{Y}_{g,h\bullet} := \frac{1}{n_{g,h}} \sum_{j=1}^{n_{g,h}} Y_{g,h,j}, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H,$$

$$\overline{Y}_{g\bullet} := \frac{1}{n_{g\bullet}} \sum_{h=1}^{H} \sum_{j=1}^{n_{g,h}} Y_{g,h,j} = \frac{1}{n_{g\bullet}} \sum_{h=1}^{H} n_{g,h}\, \overline{Y}_{g,h\bullet}, \qquad g = 1, \ldots, G,$$

$$\overline{Y}_{\bullet h} := \frac{1}{n_{\bullet h}} \sum_{g=1}^{G} \sum_{j=1}^{n_{g,h}} Y_{g,h,j} = \frac{1}{n_{\bullet h}} \sum_{g=1}^{G} n_{g,h}\, \overline{Y}_{g,h\bullet}, \qquad h = 1, \ldots, G,$$

$$\overline{Y} := \frac{1}{n} \sum_{g=1}^{G} \sum_{h=1}^{H} \sum_{j=1}^{n_{g,h}} Y_{g,h,j} = \frac{1}{n} \sum_{g=1}^{G} n_{g\bullet}\, \overline{Y}_{g\bullet} = \frac{1}{n} \sum_{h=1}^{H} n_{\bullet h}\, \overline{Y}_{\bullet h}.$$

As usual, $\widehat{m}_{g,h}$, $g = 1, \ldots, G$, $h = 1, \ldots, H$, denote the LSE of the two-way classified group means and $\widehat{\boldsymbol{m}} = \left(\widehat{m}_{1,1}, \ldots, \widehat{m}_{G,H}\right)^{\top}$.

Theorem 9.3 Least squares estimation in two-way ANOVA linear models.

The fitted values and the LSE of the group means in two-way ANOVA linear models are given as follows (always for $g = 1, \ldots, G$, $h = 1, \ldots, H$, $j = 1, \ldots, n_{g,h}$).

(i) ***Interaction model*** $\mathsf{M}_{ZW}\colon\ \sim \mathsf{Z}+\mathsf{W}+\mathsf{Z}{:}\mathsf{W}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{g,h\bullet}.$$

(ii) ***Additive model*** M_{Z+W}: $\sim \mathsf{Z} + \mathsf{W}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y},$$

but only in case of balanced *data*[6] *($n_{g,h} = J$ for all $g = 1, \ldots, G$, $h = 1, \ldots, H$).*

(iii) ***Model of effect of*** Z ***only*** M_Z: $\sim \mathsf{Z}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{g\bullet}.$$

(iv) ***Model of effect of*** W ***only*** M_W: $\sim \mathsf{W}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{\bullet h}.$$

(v) ***Intercept only model*** M_0: ~ 1

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}.$$

Note. There exists no simple expression to calculate the fitted values in the additive model in case of unbalanced data. See Searle (1987, Section 4.9) for more details.

Proof.

Only the fitted values in the *additive* model must be derived now.

Models M_{ZW}, M_Z, M_W are, in fact, one-way ANOVA models where we already know that the fitted values are equal to the corresponding group means.

Also model M_0 is nothing new.

Fitted values in the *additive* model can be calculated by solving the normal equations corresponding to the parameterization

$$m_{g,h} = \alpha_0 + \alpha_g^Z + \alpha_h^W, \qquad g = 1, \ldots, G,\ h = 1, \ldots, H.$$

while imposing the identifying constraints

$$\sum_{g=1}^{G} \alpha_g^Z = 0, \qquad \sum_{h=1}^{H} \alpha_h^W = 0.$$

For the *additive* model with the *balanced* data ($n_{g,h} = J$ for all $g = 1, \ldots, G$, $h = 1, \ldots, H$):

- Sum of squares to be minimized

$$\mathsf{SS}(\boldsymbol{\alpha}) = \sum_g \sum_h \sum_j \left(Y_{g,h,j} - \alpha_0 - \alpha_g^Z - \alpha_h^W\right)^2.$$

- Normal equations $\equiv$ derivatives of $\mathsf{SS}(\boldsymbol{\alpha})$ divided by (-2) and set to zero:

$$\begin{aligned}
\sum_g \sum_h \sum_j Y_{g,h,j} - GHJ\alpha_0 - HJ\sum_g \alpha_g^Z - GJ\sum_h \alpha_h^W &= 0, \\
\sum_h \sum_j Y_{g,h,j} - HJ\alpha_0 - HJ\alpha_g^Z - J\sum_h \alpha_h^W &= 0, \qquad g = 1, \ldots, G, \\
\sum_g \sum_j Y_{g,h,j} - GJ\alpha_0 - J\sum_g \alpha_g^Z - GJ\alpha_h^W &= 0, \qquad h = 1, \ldots, H.
\end{aligned}$$

[6] *vyvážená data*

- After exploiting the identifying constraints:

$$\sum_g \sum_h \sum_j Y_{g,h,j} - GHJ\alpha_0 = 0,$$
$$\sum_h \sum_j Y_{g,h,j} - HJ\alpha_0 - HJ\alpha_g^Z = 0, \qquad g = 1, \ldots, G,$$
$$\sum_g \sum_j Y_{g,h,j} - GJ\alpha_0 - GJ\alpha_h^W = 0, \qquad h = 1, \ldots, H.$$

- Hence $\widehat{\alpha}_0 = \overline{Y}$,
$\widehat{\alpha}_g^Z = \overline{Y}_{g\bullet} - \overline{Y}, \qquad g = 1, \ldots, G,$
$\widehat{\alpha}_h^W = \overline{Y}_{\bullet h} - \overline{Y}, \qquad h = 1, \ldots, H.$
- And then $\widehat{m}_{g,h} = \widehat{\alpha}_0 + \widehat{\alpha}_g^Z + \widehat{\alpha}_h^W = \overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y}$,
$g = 1, \ldots, G,\ h = 1, \ldots, H.$

❑

Consequence of Theorem 9.3: LSE of the means of the means in the interaction and the additive model with balanced data.

With balanced data ($n_{g,h} = J$ for all $g = 1, \ldots, G$, $h = 1, \ldots, H$), the LSE of the means of the means by the first factor (parameters $\overline{m}_{1\bullet}, \ldots, \overline{m}_{G\bullet}$) or by the second factor (parameters $\overline{m}_{\bullet 1}, \ldots, \overline{m}_{\bullet H}$) satisfy in both the interaction and the additive two-way ANOVA linear models the following:

$$\widehat{\overline{m}}_{g\bullet} = \overline{Y}_{g\bullet}, \qquad g = 1, \ldots, G,$$
$$\widehat{\overline{m}}_{\bullet h} = \overline{Y}_{\bullet h}, \qquad h = 1, \ldots, H.$$

If additionally normality is assumed then $\widehat{\overline{\boldsymbol{m}}}^Z := \big(\widehat{\overline{m}}_{1\bullet}, \ldots, \widehat{\overline{m}}_{G\bullet}\big)^\top$ *and* $\widehat{\overline{\boldsymbol{m}}}^W := \big(\widehat{\overline{m}}_{\bullet 1}, \ldots, \widehat{\overline{m}}_{\bullet H}\big)^\top$ *satisfy*

$$\widehat{\overline{\boldsymbol{m}}}^Z \,|\, \mathbb{Z}, \mathbb{W} \sim \mathcal{N}_G\big(\overline{\boldsymbol{m}}^Z, \sigma^2 \mathbb{V}^Z\big), \qquad \widehat{\overline{\boldsymbol{m}}}^W \,|\, \mathbb{Z}, \mathbb{W} \sim \mathcal{N}_H\big(\overline{\boldsymbol{m}}^W, \sigma^2 \mathbb{V}^W\big),$$

where

$$\overline{\boldsymbol{m}}^Z = \begin{pmatrix} \overline{m}_{1\bullet} \\ \vdots \\ \overline{m}_{G\bullet} \end{pmatrix}, \quad \mathbb{V}^Z = \begin{pmatrix} \frac{1}{JH} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{JH} \end{pmatrix},$$

$$\overline{\boldsymbol{m}}^W = \begin{pmatrix} \overline{m}_{\bullet 1} \\ \vdots \\ \overline{m}_{\bullet H} \end{pmatrix}, \quad \mathbb{V}^W = \begin{pmatrix} \frac{1}{JG} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{JG} \end{pmatrix}.$$

Proof. All parameters $\overline{m}_{g\bullet}$, $g = 1, \ldots, G$, and $\overline{m}_{\bullet h}$, $h = 1, \ldots, H$ are linear combinations of the group means (of the response mean vector $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}, \mathbb{W}\big)$) and hence are estimable with the LSE being an appropriate linear combination of the LSE of the group means. With *balanced* data, we get for the the considered models (calculation shown only for LSE of $\overline{m}_{g\bullet}$, $g = 1, \ldots, G$):

(i) **Interaction model**

$$\widehat{\overline{m}}_{g\bullet} = \frac{1}{H}\sum_{h=1}^{H}\widehat{m}_{g,h} = \frac{1}{H}\sum_{h=1}^{H}\overline{Y}_{g,h\bullet} = \frac{1}{H\,J}\sum_{h=1}^{H}J\,\overline{Y}_{g,h\bullet} = \frac{1}{n_{g\bullet}}\sum_{h=1}^{H}n_{g,h}\,\overline{Y}_{g,h\bullet} = \overline{Y}_{g\bullet}.$$

(ii) **Additive model**

$$\begin{aligned}\widehat{\overline{m}}_{g\bullet} = \frac{1}{H}\sum_{h=1}^{H}\widehat{m}_{g,h} &= \frac{1}{H}\sum_{h=1}^{H}\big(\overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y}\big)\\ &= \overline{Y}_{g\bullet} + \frac{1}{H}\sum_{h=1}^{H}\overline{Y}_{\bullet h} - \overline{Y} = \overline{Y}_{g\bullet} + \frac{1}{H\,G\,J}\sum_{h=1}^{H}G\,J\overline{Y}_{\bullet h} - \overline{Y}\\ &= \overline{Y}_{g\bullet} + \underbrace{\frac{1}{n}\sum_{h=1}^{H}n_{\bullet h}\overline{Y}_{\bullet h}}_{\overline{Y}} - \overline{Y} = \overline{Y}_{g\bullet}.\end{aligned}$$

Further, $\mathbb{E}\big(\overline{Y}_{g\bullet}\,\big|\,\mathbb{Z},\,\mathbb{W}\big) = \overline{m}_{g\bullet}$ follows from properties of the LSE which are unbiased or from direct calculation. Next,

$$\mathsf{var}\big(\overline{Y}_{g\bullet}\,\big|\,\mathbb{Z},\,\mathbb{W}\big) = \mathsf{var}\left[\frac{1}{J\,H}\sum_{h=1}^{H}\sum_{j=1}^{J}Y_{g,h,j}\,\middle|\,\mathbb{Z},\,\mathbb{W}\right] = \frac{\sigma^2}{J\,H}$$

follows from the linear model assumption $\mathsf{var}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z},\,\mathbb{W}\big) = \sigma^2\,\mathbf{I}_n$.

Finally, normality of $\overline{Y}_{g\bullet}$ in case of a normal linear model, follows from the general LSE theory. ❑

9.2.11 Sums of squares and ANOVA tables with balanced data

Sums of squares

As already mentioned in Section 9.2.9, the considered models form two sequence of nested submodels:

(i) $\mathsf{M}_0 \subset \mathsf{M}_Z \subset \mathsf{M}_{Z+W} \subset \mathsf{M}_{ZW}$;

(ii) $\mathsf{M}_0 \subset \mathsf{M}_W \subset \mathsf{M}_{Z+W} \subset \mathsf{M}_{ZW}$.

Corresponding differences in the residual sums of squares (that enter the numerator of the respective F-statistic) are given as squared Euclidean norms of the fitted values from the models being compared (Section 5.1). In particular, in case of *balanced* data ($n_{g,h} = J$, $g = 1,\ldots,G$, $h = 1,\ldots,H$), we get

$$\mathsf{SS}\big(\mathsf{Z}+\mathsf{W}+\mathsf{Z}{:}\mathsf{W}\,\big|\,\mathsf{Z}+\mathsf{W}\big) = \sum_{g=1}^{G}\sum_{h=1}^{H}J\,\big(\overline{Y}_{g,h\bullet} - \overline{Y}_{g\bullet} - \overline{Y}_{\bullet h} + \overline{Y}\big)^2,$$

$$\mathsf{SS}(\mathsf{Z}+\mathsf{W}\,|\,\mathsf{W}) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet}+\overline{Y}_{\bullet h}-\overline{Y}-\overline{Y}_{\bullet h}\right)^2 = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet}-\overline{Y}\right)^2,$$

$$\mathsf{SS}(\mathsf{Z}+\mathsf{W}\,|\,\mathsf{Z}) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet}+\overline{Y}_{\bullet h}-\overline{Y}-\overline{Y}_{g\bullet}\right)^2 = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{\bullet h}-\overline{Y}\right)^2,$$

$$\mathsf{SS}(\mathsf{Z}\,|\,1) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet}-\overline{Y}\right)^2,$$

$$\mathsf{SS}(\mathsf{W}\,|\,1) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{\bullet h}-\overline{Y}\right)^2.$$

We see,

$$\mathsf{SS}(\mathsf{Z}+\mathsf{W}\,|\,\mathsf{W}) = \mathsf{SS}(\mathsf{Z}\,|\,1),$$
$$\mathsf{SS}(\mathsf{Z}+\mathsf{W}\,|\,\mathsf{Z}) = \mathsf{SS}(\mathsf{W}\,|\,1).$$

Nevertheless, note that this does not hold in case of *unbalanced* data.

Notation *(Sums of squares in two-way classification).*

In case of two-way classification and *balanced* data, we will use the following notation.

$$\mathsf{SS}_Z := \sum_{g=1}^{G}\sum_{h=1}^{H} J\big(\overline{Y}_{g\bullet}-\overline{Y}\big)^2,$$
$$\mathsf{SS}_W := \sum_{g=1}^{G}\sum_{h=1}^{H} J\big(\overline{Y}_{\bullet h}-\overline{Y}\big)^2,$$
$$\mathsf{SS}_{ZW} := \sum_{g=1}^{G}\sum_{h=1}^{H} J\big(\overline{Y}_{g,h\bullet}-\overline{Y}_{g\bullet}-\overline{Y}_{\bullet h}+\overline{Y}\big)^2,$$

$$\mathsf{SS}_T := \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{j=1}^{J}\big(Y_{g,h,j}-\overline{Y}\big)^2,$$
$$\mathsf{SS}_e^{ZW} := \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{j=1}^{J}\big(Y_{g,h,j}-\overline{Y}_{g,h\bullet}\big)^2.$$

Notes.

- Quantities SS_Z, SS_W, SS_{ZW} are differences of the residual sums of squares of two models that differ by terms Z, W or Z:W, respectively.
- Quantity SS_T is a classical total sum of squares.
- Quantity SS_e^{ZW} is a residual sum of squares from the interaction model.

Lemma 9.4 Breakdown of the total sum of squares in a balanced two-way classification.

In case of a balanced *two-way classification, the following identity holds*

$$\mathsf{SS}_T = \mathsf{SS}_Z + \mathsf{SS}_W + \mathsf{SS}_{ZW} + \mathsf{SS}_e^{ZW}.$$

Proof.

Decomposition in the lemma corresponds to the numerator sum of squares of the F-statistics when testing a series of submodels

$$\mathsf{M}_0 \subset \mathsf{M}_Z \subset \mathsf{M}_{Z+W} \subset \mathsf{M}_{ZW}$$

or a series of submodels

$$\mathsf{M}_0 \subset \mathsf{M}_W \subset \mathsf{M}_{Z+W} \subset \mathsf{M}_{ZW}.$$

Let $\mathcal{M}_0$, $\mathcal{M}_Z$, $\mathcal{M}_W$, $\mathcal{M}_{Z+W}$, $\mathcal{M}_{ZW}$ be the regression spaces of the models M_0, M_Z, M_W, M_{Z+W}, M_{ZW}, respectively.

That is, $\mathsf{SS}_T = \|\boldsymbol{U}^0\|^2$, where $\boldsymbol{U}^0$ are residuals of model M_0 and

$$\boldsymbol{U}_0 = \boldsymbol{D}_1 + \boldsymbol{D}_2 + \boldsymbol{D}_3 + \boldsymbol{U}^{ZW},$$

where $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, $\boldsymbol{D}_3$, $\boldsymbol{U}^{ZW}$ are mutually orthogonal projections of $\boldsymbol{Y}$ into subspaces of $\mathbb{R}^n$:

(i) $\boldsymbol{D}_1$: projection into $\mathcal{M}_Z \setminus \mathcal{M}_0$, $\|\boldsymbol{D}_1\|^2 = \mathsf{SS}_Z$.

(ii) $\boldsymbol{D}_2$: projection into $\mathcal{M}_{Z+W} \setminus \mathcal{M}_Z$, $\|\boldsymbol{D}_2\|^2 = \mathsf{SS}_W$.

(iii) $\boldsymbol{D}_3$: projection into $\mathcal{M}_{ZW} \setminus \mathcal{M}_{Z+W}$, $\|\boldsymbol{D}_3\|^2 = \mathsf{SS}_{ZW}$.

(iv) $\boldsymbol{U}^{ZW}$: projection into $\mathbb{R}^n \setminus \mathcal{M}_{ZW}$ (residual space of M_{ZW}).

From orthogonality: $\mathsf{SS}_T = \mathsf{SS}_Z + \mathsf{SS}_W + \mathsf{SS}_{ZW} + \mathsf{SS}_e^{ZW}$.

❑

ANOVA tables

As consequence of the above considerations, it holds for *balanced* data:

(i) Equally labeled rows in the type I ANOVA table are the same irrespective of whether the table is formed in the order Z + W + Z:W or in the order W + Z + Z:W.

(ii) Type I and type II ANOVA tables are the same.

Both type I and type II ANOVA table then take the form

Effect (Term)	Degrees of freedom	Effect sum of squares	Effect mean square	F-stat.	P-value
Z	$G-1$	SS_Z	$\star$	$\star$	$\star$
W	$H-1$	SS_W	$\star$	$\star$	$\star$
Z:W	$GH-G-H+1$	SS_{ZW}	$\star$	$\star$	$\star$
Residual	$n-GH$	SS_e^{ZW}	$\star$		

9.3 Higher-way classification

Situation of three or more, let say $p \geq 3$ factors whose influence on the response expectation is of interest could further be examined. This would lead to a linear model with p categorical covariates. Each of the covariates can be parameterized by the means of (pseudo)contrast as explained in Section 7.4. In general, higher order (up to order of p) interactions can be included in the model. Depending on included interactions, models with different interpretation with respect to the structure of higher-order classified group means are obtained. Nevertheless, more details go beyond the scope of this course. More can be learned, for example, in the *Experimental Design (NMST436)* course.

In future, possibly something brief can be included. See Seber and Lee (2003, Section 8.6).

Chapter 10

Simultaneous Inference in a Linear Model

As usual, we will assume that data are represented by a set of n random vectors $\big(Y_i,\, \boldsymbol{X}_i^\top\big)^\top$, $\boldsymbol{X}_i = \big(X_{i,0}, \ldots, X_{i,k-1}\big)^\top$, $i = 1, \ldots, n$, that satisfy a linear model. Throughout the whole chapter, *normality* will also be assumed. That is, we assume that

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big), \qquad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \leq k < n,$$

where $\boldsymbol{Y} = \big(Y_1, \ldots, Y_n\big)^\top$, $\mathbb{X}$ is a matrix with vectors $\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top$ in its rows and $\boldsymbol{\beta} = \big(\beta_0, \ldots, \beta_{k-1}\big)^\top \in \mathbb{R}^k$ and $\sigma^2 > 0$ are unknown parameters. Further, we will assume that a matrix $\mathbb{L}_{m\times k}$ ($m > 1$) with rows $\mathbf{l}_1^\top, \ldots, \mathbf{l}_m^\top$ (all non-zero vectors) is given such that

$$\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta} = \big(\mathbf{l}_1^\top\boldsymbol{\beta},\, \ldots,\, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \big(\theta_1,\, \ldots,\, \theta_m\big)^\top$$

is an estimable vector parameter of the linear model. Our interest will lie in a *simultaneous inference* on elements of the parameter $\boldsymbol{\theta}$. This means, we will be interested in

(i) deriving confidence regions for a vector parameter $\boldsymbol{\theta}$;

(ii) testing a null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ for given $\boldsymbol{\theta}^0 \in \mathbb{R}^m$.

10.1 Basic simultaneous inference

If matrix $\mathbb{L}_{m\times k}$ is such that

(i) $m \leq r$;

(ii) its rows, i.e., vectors $\mathbf{l}_1, \ldots, \mathbf{l}_m \in \mathbb{R}^k$ are linearly independent,

then we already have a tool for a simultaneous inference on $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$. It is based on point (x) of Theorem 3.2 (Least squares estimators under the normality). It provides a confidence region for $\boldsymbol{\theta}$ with a coverage of $1-\alpha$ which is

$$\left\{\boldsymbol{\theta} \in \mathbb{R}^m : \ \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^\top \left\{\mathsf{MS}_e\, \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1} \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) \ < \ m\, \mathcal{F}_{m,n-r}(1-\alpha)\right\}, \tag{10.1}$$

where $\widehat{\boldsymbol{\theta}} = \mathbb{L}\boldsymbol{b}$ is the LSE of $\boldsymbol{\theta}$. The null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ is tested using the statistic

$$Q_0 = \frac{1}{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^\top \left\{\mathsf{MS}_e\, \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-}\mathbb{L}^\top\right\}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right), \tag{10.2}$$

which under the null hypothesis follows an $\mathcal{F}_{m,n-r}$ distribution and the critical region of a test on the level of α is

$$C(\alpha) = \left[\mathcal{F}_{m,n-r}(1-\alpha),\ \infty\right). \tag{10.3}$$

The P-value if $Q_0 = q_0$ is then given as $p = 1 - \mathsf{CDF}_{\mathcal{F},\,m,\,n-r}(q_0)$. Note that the confidence region (10.1) and the test based on the statistic Q_0 and the critical region (10.3) are mutually dual. That is, the null hypothesis is rejected on a level of α if and only if $\boldsymbol{\theta}^0$ is not covered by the confidence region (10.1) with a coverage $1-\alpha$.

10.2 Multiple comparison procedures

10.2.1 Multiple testing

The null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ ($\boldsymbol{\theta}^0 = \left(\theta_1^0, \ldots, \theta_m^0\right)^\top$) on a vector parameter $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_m\right)^\top$ can also be written as $\mathrm{H}_0: \quad \theta_1 = \theta_1^0 \quad \& \quad \cdots \quad \& \quad \theta_m = \theta_m^0$.

Definition 10.1 Multiple testing problem, elementary null hypotheses, global null hypothesis.

A testing problem with the null hypothesis

$$H_0: \quad \theta_1 = \theta_1^0 \quad \& \quad \ldots \quad \& \quad \theta_m = \theta_m^0, \tag{10.4}$$

is called the multiple testing problem[1] *with the* m elementary hypotheses[2]

$$H_1: \quad \theta_1 = \theta_1^0, \quad \ldots, \quad H_m: \quad \theta_m = \theta_m^0.$$

Hypothesis H_0 *is called in this context also as a* global null hypothesis.

Note. The above definition of the multiple testing problem is a simplified definition of a general multiple testing problem where the elementary null hypotheses are not necessarily simple hypotheses. Further, general multiple testing procedures consider also problems where the null hypothesis H_0 is not necessarily given as a conjunction of the elementary hypotheses. Nevertheless, for our purposes in context of this lecture, Definition 10.1 will suffice. Also subsequent theory of multiple comparison procedures will be provided in a simplified way in an extent needed for its use in context of the multiple testing problem according to Definition 10.1 and in context of a linear model.

Notation.

- When dealing with a multiple testing problem, we will also write

$$\mathrm{H}_0 \quad \equiv \quad \mathrm{H}_1 \quad \& \quad \ldots \quad \& \quad \mathrm{H}_m$$

or

$$\mathrm{H}_0 \quad \equiv \quad \mathrm{H}_1, \quad \ldots, \quad \mathrm{H}_m$$

or

$$\mathrm{H}_0 = \bigcap_{j=1}^{m} \mathrm{H}_j.$$

- In context of a multiple testing, subscript 1 at H_1 will never indicate an alternative hypothesis. A symbol $^{\complement}$ will rather be used to indicate an alternative hypothesis.
- The alternative hypothesis of a multiple testing problem with the null hypothesis (10.4) will always be given by a complement of the parameter space under the global null hypothesis, i.e.,

$$\mathrm{H}_0^{\complement}: \ \theta_1 \neq \theta_1^0 \quad \mathrm{OR} \quad \ldots \quad \mathrm{OR} \quad \theta_m \neq \theta_m^0,$$

$$\equiv \quad \mathrm{H}_1^{\complement} \quad \mathrm{OR} \quad \ldots \quad \mathrm{OR} \quad \mathrm{H}_m^{\complement},$$

[1] *problém vícenásobného testování* [2] *elementární hypotézy*

where $\mathrm{H}_j^{\complement}: \ \theta_j \neq \theta_j^0$, $j = 1, \ldots, m$. We will also write

$$\mathrm{H}_0^{\complement} = \bigcup_{j=1}^{m} \mathrm{H}_j^{\complement}.$$

- Different ways of indexing the elementary null hypotheses will also be used (e.g., a double subscript) depending on a problem at hand.

Example 10.1 (Multiple testing problem for one-way classified group means).
Suppose that a normal linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\big)$ *is used to model dependence of the response* Y *on a single categorical covariate* Z *with a sample space* $\mathcal{Z} = \{1, \ldots, G\}$*, where the regression space* $\mathcal{M}\big(\mathbb{X}\big)$ *of a vector dimension* G *parameterizes the one-way classified group means*

$$m_1 := \mathbb{E}\big(Y \,\big|\, Z = 1\big), \ \ldots, \ m_G = \mathbb{E}\big(Y \,\big|\, Z = G\big).$$

If we restrict ourselves to full-rank parameterizations (see Section 7.4.4), the regression coefficients vector is $\boldsymbol{\beta} = \big(\beta_0, \boldsymbol{\beta^Z}^\top\big)^\top$, $\boldsymbol{\beta^Z} = \big(\beta_1, \ldots, \beta_{G-1}\big)^\top$ *and the group means are parameterized as*

$$m_g = \beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta^Z}, \qquad g = 1, \ldots, G,$$

where

$$\mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix}$$

is a chosen $G \times (G-1)$ *(pseudo)contrast matrix.*

The null hypothesis H_0*:* $m_1 = \cdots = m_G$ *on equality of the* G *group means can be specified as a multiple testing problem with* $m = \binom{G}{2}$ *elementary hypotheses (double subscript will be used to index them):*

$$H_{1,2}\colon m_1 = m_2, \quad \ldots, \quad H_{G-1,G}\colon m_{G-1} = m_G.$$

The elementary null hypotheses can now be written in terms of a vector estimable parameter

$$\boldsymbol{\theta} = \big(\theta_{1,2}, \ldots, \theta_{G-1,G}\big)^\top,$$

$$\theta_{g,h} = m_g - m_h = \big(\boldsymbol{c}_g - \boldsymbol{c}_h\big)^\top\boldsymbol{\beta^Z}, \qquad g = 1, \ldots, G-1, \ h = g+1, \ldots, G$$

as

$$H_{1,2}\colon \theta_{1,2} = 0, \quad \ldots, \quad H_{G-1,G}\colon \theta_{G-1,G} = 0,$$

or written directly in term of the group means as

$$H_{1,2}\colon m_1 - m_2 = 0, \quad \ldots, \quad H_{G-1,G}\colon m_{G-1} - m_G = 0,$$

The global null hypothesis is H_0*:* $\boldsymbol{\theta} = \mathbf{0}$*, where* $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$*. Here,* $\mathbb{L}$ *is an* $\binom{G}{2} \times G$ *matrix*

$$\mathbb{L} = \begin{pmatrix} 0 & \big(\boldsymbol{c}_1 - \boldsymbol{c}_2\big)^\top \\ \vdots & \vdots \\ 0 & \big(\boldsymbol{c}_{G-1} - \boldsymbol{c}_G\big)^\top \end{pmatrix}.$$

Since $\mathsf{rank}\big(\mathbb{C}\big) = G - 1$*, we have* $\mathsf{rank}\big(\mathbb{L}\big) = G - 1$*. We then have*

- *For $G \geq 4$, $m = \binom{G}{2} > G$. That is, in this case, the number of elementary null hypotheses is higher than the rank of the underlying linear model.*
- *For $G \geq 3$, the matrix $\mathbb{L}$ has linearly dependent rows.*

That is, for $G \geq 3$, we can

(i) neither calculate a simultaneous confidence region (10.1) for $\boldsymbol{\theta}$;

(ii) nor use the test statistic (10.2) to test H_0: $\boldsymbol{\theta} = \mathbf{0}$.

In this chapter,

(i) we develop procedures that allow to test the null hypothesis H_0: $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$ and provide a simultaneous confidence region for $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ even if the rows of the matrix $\mathbb{L}$ are linearly dependent or its rank is higher than the rank of the underlying linear model;

(ii) the test procedure will also decide which of the elementary hypotheses is/are responsible (in a certain sense) for rejection of the global null hypothesis;

(iii) developed confidence regions will have a more appealing form of a product of intervals.

10.2.2 Simultaneous confidence intervals

Suppose that a distribution of the random vector $\boldsymbol{D}$ depends on a (vector) parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top \in \Theta_1 \times \cdots \times \Theta_m = \Theta \subseteq \mathbb{R}^m$.

Definition 10.2 Simultaneous confidence intervals.
(Random) intervals (θ_j^L, θ_j^U), $j = 1, \ldots, m$, where $\theta_j^L = \theta_j^L(\boldsymbol{D})$ and $\theta_j^U = \theta_j^U(\boldsymbol{D})$, $j = 1, \ldots, m$, are called simultaneous confidence intervals[3] *for parameter $\boldsymbol{\theta}$ with a coverage of $1 - \alpha$ if for any $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_m^0)^\top \in \Theta$,*

$$\mathsf{P}\Big((\theta_1^L, \theta_1^U) \times \cdots \times (\theta_m^L, \theta_m^U) \ni \boldsymbol{\theta}^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \geq 1 - \alpha.$$

Notes.

- The condition in the definition can also be written as

$$\mathsf{P}\Big(\forall\, j = 1, \ldots, m : (\theta_j^L, \theta_j^U) \ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \geq 1 - \alpha.$$

- The product of the simultaneous confidence intervals indeed forms a confidence region in a classical sense.

End of Lecture #18 (30/11/2016)
Start of Lecture #19 (30/11/2016)

Example 10.2 (Bonferroni simultaneous confidence intervals).
Let for each $j = 1, \ldots, m$, (θ_j^L, θ_j^U) be a classical confidence interval for θ_j with a coverage of $1 - \frac{\alpha}{m}$. That is,

$$\forall\, j = 1, \ldots, m,\ \forall\, \theta_j^0 \in \Theta_j : \quad \mathsf{P}\Big((\theta_j^L, \theta_j^U) \ni \theta_j^0;\ \theta_j = \theta_j^0\Big) \geq 1 - \frac{\alpha}{m}.$$

[3] *simultánní intervaly spolehlivosti*

We then have

$$\forall\, j = 1, \ldots, m,\ \forall\, \theta_j^0 \in \Theta_j: \quad \mathsf{P}\Big(\big(\theta_j^L,\, \theta_j^U\big) \not\ni \theta_j^0;\, \theta_j = \theta_j^0\Big) \leq \frac{\alpha}{m}.$$

Further, using elementary property of a probability (for any $\boldsymbol{\theta}^0 \in \Theta$*)*

$$\mathsf{P}\Big(\exists\, j = 1, \ldots, m: \quad \big(\theta_j^L,\, \theta_j^U\big) \not\ni \theta_j^0;\, \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \;\leq\; \sum_{j=1}^{m} \mathsf{P}\Big(\big(\theta_j^L,\, \theta_j^U\big) \not\ni \theta_j^0;\, \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big)$$

$$\leq \sum_{j=1}^{m} \frac{\alpha}{m} \;=\; \alpha.$$

Hence,

$$\mathsf{P}\Big(\forall\, j = 1, \ldots, m: \quad \big(\theta_j^L,\, \theta_j^U\big) \ni \theta_j^0;\, \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \;\geq\; 1 - \alpha.$$

That is, intervals $\big(\theta_j^L,\, \theta_j^U\big)$, $j = 1, \ldots, m$, *are simultaneous confidence intervals for parameter* $\boldsymbol{\theta}$ *with a coverage of* $1 - \alpha$. *Simultaneous confidence intervals constructed in this way from univariate confidence intervals are called* Bonferroni *simultaneous confidence intervals. Their disadvantage is that they are often seriously* conservative, *i.e., having a coverage (much) higher than requested* $1 - \alpha$.

10.2.3 Multiple comparison procedure, P-values adjusted for multiple comparison

Suppose again that a distribution of the random vector $\boldsymbol{D}$ depends on a (vector) parameter $\boldsymbol{\theta} = \big(\theta_1,\, \ldots,\, \theta_m\big)^\top \in \Theta_1 \times \cdots \times \Theta_m = \Theta \subseteq \mathbb{R}^m$. Let for each $0 < \alpha < 1$ a procedure be given to construct the simultaneous confidence intervals $\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big)$, $j = 1, \ldots, m$, for parameter $\boldsymbol{\theta}$ with a coverage of $1 - \alpha$. Let for each $j = 1, \ldots, m$, the procedure creates intervals satisfying a monotonicity condition

$$1 - \alpha_1 < 1 - \alpha_2 \qquad \Longrightarrow \qquad \big(\theta_j^L(\alpha_1),\, \theta_j^U(\alpha_1)\big) \;\subseteq\; \big(\theta_j^L(\alpha_2),\, \theta_j^U(\alpha_2)\big).$$

Definition 10.3 Multiple comparison procedure.

Multiple comparison procedure (MCP)[4] *for a multiple testing problem with the elementary null hypotheses* $H_j\colon\ \theta_j = \theta_j^0$, $j = 1, \ldots, m$, *based on given procedure for construction of simultaneous confidence intervals for parameter* $\boldsymbol{\theta}$ *is the testing procedure that for given* $0 < \alpha < 1$

(i) rejects the global null hypothesis $H_0\colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$ *if and only if*

$$\big(\theta_1^L(\alpha),\, \theta_1^U(\alpha)\big) \;\times\; \cdots \;\times\; \big(\theta_m^L(\alpha),\, \theta_m^U(\alpha)\big) \not\ni \boldsymbol{\theta}^0;$$

(ii) for $j = 1, \ldots, m$, *rejects the* j*th elementary hypothesis* $H_j\ :\ \theta_j = \theta_j^0$ *if and only if*

$$\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big) \not\ni \theta_j^0.$$

[4] *procedura vícenásobného srovnávání*

Note. Since $\big(\theta_1^L(\alpha),\, \theta_1^U(\alpha)\big) \times \cdots \times \big(\theta_m^L(\alpha),\, \theta_m^U(\alpha)\big) \not\ni \boldsymbol{\theta}^0$ if and only if there exists $j = 1, \ldots, m$, such that $\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big) \not\ni \theta_j^0$, the MCP rejects, for given $0 < \alpha < 1$, the global null hypothesis $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ if and only if, it rejects at least one out of m elementary null hypotheses.

Note *(Control of the type-I error rate).*

Classical duality between confidence regions and testing procedures provides that for any $0 < \alpha < 1$, the multiple comparison procedure defines a statistical test which

(i) controls the type-I error rate with respect to the global null hypothesis $\mathrm{H}_0 \colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$, i.e.,

$$\mathsf{P}\big(\mathrm{H}_0 \text{ rejected};\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\big) \leq \alpha;$$

(ii) at the same time, for each $j = 1, \ldots, m$, controls the type-I error rate with respect to the elementary hypothesis $\mathrm{H}_j \colon\ \theta_j = \theta_j^0$, i.e.,

$$\mathsf{P}\big(\mathrm{H}_j \text{ rejected};\ \theta_j = \theta_j^0\big) \leq \alpha.$$

Definition 10.4 P-values adjusted for multiple comparison.

P-values adjusted for multiple comparison *for a multiple testing problem with the elementary null hypotheses* $H_j \colon\ \theta_j = \theta_j^0$, $j = 1, \ldots, m$, *based on given procedure for construction of simultaneous confidence intervals for parameter* $\boldsymbol{\theta}$ *are values* $p_1^{adj}, \ldots, p_m^{adj}$ *defined as*

$$p_j^{adj} = \inf\Big\{\alpha :\ \big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big) \not\ni \theta_j^0\Big\}, \qquad j = 1, \ldots, m.$$

Notes. The following is clear from construction:

- The multiple comparison procedure rejects for given $0 < \alpha < 1$ the jth elementary hypothesis $\mathrm{H}_j \colon\ \theta_j = \theta_j^0$ ($j = 1, \ldots, m$) if and only if $p_j^{adj} \leq \alpha$.
- Since the global null hypothesis $\mathrm{H}_0 \colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$ is rejected by the MCP if and only if at least one elementary hypothesis is rejected, we have that the global null hypothesis is for given α rejected if and only if

  $$\min\big\{p_1^{adj}, \ldots, p_m^{adj}\big\} \leq \alpha.$$

  That is,

  $$p^{adj} := \min\big\{p_1^{adj}, \ldots, p_m^{adj}\big\}$$

  is the P-value of a test of the global null hypothesis based on the considered MCP.

Example 10.3 (Bonferroni multiple comparison procedure, Bonferroni adjusted P-values).

Let for $0 < \alpha < 1$, $\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big)$, $j = 1, \ldots, m$, *be the confidence intervals for parameters* $\theta_1, \ldots, \theta_m$, *each with a (univariate) coverage of* $1 - \frac{\alpha}{m}$. *That is,*

$$\forall\, j = 1, \ldots, m,\ \forall\, \theta_j^0 \in \Theta_j :\quad \mathsf{P}\Big(\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big) \ni \theta_j^0;\, \theta_j = \theta_j^0\Big) \geq 1 - \frac{\alpha}{m}.$$

As shown in Example 10.2, $\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big)$, $j = 1, \ldots, m$, are the Bonferroni simultaneous confidence intervals for parameter $\boldsymbol{\theta} = \big(\theta_1,\, \ldots,\, \theta_m\big)^\top$ with a coverage of $1 - \alpha$.

Let for $j = 1, \ldots, m$, p_j^{uni} be a P-value related to the (single) test of the (jth elementary) hypothesis $H_j\colon\ \theta_j = \theta_j^0$ being dual to the confidence interval $\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big)$. That is,

$$p_j^{uni} = \inf\Big\{\frac{\alpha}{m}:\ \big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big) \not\ni \theta_j^0\Big\}.$$

Hence,

$$\min\big\{m\, p_j^{uni},\, 1\big\} = \inf\Big\{\alpha:\ \big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big) \not\ni \theta_j^0\Big\}.$$

That is, the P-values adjusted for multiple comparison based on the Bonferroni simultaneous confidence intervals are

$$p_j^B = \min\big\{m\, p_j^{uni},\, 1\big\}, \qquad j = 1, \ldots, m.$$

The related multiple comparison procedure is called the Bonferroni MCP.

Conservativeness of the Bonferroni MCP is seen, for instance, on the fact that the global null hypothesis $H_0\colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$ is rejected for given $0 < \alpha < 1$ if and only if, at least one of the elementary hypothesis is rejected by its single test on a significance level of α/m which approaches zero as m, the number of elementary hypotheses, increases.

10.2.4 Bonferroni simultaneous inference in a normal linear model

Consider a linear model

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big), \qquad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \le k < n.$$

Let

$$\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta} = \big(\mathbf{l}_1^\top\boldsymbol{\beta},\, \ldots,\, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \big(\theta_1,\, \ldots,\, \theta_m\big)^\top$$

be an estimable vector parameter of the linear model. At this point, we shall only require that $\mathbf{l}_j \neq \mathbf{0}_k$ for each $j = 1, \ldots, m$. Nevertheless, we allow for $m > r$ and also for possibly linearly dependent vectors $\mathbf{l}_1,\, \ldots,\, \mathbf{l}_m$.

As usual, let $\widehat{\boldsymbol{\theta}} = \mathbb{L}\boldsymbol{b} = \big(\mathbf{l}_1^\top\boldsymbol{b},\, \ldots,\, \mathbf{l}_m^\top\boldsymbol{b}\big)^\top = \big(\widehat{\theta}_1,\, \ldots,\, \widehat{\theta}_m\big)^\top$ be the LSE of the vector $\boldsymbol{\theta}$ and let MS_e be the residual mean square of the model.

It follows from properties of the LSE under normality that for given α, the $\Big(1 - \dfrac{\alpha}{m}\Big)\,100\%$ confidence intervals for parameters $\theta_1,\, \ldots,\, \theta_m$ have the lower and the upper bounds given as

$$\begin{aligned}
\theta_j^L(\alpha) &= \mathbf{l}_j^\top\boldsymbol{b} - \sqrt{\mathsf{MS}_e\, \mathbf{l}_j^\top\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbf{l}_j}\ \ \mathsf{t}_{n-r}\Big(1 - \frac{\alpha}{2\,m}\Big),\\
\theta_j^U(\alpha) &= \mathbf{l}_j^\top\boldsymbol{b} + \sqrt{\mathsf{MS}_e\, \mathbf{l}_j^\top\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbf{l}_j}\ \ \mathsf{t}_{n-r}\Big(1 - \frac{\alpha}{2\,m}\Big), \qquad j = 1, \ldots, m.
\end{aligned} \tag{10.5}$$

By the Bonferroni principle, intervals $\big(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\big)$, $j = 1, \ldots, m$, are simultaneous confidence intervals for parameter $\boldsymbol{\theta}$ with a coverage of $1 - \alpha$.

For each $j = 1, \ldots, m$, the confidence interval (10.5) is dual to the (single) test of the (jth elementary) hypothesis $\mathrm{H}_j\colon\ \theta_j = \theta_j^0$ based on the statistic

$$T_j(\theta_j^0) = \frac{\mathbf{l}_j^\top\boldsymbol{b} - \theta_j^0}{\sqrt{\mathsf{MS}_e\, \mathbf{l}_j^\top\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbf{l}_j}},$$

(which under the hypothesis H_j follows the Student t_{n-r} distribution) while having the critical region of the test on a level of α/m as

$$C_j = \left(-\infty,\ -\mathsf{t}_{n-r}\Big(1-\frac{\alpha}{2\,m}\Big)\right] \bigcup \left[\mathsf{t}_{n-r}\Big(1-\frac{\alpha}{2\,m}\Big),\ \infty\right).$$

The related univariate P-values are then calculated as

$$p_j^{uni} = 2\,\mathsf{CDF}_{t,\,n-r}\big(-|t_{j,0}|\big),$$

where $t_{j,0}$ is the value of the statistic $T_j(\theta_j^0)$ attained with given data. Hence the Bonferroni adjusted P-values for a multiple testing problem with the elementary null hypotheses $\mathrm{H}_j\colon\ \theta_j = \theta_j^0$, $j = 1, \ldots, m$, are

$$p_j^B = \min\Big\{2\,m\,\mathsf{CDF}_{t,\,n-r}\big(-|t_{j,0}|\big),\ 1\Big\}, \qquad j = 1, \ldots, m.$$

10.3 Tukey's T-procedure

Method presented in this section is due to John Wilder Tukey (1915 – 2000) who published the initial version of the method in 1949 (Tukey, 1949).

10.3.1 Tukey's pairwise comparisons theorem

Lemma 10.1 Studentized range.

Let $T_1, \ldots, T_m$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, $\sigma^2 > 0$. Let

$$R = \max_{j=1,\ldots,m} T_j - \min_{j=1,\ldots,m} T_j$$

be the range of the sample. Let S^2 be the estimator of σ^2 such that S^2 and $\boldsymbol{T} = \left(T_1, \ldots, T_m\right)^\top$ are independent and

$$\frac{\nu\, S^2}{\sigma^2} \sim \chi^2_\nu \quad \text{for some} \quad \nu > 0.$$

Let

$$Q = \frac{R}{S}.$$

The distribution of the random variable Q then depends on neither μ, nor σ.

Proof.

- We can write:

$$\frac{R}{S} = \frac{\frac{1}{\sigma}\left\{\max_j(T_j - \mu) - \min_j(T_j - \mu)\right\}}{\frac{S}{\sigma}} = \frac{\max_j\left(\frac{T_j - \mu}{\sigma}\right) - \min_j\left(\frac{T_j - \mu}{\sigma}\right)}{\frac{S}{\sigma}}.$$

- Distribution of both the numerator and the denominator depends on neither μ, nor σ since
  - For all $j = 1, \ldots, m$ $\quad \frac{T_j - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.
  - Distribution of $\frac{S}{\sigma}$ is a transformation of the χ^2_ν distribution.

❑

Note. The distribution of the random variable $Q = \frac{R}{S}$ from Lemma 10.1 still depends on m (the sample size of $\boldsymbol{T}$) and ν (degrees of freedom of the χ^2 distribution related to the variance estimator S^2).

Definition 10.5 Studentized range.

The random variable $Q = \dfrac{R}{S}$ *from Lemma 10.1 will be called* studentized range[5] *of a sample of size* m *with* ν *degrees of freedom and its distribution will be denoted as* $\mathsf{q}_{m,\nu}$.

Notation.

- For $0 < p < 1$, the $p\,100\%$ quantile of the random variable Q with distribution $\mathsf{q}_{m,\nu}$ will be denoted as $\mathsf{q}_{m,\nu}(p)$.
- The distribution function of the random variable Q with distribution $\mathsf{q}_{m,\nu}$ will be denoted $\mathsf{CDF}_{\mathsf{q},m,\nu}(\cdot)$.

Theorem 10.2 Tukey's pairwise comparisons theorem, balanced version.

Let $T_1, \ldots, T_m$ *be* independent *random variables and let* $T_j \sim \mathcal{N}(\mu_j,\, v^2\sigma^2)$, $j = 1, \ldots, m$, *where* $v^2 > 0$ *is a* known *constant. Let* S^2 *be the estimator of* σ^2 *such that* S^2 *and* $\boldsymbol{T} = \big(T_1, \ldots, T_m\big)^\top$ *are* independent *and*

$$\frac{\nu\, S^2}{\sigma^2} \sim \chi^2_\nu \quad \textit{for some} \quad \nu > 0.$$

Then

$$\mathsf{P}\Big(\textit{for all } j \neq l\text{:}\quad \big|T_j - T_l \;-\; (\mu_j - \mu_l)\big| \;<\; \mathsf{q}_{m,\nu}(1-\alpha)\,\sqrt{v^2\; S^2}\Big) \;=\; 1 - \alpha.$$

Proof.

- It follows from the assumptions that random variables $\dfrac{T_j - \mu_j}{v}$, $j = 1, \ldots, m$, are i.i.d. with the distribution $\mathcal{N}(0,\, \sigma^2)$.
- Let $R = \max\limits_j\left(\dfrac{T_j - \mu_j}{v}\right) \;-\; \min\limits_j\left(\dfrac{T_j - \mu_j}{v}\right)$.
  $\Rightarrow \dfrac{R}{S} \sim q_{m,\,\nu}$.
- Hence for any $0 < \alpha < 1$ ($\mathsf{q}_{m,\nu}$ is a continuous distribution):

$$\begin{aligned}
1 - \alpha &= \mathsf{P}\left(\frac{\max\limits_j\left(\dfrac{T_j - \mu_j}{v}\right) \;-\; \min\limits_j\left(\dfrac{T_j - \mu_j}{v}\right)}{S} \;<\; q_{m,\nu}(1-\alpha)\right) \\
&= \mathsf{P}\left(\frac{\max\limits_j(T_j - \mu_j) \;-\; \min\limits_j(T_j - \mu_j)}{v\, S} \;<\; q_{m,\nu}(1-\alpha)\right) \\
&= \mathsf{P}\Big(\max_j(T_j - \mu_j) \;-\; \min_j(T_j - \mu_j) \;<\; v\, S\, q_{m,\nu}(1-\alpha)\Big)
\end{aligned}$$

[5] *studentizované rozpětí*

$$= \mathsf{P}\Big(\text{for all } j \neq l \quad \big|(T_j - \mu_j) - (T_l - \mu_l)\big| < v\, S\, q_{m,\nu}(1-\alpha)\Big)$$
$$= \mathsf{P}\Big(\text{for all } j \neq l \quad \big|T_j - T_l - (\mu_j - \mu_l)\big| < q_{m,\nu}(1-\alpha)\sqrt{v^2\, S^2}\Big).$$

❑

Theorem 10.3 Tukey's pairwise comparisons theorem, general version.

Let $T_1, \ldots, T_m$ *be* independent *random variables and let* $T_j \sim \mathcal{N}(\mu_j,\, v_j^2\sigma^2)$, $j = 1, \ldots, m$, *where* $v_j^2 > 0$, $j = 1, \ldots, m$ *are* known *constants. Let* S^2 *be the estimator of* σ^2 *such that* S^2 *and* $\boldsymbol{T} = \big(T_1, \ldots, T_m\big)^\top$ *are* independent *and*

$$\frac{\nu\, S^2}{\sigma^2} \sim \chi^2_\nu \quad \textit{for some} \quad \nu > 0.$$

Then

$$\mathsf{P}\left(\textit{for all } j \neq l \quad \big|T_j - T_l - (\mu_j - \mu_l)\big| < \mathsf{q}_{m,\nu}(1-\alpha)\sqrt{\frac{v_j^2 + v_l^2}{2}\, S^2}\right) \quad \geq \quad 1 - \alpha.$$

Proof. **Proof/calculations were skipped and are not requested for the exam.**

See Hayter (1984).

❑

Notes.

- Tukey suggested that statement of Theorem 10.3 holds already in 1953 (in an unpublished manuscript Tukey, 1953) without proving it. Independently, it was also suggested by Kramer (1956). Consequently, the statement of Theorem 10.3 was called as Tukey–Kramer conjecture.
- The proof is not an easy adaptation of the proof of the balanced version.

10.3.2 Tukey's honest significance differences (HSD)

A method of multiple comparison that will now be developed appears under several different names in the literature: Tukey's method, Tukey–Kramer method, Tukey's range test, Tukey's honest significance differences (HSD) test.

Assumptions.

In the following, we assume that

$$\boldsymbol{T} = \big(T_1, \ldots, T_m\big)^\top \sim \mathcal{N}_m(\boldsymbol{\mu},\, \sigma^2\, \mathbb{V}),$$

where

- $\boldsymbol{\mu} = \big(\mu_1, \ldots, \mu_m\big)^\top \in \mathbb{R}^m$ and $\sigma^2 > 0$ are unknown parameters;
- $\mathbb{V}$ is a *known diagonal* matrix with $v_1^2, \ldots, v_m^2$ on a diagonal.

That is, $T_1, \ldots, T_m$ are *independent* and $T_j \sim \mathcal{N}(\mu_j, \sigma^2 v_j)$, $j = 1, \ldots, m$. Further, we will assume that an estimator S^2 of σ^2 is available which is independent of $\boldsymbol{T}$ and which satisfies $\nu S^2/\sigma^2 \sim \chi^2_\nu$ for some $\nu > 0$.

Multiple comparison problem.

A multiple comparison procedure that will be developed aims in testing $m^\star = \binom{m}{2}$ elementary hypotheses on all pairwise differences between the means $\mu_1, \ldots, \mu_m$. Let

$$\theta_{j,l} = \mu_j - \mu_l, \qquad j = 1, \ldots, m-1,\ l = j+1, \ldots, m,$$
$$\boldsymbol{\theta} = \big(\theta_{1,2}, \theta_{1,3}, \ldots, \theta_{m-1,m}\big)^\top.$$

The elementary hypotheses of a multiple testing problem that we shall consider are

$$\mathrm{H}_{j,l}\colon \ \theta_{j,l}(= \mu_j - \mu_l) = \theta^0_{j,l}, \qquad j = 1, \ldots, m-1,\ l = j+1, \ldots, m,$$

for some $\boldsymbol{\theta}^0 = \big(\theta^0_{1,2}, \theta^0_{1,3}, \ldots, \theta^0_{m-1,m}\big)^\top \in \mathbb{R}^{m^\star}$. The global null hypothesis is as usual $\mathrm{H}_0\colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$.

Note. The most common multiple testing problem in this context is with $\boldsymbol{\theta}^0 = \mathbf{0}_{m^\star}$ which corresponds to all pairwise comparisons of the means $\mu_1, \ldots, \mu_m$. The global null hypothesis then states that all the means are equal.

Some derivations

Using either of the Tukey's pairwise comparison theorems (Theorems 10.2 and 10.3), we have (for chosen $0 < \alpha < 1$):

$$\mathsf{P}\left(\text{for all } j \neq l \quad \big|T_j - T_l - (\mu_j - \mu_l)\big| < \mathsf{q}_{m,\nu}(1-\alpha)\sqrt{\frac{v_j^2 + v_l^2}{2}\, S^2}\right) \ \geq\ 1 - \alpha,$$

with equality of the above probability to $1-\alpha$ in the balanced case of $v_1^2 = \cdots = v_m^2$. That is, we have,

$$\mathsf{P}\left(\text{for all } j \neq l \quad \left|\frac{T_j - T_l - (\mu_j - \mu_l)}{\sqrt{\frac{v_j^2+v_l^2}{2}\, S^2}}\right| < \mathsf{q}_{m,\nu}(1-\alpha)\right) \ \geq\ 1 - \alpha.$$

Let for $j \neq l$ and for $\theta^0_{j,l} \in \mathbb{R}$

$$T_{j,l}(\theta^0_{j,l}) := \frac{T_j - T_l - \theta^0_{j,l}}{\sqrt{\frac{v_j^2 + v_l^2}{2}\, S^2}}.$$

That is

$$1-\alpha \;\le\; \mathsf{P}\Bigg(\text{for all } j\neq l \quad \big|T_{j,l}(\theta^0_{j,l})\big| < \mathsf{q}_{m,\nu}(1-\alpha);\; \boldsymbol{\theta}=\boldsymbol{\theta}^0\Bigg) \tag{10.6}$$

$$= \mathsf{P}\left(\text{for all } j\neq l \quad \left|\frac{T_j - T_l - \theta^0_{j,l}}{\sqrt{\frac{v_j^2+v_l^2}{2}\,S^2}}\right| < \mathsf{q}_{m,\nu}(1-\alpha);\; \boldsymbol{\theta}=\boldsymbol{\theta}^0\right)$$

$$= \mathsf{P}\Big(\text{for all } j\neq l \quad \big(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\big) \ni \theta^0_{j,l};\; \boldsymbol{\theta}=\boldsymbol{\theta}^0\Big), \tag{10.7}$$

where

$$\begin{aligned} \theta^{TL}_{j,l}(\alpha) &= T_j - T_l - \mathsf{q}_{m,\nu}(1-\alpha)\sqrt{\tfrac{v_j^2+v_l^2}{2}\,S^2}, \\ \theta^{TU}_{j,l}(\alpha) &= T_j - T_l + \mathsf{q}_{m,\nu}(1-\alpha)\sqrt{\tfrac{v_j^2+v_l^2}{2}\,S^2}, \qquad j<l. \end{aligned} \tag{10.8}$$

Theorem 10.4 Tukey's honest significance differences.

Random intervals given by (10.8) are simultaneous confidence intervals for parameters $\theta_{j,l}=\mu_j-\mu_l$, $j=1,\ldots,m-1$, $l=j+1,\ldots,m$ with a coverage of $1-\alpha$.

In the balanced case of $v_1^2=\cdots=v_m^2$, the coverage is exactly equal to $1-\alpha$, i.e., for any $\boldsymbol{\theta}^0\in\mathbb{R}^{m^\star}$

$$\mathsf{P}\Big(\textit{for all } j\neq l \quad \big(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\big) \ni \theta^0_{j,l};\; \boldsymbol{\theta}=\boldsymbol{\theta}^0\Big) \;=\; 1-\alpha.$$

Related P-values for a multiple testing problem with elementary hypotheses $H_{j,l}\colon\ \theta_{j,l}=\theta^0_{j,l}$, $\theta^0_{j,l}\in\mathbb{R}$, $j<l$, adjusted for multiple comparison are given by

$$p^T_{j,l} = 1 \;-\; \mathsf{CDF}_{\mathsf{q},m,\nu}\Big(\big|t^0_{j,l}\big|\Big), \qquad j<l,$$

where $t^0_{j,l}$ is a value of $T_{j,l}(\theta^0_{j,l}) = \dfrac{T_j-T_l-\theta^0_{j,l}}{\sqrt{\frac{v_j^2+v_l^2}{2}\,S^2}}$ attained with given data.

Proof.

The fact that $\big(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\big)$, $j<l$, are simultaneous confidence intervals for parameters $\theta_{j,l}=\mu_j-\mu_l$ with a coverage of $1-\alpha$ follows from (10.7).

The fact that the coverage of the simultaneous confidence intervals is exactly equal to $1-\alpha$ in a balanced case follows from the fact that inequality in (10.6) is equality in a balanced case.

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with the elementary hypotheses $\mathrm{H}_{j,l}\colon\ \theta_{j,l}=\theta^0_{j,l}$, $j<l$, follows from noting the following (for each $j<l$):

$$\big(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\big) \not\ni \theta^0_{j,l} \quad\Longleftrightarrow\quad \Big|T_{j,l}\big(\theta^0_{j,l}\big)\Big| \ge \mathsf{q}_{m,\,\nu}(1-\alpha)$$

It now follows from monotonicity of the quantiles of a continuous Studentized range distribution that

$$p^T_{j,l} = \inf\Big\{\alpha:\ \big(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\big) \not\ni \theta^0_{j,l}\Big\} = \inf\Big\{\alpha:\ \Big|T_{j,l}\big(\theta^0_{j,l}\big)\Big| \ge \mathsf{q}_{m,\,\nu}(1-\alpha)\Big\}$$

is attained for $p_{j,l}^T$ satisfying

$$\left|T_{j,l}(\theta_{j,l}^0)\right| = \mathsf{q}_{m,\nu}(1-p_{j,l}^T).$$

That is, if $t_{j,l}^0$ is a value of the statistic $T_{j,l}(\theta_{j,l}^0)$ attained with given data, we have

$$p_{j,l}^T = 1 - \mathsf{CDF}_{\mathsf{q},m,\nu}\left(\left|t_{j,l}^0\right|\right).$$

❑

End of Lecture #19 (30/11/2016)

Start of Lecture #20 (01/12/2016)

10.3.3 Tukey's HSD in a linear model

In context of a normal linear model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k < n$, the Tukey's honest significance differences are applicable in the following situation.

- $\mathbb{L}_{m\times k}$ is a matrix with non-zero rows $\mathbf{l}_1^\top, \ldots, \mathbf{l}_m^\top$ such that the parameter

  $$\boldsymbol{\eta} = \mathbb{L}\boldsymbol{\beta} = \left(\mathbf{l}_1^\top\boldsymbol{\beta}, \ldots, \mathbf{l}_m^\top\boldsymbol{\beta}\right)^\top = \left(\eta_1, \ldots, \eta_m\right)^\top$$

  is estimable.

- Matrix $\mathbb{L}$ is such that

  $$\mathbb{V} := \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^- \mathbb{L}^\top = \left(v_{j,l}\right)_{j,l=1,\ldots,m}$$

  is a diagonal matrix with $v_j^2 := v_{j,j}$, $j = 1, \ldots, m$.

With $\boldsymbol{b} = \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{X}^\top\boldsymbol{Y}$ and the residual mean square MS_e of the fitted linear model, we have (conditionally, given the model matrix $\mathbb{X}$):

$$\boldsymbol{T} := \widehat{\boldsymbol{\eta}} = \left(\mathbf{l}_1^\top\boldsymbol{b}, \ldots, \mathbf{l}_m^\top\boldsymbol{b}\right)^\top = \mathbb{L}\boldsymbol{b} \sim \mathcal{N}_m(\boldsymbol{\eta}, \sigma^2\mathbb{V}), \qquad \frac{(n-r)\mathsf{MS}_e}{\sigma^2} \sim \chi^2_{n-r},$$

$$\widehat{\boldsymbol{\eta}} \text{ and } \mathsf{MS}_e \text{ independent.}$$

Hence the Tukey's T-procedure can be used for a multiple comparison problem on (also estimable) parameters

$$\theta_{j,l} = \eta_j - \eta_l = \left(\mathbf{l}_j - \mathbf{l}_l\right)^\top\boldsymbol{\beta}, \qquad j < l.$$

The Tukey's simultaneous confidence intervals for parameters $\theta_{j,l}$, $j < l$, with a coverage of $1-\alpha$ have then the lower and the upper bound given as

$$\begin{aligned}
\theta_{j,l}^{TL}(\alpha) &= \widehat{\eta}_j - \widehat{\eta}_l - \mathsf{q}_{m,n-r}(1-\alpha)\sqrt{\tfrac{v_j^2+v_l^2}{2}\,\mathsf{MS}_e}, \\
\theta_{j,l}^{TU}(\alpha) &= \widehat{\eta}_j - \widehat{\eta}_l + \mathsf{q}_{m,n-r}(1-\alpha)\sqrt{\tfrac{v_j^2+v_l^2}{2}\,\mathsf{MS}_e}, \qquad j < l.
\end{aligned}$$

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with elementary hypotheses

$$\mathrm{H}_{j,l}\colon\ \theta_{j,l} = \theta_{j,l}^0, \qquad j < l,$$

for chosen $\theta_{j,l}^0 \in \mathbb{R}$, is based on statistics

$$T_{j,l}(\theta_{j,l}^0) = \frac{\widehat{\eta}_j - \widehat{\eta}_l - \theta_{j,l}^0}{\sqrt{\dfrac{v_j^2+v_l^2}{2}\,\mathsf{MS}_e}}, \qquad j < l.$$

The above procedure is in particular applicable if all involved covariates are *categorical* and the model corresponds to one-way, two-way or higher-way classification. If normal and homoscedastic errors in the underlying linear model are assumed, the Tukey's HSD method can then be used to develop a multiple comparison procedure for differences between the group means or between the means of the group means.

One-way classification

Let $\boldsymbol{Y} = \big(Y_{1,1}, \ldots, Y_{G,n_G}\big)^\top$, $n = \sum_{g=1}^{G} n_g$, and

$$\begin{aligned} Y_{g,j} &\sim \mathcal{N}(m_g,\ \sigma^2), \\ Y_{g,j} &\text{ independent for } g = 1, \ldots, G,\ j = 1, \ldots, n_g, \end{aligned}$$

We then have (see Theorem 9.1, with random covariates conditionally given the covariate values)

$$\boldsymbol{T} := \begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_G \end{pmatrix} \sim \mathcal{N}_G \left(\begin{pmatrix} m_1 \\ \vdots \\ m_G \end{pmatrix}, \quad \sigma^2 \begin{pmatrix} \frac{1}{n_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{n_G} \end{pmatrix} \right).$$

Moreover, the mean square error MS_e of the underlying one-way ANOVA linear model satisfies, with $\nu_e = n - G$,

$$\frac{\nu_e\, \mathsf{MS}_e}{\sigma^2} \sim \chi^2_{\nu_e}, \qquad \mathsf{MS}_e \text{ and } \boldsymbol{T} \text{ independent}$$

(due to the fact that $\boldsymbol{T}$ is the LSE of the vector of group means $\boldsymbol{m} = \big(m_1, \ldots, m_G\big)^\top$). Hence the Tukey's simultaneous confidence intervals for $\theta_{g,h} = m_g - m_h$, $g = 1, \ldots, G-1$, $h = g+1, \ldots, G$ with a coverage of $1 - \alpha$, have then the lower and upper bounds given as

$$\overline{Y}_g - \overline{Y}_h \ \pm\ \mathsf{q}_{G,\,n-G}(1-\alpha) \sqrt{\frac{1}{2}\left(\frac{1}{n_g} + \frac{1}{n_h}\right) \mathsf{MS}_e}, \qquad g < h.$$

In case of a balanced data ($n_1 = \cdots = n_G$), the coverage of those intervals is even exactly equal to $1 - \alpha$, otherwise, the intervals are conservative (having a coverage greater than $1 - \alpha$).

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with elementary hypotheses

$$\mathrm{H}_{g,h}\colon\ \theta_{g,h} = \theta^0_{g,h}, \qquad g < h,$$

for chosen $\theta^0_{g,h} \in \mathbb{R}$, is based on statistics

$$T_{g,h}(\theta^0_{g,h}) = \frac{\overline{Y}_g - \overline{Y}_h - \theta^0_{g,h}}{\sqrt{\frac{1}{2}\left(\frac{1}{n_g} + \frac{1}{n_h}\right) \mathsf{MS}_e}}, \qquad g < h.$$

Note. The R function `TukeyHSD` applied to objects obtained using the function `aov` (performs LSE based inference for linear models involving only categorical covariates) provides a software implementation of the Tukey's T multiple comparison described here.

Two-way classification

Let $\boldsymbol{Y} = \big(Y_{1,1,1},\, \ldots,\, Y_{G,H,n_{G,H}}\big)^\top$, $n = \sum_{g=1}^{G} \sum_{h=1}^{H} n_{g,h}$, and

$$Y_{g,h,j} \;\sim\; \mathcal{N}(m_{g,h},\, \sigma^2),$$

$$Y_{g,h,j} \text{ independent for } g = 1, \ldots, G,\; h = 1, \ldots, H,\; j = 1, \ldots, n_{g,h},$$

Let, as usual,

$$n_{g\bullet} = \sum_{h=1}^{H} n_{g,h}, \qquad \overline{Y}_{g\bullet} = \frac{1}{n_{g\bullet}} \sum_{h=1}^{H} \sum_{j=1}^{n_{g,h}} Y_{g,h,j},$$

$$\overline{m}_{g\bullet} = \frac{1}{H} \sum_{h=1}^{H} m_{g,h}, \qquad \overline{m}^{wt}_{g\bullet} = \frac{1}{n_{g\bullet}} \sum_{h=1}^{H} n_{g,h} m_{g,h}, \qquad g = 1, \ldots, G.$$

Balanced data

In case of *balanced data* ($n_{g,h} = J$ for all g, h), we have $n_{g\bullet} = J\,H$, $\overline{m}^{wt}_{g} = \overline{m}_{g}$. Further,

$$\boldsymbol{T} := \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{G\bullet} \end{pmatrix} \sim \mathcal{N}_G \left(\begin{pmatrix} \overline{m}_{1\bullet} \\ \vdots \\ \overline{m}_{G\bullet} \end{pmatrix},\; \sigma^2 \begin{pmatrix} \frac{1}{J\,H} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{J\,H} \end{pmatrix} \right),$$

see Consequence of Theorem 9.3. Further, let MS_e^{ZW} and MS_e^{Z+W} be the residual mean squares from the interaction model and the additive model, respectively, $\nu_e^{ZW} = n - G\,H$, and $\nu_e^{Z+W} = n - G - H + 1$ degrees of freedom, respectively. We have shown in the proof of Consequence of Theorem 9.3 that for both the interaction model and the additive model, the sample means $\overline{Y}_{1\bullet}, \ldots, \overline{Y}_{G\bullet}$ are LSE's of estimable parameters $\overline{m}_{1\bullet}, \ldots, \overline{m}_{G\bullet}$ and hence, for both models, vector $\boldsymbol{T}$ is *independent* of the corresponding residual mean square. Further, depending on whether the interaction model or the additive model is assumed, we have

$$\frac{\nu_e^\star\, \mathsf{MS}_e^\star}{\sigma^2} \sim \chi^2_{\nu_e^\star},$$

where $\mathsf{MS}_e^\star$ is the residual mean square of the model that is assumed (MS_e^{ZW} or MS_e^{Z+W}) and $\nu_e^\star$ the corresponding degrees of freedom (ν_e^{ZW} or ν_e^{Z+W}). Hence the Tukey's simultaneous confidence intervals for $\theta_{g_1,g_2} = \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet}$, $g_1 = 1, \ldots, G-1$, $g_2 = g_1 + 1, \ldots, G$ have then the lower and upper bounds given as

$$\overline{Y}_{g_1\bullet} - \overline{Y}_{g_2\bullet} \;\pm\; \mathsf{q}_{G,\nu_e^\star}(1-\alpha) \sqrt{\frac{1}{J\,H}\, \mathsf{MS}_e^\star},$$

and the coverage of those intervals is even exactly equal to $1 - \alpha$.

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with elementary hypotheses

$$\mathrm{H}_{g_1,g_2}\colon\; \theta_{g_1,g_2} = \theta^0_{g_1,g_2}, \qquad g_1 < g_2,$$

for chosen $\theta^0_{g_1,g_2} \in \mathbb{R}$, is based on statistics

$$T_{g_1,g_2}(\theta^0_{g_1,g_2}) = \frac{\overline{Y}_{g_1\bullet} - \overline{Y}_{g_2\bullet} - \theta^0_{g_1,g_2}}{\sqrt{\frac{1}{J\,H}\, \mathsf{MS}_e^\star}}, \qquad g_1 < g_2.$$

Unbalanced data

Beginning of skipped part

With unbalanced data, direct calculation shows that

$$\boldsymbol{T} := \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{G\bullet} \end{pmatrix} \sim \mathcal{N}_G \left(\begin{pmatrix} \overline{m}^{wt}_{1\bullet} \\ \vdots \\ \overline{m}^{wt}_{G\bullet} \end{pmatrix}, \ \sigma^2 \begin{pmatrix} \frac{1}{n_{1\bullet}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{n_{G\bullet}} \end{pmatrix} \right).$$

Further, the sample means $\overline{Y}_{1\bullet}, \dots, \overline{Y}_{G\bullet}$ are LSE's of the estimable parameters $\overline{m}^{wt}_{1\bullet}$, ..., $\overline{m}^{wt}_{G\bullet}$ in both the interaction and the additive model. This is obvious for the interaction model since there we know the fitted values ($\equiv$ LSE's of the group means $m_{g,h}$). Those are $\widehat{Y}_{g,h,j} = \overline{Y}_{g,h\bullet}$, $g = 1, \dots, G$, $h = 1, \dots, H$, $j = 1, \dots, n_{g,h}$ (Theorem 9.3). Hence the sample means $\overline{Y}_{1\bullet}$, ..., $\overline{Y}_{G\bullet}$, which are their linear combinations, are LSE's of the corresponding linear combinations of the group means $m_{g,h}$. Those are the weighted means of the means $\overline{m}^{wt}_{1\bullet}$, ..., $\overline{m}^{wt}_{G\bullet}$. To show that the sample means $\overline{Y}_{1\bullet}$, ..., $\overline{Y}_{G\bullet}$ are the LSE's for the estimable parameters $\overline{m}^{wt}_{1\bullet}$, ..., $\overline{m}^{wt}_{G\bullet}$ in the additive model would, nevertheless, require additional derivations.

For the rest, we can proceed in the same way as in the balanced case. That is, let $\mathsf{MS}^\star_e$ and $\nu^\star_e$ denote the residual mean square and the residual degrees of freedom of the model that can be assumed (interaction or additive). Owing to the fact that $\boldsymbol{T}$ is a vector of the LSE's of the estimable parameters for both models, it is independent of $\mathsf{MS}^\star_e$. The Tukey's T multiple comparison procedure is now applicable for inference on parameters

$$\theta^{wt}_{g_1,g_2} = \overline{m}^{wt}_{g_1\bullet} - \overline{m}^{wt}_{g_2\bullet}, \qquad g_1 = 1, \dots, G-1, \ \ g_2 = g_1+1, \dots, G.$$

The Tukey's simultaneous confidence intervals for $\theta^{wt}_{g_1,g_2} = \overline{m}^{wt}_{g_1\bullet} - \overline{m}^{wt}_{g_2\bullet}$, $g_1 = 1, \dots, G-1$, $g_2 = g_1+1, \dots, G$, with a coverage of $1-\alpha$, have the lower and upper bounds given as

$$\overline{Y}_{g_1\bullet} - \overline{Y}_{g_2\bullet} \ \pm \ \mathsf{q}_{G,\nu^\star_e}(1-\alpha)\sqrt{\frac{1}{2}\left(\frac{1}{n_{g_1\bullet}} + \frac{1}{n_{g_2\bullet}}\right)\mathsf{MS}^\star_e}.$$

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with elementary hypotheses

$$\mathrm{H}_{g_1,g_2}\colon \ \theta^{wt}_{g_1,g_2} = \theta^{wt,0}_{g_1,g_2}, \qquad g_1 < g_2,$$

for chosen $\theta^{wt,0}_{g_1,g_2} \in \mathbb{R}$, is based on statistics

$$T_{g_1,g_2}(\theta^{wt,0}_{g_1,g_2}) = \frac{\overline{Y}_{g_1\bullet} - \overline{Y}_{g_2\bullet} - \theta^{wt,0}_{g_1,g_2}}{\sqrt{\frac{1}{2}\left(\frac{1}{n_{g_1\bullet}} + \frac{1}{n_{g_2\bullet}}\right)\mathsf{MS}^\star_e}}, \qquad g_1 < g_2.$$

Notes.

- Analogous procedure applies for the inference on the means of the means

$$\overline{m}_{\bullet h} = \frac{1}{G}\sum_{g=1}^{G} m_{g,h}, \qquad \overline{m}^{wt}_{\bullet h} = \frac{1}{n_{\bullet h}}\sum_{g=1}^{G} n_{g,h} m_{g,h}, \qquad h = 1, \dots, H,$$

  by the second factor of the two-way classification.

- The weighted means of the means $\overline{m}^{wt}_{g\bullet}$ or $\overline{m}^{wt}_{\bullet h}$ have a reasonable interpretation only in certain special situations. If this is not the case, the Tukey's multiple comparison with unbalanced data does not make much sense.

- Even with *unbalanced* data, we can, of course, calculate the LSE's of the (unweighted) means of the means $\overline{m}_{g\bullet}$ or $\overline{m}_{\bullet h}$. Nevertheless, those LSE's are correlated with unbalanced data and hence we cannot apply the Tukey's procedure.

Note *(Tukey's HSD in the* R *software).*

The R function `TukeyHSD` provides the Tukey's T-procedure also for the two-way classification (for both the additive and the interaction model). For balanced data, it performs a simultaneous inference on parameters $\theta_{g_1,g_2} = \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet}$ (and analogous parameters with respect to the second factor) in a way described here. For unbalanced data, it performs a simultaneous inference on parameters $\theta^{wt}_{g_1,g_2} = \overline{m}^{wt}_{g_1\bullet} - \overline{m}^{wt}_{g_2\bullet}$ as described here, nevertheless, only for the first factor mentioned in the model formula. Inference on different parameters is provided with respect to the second factor in the model formula. That is, with unbalanced data, output from the R function `TukeyHSD` and interpretation of the results depend on the order of the factors in the model formula.

`TukeyHSD` with two-way classification for the second factor uses "new" observations that adjust for the effect of the first factor. That is, it is worked with "new" observations $Y^{\star}_{g,h,j}$, given as

$$Y^{\star}_{g,h,j} = Y_{g,h,j} - \overline{Y}_{g\bullet} + \overline{Y}, \qquad g = 1,\dots,G,\ h = 1,\dots,H,\ j = 1,\dots,n_{g,h}.$$

The Tukey's T procedure is then applied to the sample means

$$\overline{Y}^{\star}_{\bullet h} \ = \ \overline{Y}_{\bullet h} \ - \ \frac{1}{n_{\bullet h}} \sum_{g=1}^{G} n_{g,h}\, \overline{Y}_{g\bullet} \ + \ \overline{Y}, \qquad h = 1,\dots,H,$$

whose expectations are

$$\overline{m}^{wt}_{\bullet h} \ - \ \frac{1}{n_{\bullet h}} \sum_{g=1}^{G} n_{g,h} \overline{m}^{wt}_{g\bullet} \ + \ \frac{1}{n} \sum_{g=1}^{G} \sum_{h_2=1}^{H} n_{g,h_2} m_{g,h_2}, \qquad h = 1,\dots,H,$$

which, with unbalanced data, are *not* equal to $\overline{m}^{wt}_{\bullet h}$.

End of skipped part

10.4 Hothorn-Bretz-Westfall procedure

The multiple comparison procedure presented in this section is applicable for any parametric model where the parameters estimators follow either exactly (as in the case of a normal linear model) or at least asymptotically a (multivariate) normal or t-distribution. In full generality, it was published only rather recently (Hothorn et al., 2008, 2011), nevertheless, the principal ideas behind the method are some decades older.

10.4.1 Max-abs-t distribution

Definition 10.6 Max-abs-t-distribution.
Let $\boldsymbol{T} = \big(T_1, \ldots, T_m\big)^\top \sim \mathsf{mvt}_{m,\nu}\big(\boldsymbol{\Sigma}\big)$, *where* $\boldsymbol{\Sigma}$ *is a positive* semidefinite *matrix. The distribution of a random variable*

$$H = \max_{j=1,\ldots,m} |T_j|$$

will be called the max-abs-t-distribution of dimension m *with* ν *degrees of freedom and a scale matrix* $\boldsymbol{\Sigma}$ *and will be denoted as* $\mathsf{h}_{m,\nu}(\boldsymbol{\Sigma})$.

Notation.

- For $0 < p < 1$, the $p\,100\%$ quantile of the distribution $\mathsf{h}_{m,\nu}(\boldsymbol{\Sigma})$ will be denoted as $\mathsf{h}_{m,\nu}(p;\, \boldsymbol{\Sigma})$. That is, $\mathsf{h}_{m,\nu}(p;\, \boldsymbol{\Sigma})$ is the number satisfying

$$\mathsf{P}\Big(\max_{j=1,\ldots,m} |T_j| \le \mathsf{h}_{m,\nu}(p;\, \boldsymbol{\Sigma})\Big) = p.$$

- The distribution function of the random variable with distribution $\mathsf{h}_{m,\nu}(\boldsymbol{\Sigma})$ will be denoted $\mathsf{CDF}_{\mathsf{h},m,\nu}(\cdot;\, \boldsymbol{\Sigma})$.

Notes.

- If the scale matrix $\boldsymbol{\Sigma}$ is positive definite (invertible), the random vector $\boldsymbol{T} \sim \mathsf{mvt}_{m,\nu}\big(\boldsymbol{\Sigma}\big)$ has a density w.r.t. Lebesgue measure

$$f_T(\boldsymbol{t}) = \frac{\Gamma\big(\frac{\nu+m}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)\, \nu^{\frac{m}{2}}\, \pi^{\frac{m}{2}}} \big|\boldsymbol{\Sigma}\big|^{-\frac{1}{2}} \left\{1 + \frac{\boldsymbol{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{t}}{\nu}\right\}^{-\frac{\nu+m}{2}}, \qquad \boldsymbol{t} \in \mathbb{R}^m.$$

- The distribution function $\mathsf{CDF}_{\mathsf{h},m,\nu}(\cdot;\, \boldsymbol{\Sigma})$ of a random variable $H = \max_{j=1,\ldots,m} |T_j|$ is then (for $h > 0$):

$$\mathsf{CDF}_{\mathsf{h},m,\nu}(h;\, \boldsymbol{\Sigma}) = \mathsf{P}\Big(\max_{j=1,\ldots,m} |T_j| \le h\Big) = \mathsf{P}\Big(\forall j = 1, \ldots, m \quad |T_j| \le h\Big)$$

$$= \int_{-h}^{h} \cdots \int_{-h}^{h} f_T(t_1, \ldots, t_m)\, \mathsf{d}t_1 \cdots \mathsf{d}t_m.$$

- That is, when calculating the CDF of the random variable H having the max-abs-t distribution, it is necessary to calculate integrals from a density of a multivariate t-distribution.
  - Computationally efficient methods not available until 90's of the 20th century.
  - Nowadays, see, e.g., Genz and Bretz (2009) and the R packages `mvtnorm` or `mnormt`.
  - Calculation of $\mathsf{CDF}_{\mathsf{h},m,\nu}(\cdot;\, \boldsymbol{\Sigma})$ is also possible with a singular scale matrix $\boldsymbol{\Sigma}$.

10.4.2 General multiple comparison procedure for a linear model

Assumptions.

In the following, we consider a normal linear model

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k.$$

Further, let

$$\mathbb{L}_{m\times k} = \begin{pmatrix} \mathbf{l}_1^\top \\ \vdots \\ \mathbf{l}_m^\top \end{pmatrix}$$

be a matrix such that

$$\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta} = \big(\mathbf{l}_1^\top\boldsymbol{\beta},\, \ldots,\, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \big(\theta_1,\, \ldots,\, \theta_m\big)^\top$$

is an estimable vector parameter with $\mathbf{l}_1 \neq \mathbf{0}_k,\, \ldots,\, \mathbf{l}_m \neq \mathbf{0}_k$.

Notes.

- The number m of the estimable parameters of interest may be arbitrary, i.e., even greater than r or k.
- The rows of the matrix $\mathbb{L}$ may be linearly dependent vectors.

Multiple comparison problem.

A multiple comparison procedure that will be developed aims in providing a simultaneous inference on m estimable parameters $\theta_1,\, \ldots,\, \theta_m$ with the multiple testing problem composed of m elementary hypotheses

$$\mathrm{H}_j\colon\ \ \theta_j = \theta_j^0, \qquad j = 1,\, \ldots,\, m,$$

for some $\boldsymbol{\theta}^0 = \big(\theta_1^0,\, \ldots,\, \theta_m^0\big)^\top \in \mathbb{R}^m$. The global null hypothesis is as usual $\mathrm{H}_0\colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$.

Notation. In the following, the following (standard) notation will be used:

- $\boldsymbol{b} = \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\boldsymbol{Y}$ (any solution to normal equations $\mathbb{X}^\top\mathbb{X}\boldsymbol{b} = \mathbb{X}^\top\boldsymbol{Y}$);
- $\widehat{\boldsymbol{\theta}} = \mathbb{L}\boldsymbol{b} = \big(\mathbf{l}_1^\top\boldsymbol{b},\, \ldots,\, \mathbf{l}_m^\top\boldsymbol{b}\big)^\top = \big(\widehat{\theta}_1,\, \ldots,\, \widehat{\theta}_m\big)^\top$: LSE of $\boldsymbol{\theta}$;
- $\mathbb{V} = \mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{L}^\top = \big(v_{j,l}\big)_{j,l=1,\ldots,m}$ (which does not depend on a choice of a pseudoinverse $\big(\mathbb{X}^\top\mathbb{X}\big)^-$);
- $\mathbb{D} = \mathsf{diag}\left(\dfrac{1}{\sqrt{v_{1,1}}},\, \ldots,\, \dfrac{1}{\sqrt{v_{m,m}}}\right)$;
- MS_e: the residual mean square of the model with $\nu_e = n - r$ degrees of freedom.

Reminders from Chapter 3

- For $j = 1, \ldots, m$, (both conditionally given $\mathbb{X}$ and unconditionally as well):

$$Z_j := \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2\, v_{j,j}}} \;\sim\; \mathcal{N}(0,\, 1), \qquad\qquad T_j := \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\mathsf{MS}_e\, v_{j,j}}} \;\sim\; \mathsf{t}_{n-r}.$$

- Further (conditionally given $\mathbb{X}$):

$$\boldsymbol{Z} = \big(Z_1,\, \ldots,\, Z_m\big)^\top = \frac{1}{\sqrt{\sigma^2}}\, \mathbb{D}\, \big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \;\sim\; \mathcal{N}_m\big(\mathbf{0}_m,\, \mathbb{D}\mathbb{V}\mathbb{D}\big),$$

$$\boldsymbol{T} = \big(T_1,\, \ldots,\, T_m\big)^\top = \frac{1}{\sqrt{\mathsf{MS}_e}}\, \mathbb{D}\, \big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \;\sim\; \mathsf{mvt}_{m,\, n-r}\big(\mathbb{D}\mathbb{V}\mathbb{D}\big).$$

Notes.

- Matrices $\mathbb{V}$ and $\mathbb{D}\mathbb{V}\mathbb{D}$ are not necessarily invertible.
- If $\mathsf{rank}\big(\mathbb{L}\big) = m \le r$ then both matrices $\mathbb{V}$ and $\mathbb{D}\mathbb{V}\mathbb{D}$ are invertible and Theorem 3.2 further provides (both conditionally given $\mathbb{X}$ and unconditionally as well) that under H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$:

$$Q_0 = \frac{1}{m}\, \big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big)^\top\, \big(\mathsf{MS}_e\, \mathbb{V}\big)^{-1} \big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big) = \frac{1}{m}\, \boldsymbol{T}^\top\, \big(\mathbb{D}\mathbb{V}\mathbb{D}\big)^{-1}\, \boldsymbol{T} \;\sim\; \mathcal{F}_{m,\, n-r}.$$

  This was used to test the global null hypothesis $\mathrm{H}_0 :\; \boldsymbol{\theta} = \boldsymbol{\theta}^0$ and to derive the elliptical confidence sets for $\boldsymbol{\theta}$.
- It can also be shown that if $m_0 = \mathsf{rank}\big(\mathbb{L}\big)$ then under H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$:

$$Q_0 = \frac{1}{m}\, \big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big)^\top\, \big(\mathsf{MS}_e\, \mathbb{V}\big)^{+} \big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big) = \frac{1}{m}\, \boldsymbol{T}^\top\, \big(\mathbb{D}\mathbb{V}\mathbb{D}\big)^{+}\, \boldsymbol{T} \;\sim\; \mathcal{F}_{m_0,\, n-r}$$

  (both conditionally given $\mathbb{X}$ and unconditionally), where symbol $+$ denotes the Moore-Penrose pseudoinverse.

Some derivations

Let for $\theta_j^0 \in \mathbb{R}$, $j = 1, \ldots, m$,

$$T_j(\theta_j^0) = \frac{\widehat{\theta}_j - \theta_j^0}{\sqrt{\mathsf{MS}_e\, v_{j,j}}}, \qquad j = 1, \ldots, m.$$

Then, under H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$:

$$\boldsymbol{T}\big(\boldsymbol{\theta}^0\big) := \big(T_1(\theta_1^0),\, \ldots,\, T_m(\theta_m^0)\big)^\top \;\sim\; \mathsf{mvt}_{m,\, n-r}\big(\mathbb{D}\mathbb{V}\mathbb{D}\big).$$

We then have, for $0 < \alpha < 1$:

$$\begin{aligned}
1 - \alpha \;&=\; \mathsf{P}\Big(\max_{j=1,\ldots,m} \big|T_j(\theta_j^0)\big| < \mathsf{h}_{m,\, n-r}(1-\alpha;\, \mathbb{D}\mathbb{V}\mathbb{D});\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \\
&=\; \mathsf{P}\Big(\text{for all } j = 1, \ldots, m \quad \big|T_j(\theta_j^0)\big| < \mathsf{h}_{m,\, n-r}(1-\alpha;\, \mathbb{D}\mathbb{V}\mathbb{D});\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \\
&=\; \mathsf{P}\left(\text{for all } j = 1, \ldots, m \quad \left|\frac{\widehat{\theta}_j - \theta_j^0}{\sqrt{\mathsf{MS}_e\, v_{j,j}}}\right| < \mathsf{h}_{m,\, n-r}(1-\alpha;\, \mathbb{D}\mathbb{V}\mathbb{D});\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\right) \\
&=\; \mathsf{P}\Big(\text{for all } j = 1, \ldots, m \quad \big(\theta_j^{HL}(\alpha),\; \theta_j^{HU}(\alpha)\big) \ni \theta_j^0;\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big), \qquad (10.9)
\end{aligned}$$

where

$$\begin{aligned} \theta_j^{HL}(\alpha) &= \widehat{\theta}_j - \mathsf{h}_{m,\,n-r}(1-\alpha;\,\mathbb{DVD})\,\sqrt{\mathsf{MS}_e\,v_{j,j}}, \\ \theta_j^{HU}(\alpha) &= \widehat{\theta}_j + \mathsf{h}_{m,\,n-r}(1-\alpha;\,\mathbb{DVD})\,\sqrt{\mathsf{MS}_e\,v_{j,j}}, \qquad j = 1,\,\ldots,\,m. \end{aligned} \tag{10.10}$$

Theorem 10.5 Hothorn-Bretz-Westfall MCP for linear hypotheses in a normal linear model.

Random intervals given by (10.10) are simultaneous confidence intervals for parameters $\theta_j = \mathbf{l}^\top\boldsymbol{\beta}$, $j = 1,\ldots,m$, with an exact coverage of $1-\alpha$, i.e., for any $\boldsymbol{\theta}^0 = \left(\theta_1^0,\ldots,\theta_m^0\right)^\top \in \mathbb{R}^m$

$$\mathsf{P}\Big(\textit{for all } j = 1,\ldots,m \quad \Big(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Big) \ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big) \ = \ 1-\alpha.$$

Related P-values for a multiple testing problem with elementary hypotheses $H_j:\ \theta_j = \theta_j^0$, $\theta_j^0 \in \mathbb{R}$, $j = 1,\ldots,m$, adjusted for multiple comparison are given by

$$p_j^H = 1 \ - \ \mathsf{CDF}_{\mathsf{h},m,n-r}\Big(\big|t_j^0\big|;\,\mathbb{DVD}\Big), \qquad j = 1,\ldots,m,$$

where t_j^0 is a value of $T_j(\theta_j^0) = \frac{\widehat{\theta}_j - \theta_j^0}{\sqrt{\mathsf{MS}_e\,v_{j,j}}}$ attained with given data.

Proof.

The fact that $\Big(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Big)$, $j = 1,\ldots,m$, are simultaneous confidence intervals for parameters $\theta_j = \mathbf{l}_j^\top\boldsymbol{\beta}$ with an *exact* coverage of $1-\alpha$ follows from (10.9).

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with the elementary hypotheses $\mathrm{H}_j:\ \theta_j = \theta_j^0$, $j = 1,\ldots,m$, follows from noting the following (for each $j = 1,\ldots,m$):

$$\Big(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Big) \not\ni \theta_j^0 \quad \Longleftrightarrow \quad \Big|T_j\big(\theta_j^0\big)\Big| \geq \mathsf{h}_{m,\,n-r}(1-\alpha;\,\mathbb{DVD}).$$

It now follows from monotonicity of the quantiles of a continuous max-abs-t-distribution that

$$p_j^H = \inf\Big\{\alpha:\ \Big(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Big) \not\ni \theta_j^0\Big\} = \inf\Big\{\alpha:\ \Big|T_j\big(\theta_j^0\big)\Big| \geq \mathsf{h}_{m,\,n-r}(1-\alpha;\,\mathbb{DVD})\Big\}$$

is attained for p_j^H satisfying

$$\Big|T_j\big(\theta_j^0\big)\Big| = \mathsf{h}_{m,\,n-r}(1-p_j^H;\,\mathbb{DVD}).$$

That is, if t_j^0 is a value of the statistic $T_j\big(\theta_j^0\big)$ attained with given data, we have

$$p_j^H = 1 \ - \ \mathsf{CDF}_{\mathsf{h},m,n-r}\Big(\big|t_j^0\big|;\,\mathbb{DVD}\Big).$$

❑

Note *(Hothorn-Bretz-Westfall MCP in the* R *software).*

In the R software, the Hothorn-Bretz-Westfall MCP for linear hypotheses on parameters of (generalized) linear models is implemented in the package multcomp. After fitting a model (by the function `lm`), it is necessary to call sequentially the following functions:

(i) `glht`. One of its arguments specifies the linear hypothesis of interest (specification of the $\mathbb{L}$ matrix). Note that for some common hypotheses, certain keywords can be used. For example, pairwise comparison of all group means in context of the ANOVA models is achieved by specifying the keyword "`Tukey`". Nevertheless, note that invoked MCP is still that of Hothorn-Bretz-Westfall and it is not based on the Tukey's procedure. The "`Tukey`" keyword only specifies what should be compared and not how it should be compared.

(ii) `summary` (applied on an object of class `glht`) provides P-values adjusted for multiple comparison.

(iii) `confint` (applied on an object of class `glht`) provides simultaneous confidence intervals which, among other things, requires calculation of a critical value $\mathsf{h}_{m,\,n-r}(1-\alpha)$, that is also available in the output.

Note that both calculation of the P-values adjusted for multiple comparison and calculation of the critical value $\mathsf{h}_{m,\,n-r}(1-\alpha)$ needed for the simultaneous confidence intervals requires calculation of a multivariate `t` integral. This is calculated by a Monte Carlo integration (i.e., based on a certain stochastic simulation) and hence the results slightly differ if repeatedly calculated at different occasions. Setting a seed of the random number generator (`set.seed()`) is hence recommended for full reproducibility of the results.

10.5 Confidence band for the regression function

In this section, we shall assume that data are represented by i.i.d. random vectors $\big(Y_i,\, \boldsymbol{Z}_i^\top\big)^\top$, $i = 1, \ldots, n$, being sampled from a distribution of a generic random vector $\big(Y,\, \boldsymbol{Z}^\top\big)^\top \in \mathbb{R}^{1+p}$. It is further assumed that for some known transformation $\boldsymbol{t} : \mathbb{R}^p \longrightarrow \mathbb{R}^k$, a normal linear model with regressors $\boldsymbol{X}_i = \boldsymbol{t}(\boldsymbol{Z}_i)$, $i = 1, \ldots, n$, holds. That is, it is assumed that for the response vector $\boldsymbol{Y}$, the covariate matrix $\mathbb{Z}$ and the model matrix $\mathbb{X}$, where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{Z} = \begin{pmatrix} \boldsymbol{Z}_1^\top \\ \vdots \\ \boldsymbol{Z}_n^\top \end{pmatrix}, \qquad \mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{t}^\top(\boldsymbol{Z}_1) \\ \vdots \\ \boldsymbol{t}^\top(\boldsymbol{Z}_n) \end{pmatrix},$$

we have

$$\boldsymbol{Y} \,\big|\, \mathbb{Z} \;\sim\; \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big) \tag{10.11}$$

for some $\boldsymbol{\beta} \in \mathbb{R}^k$, $\sigma^2 > 0$. Remember that it follows from (10.11) that

$$Y_i \;\big|\; \boldsymbol{Z}_i \;\sim\; \mathcal{N}\big(\boldsymbol{X}_i^\top \boldsymbol{\beta},\, \sigma^2\big),$$

and the error terms $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$ are i.i.d. distributed as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The corresponding regression function is

$$\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{t}(\boldsymbol{z})\big) = \mathbb{E}\big(Y \,\big|\, \boldsymbol{Z} = \boldsymbol{z}\big) = m(\boldsymbol{z}) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}, \qquad \boldsymbol{z} \in \mathbb{R}^p.$$

It will further be assumed that the model matrix $\mathbb{X}$ is of full-rank (almost surely), i.e., $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$. As it is usual, $\widehat{\boldsymbol{\beta}}$ will be the LSE of a vector of $\boldsymbol{\beta}$ and MS_e the residual mean square.

Reminder from Section 3.3

Let $\boldsymbol{z} \in \mathbb{R}^p$ be given. Theorem 3.3 then states that a random interval with the lower and upper bounds given as

$$\boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;\pm\; \mathsf{t}_{n-k}\Big(1 - \frac{\alpha}{2}\Big) \sqrt{\mathsf{MS}_e\, \boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})},$$

is the confidence interval for $m(\boldsymbol{z}) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}$ with a coverage of $1 - \alpha$. That is, for given $\boldsymbol{z} \in \mathbb{R}^p$, for any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$,

$$\mathsf{P}\Big(\boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;\pm\; \mathsf{t}_{n-k}\Big(1 - \frac{\alpha}{2}\Big) \sqrt{\mathsf{MS}_e\, \boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})} \;\ni\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0;\; \boldsymbol{\beta} = \boldsymbol{\beta}^0\Big) \;=\; 1 - \alpha.$$

Theorem 10.6 Confidence band for the regression function.

Let $\big(Y_i,\, \boldsymbol{Z}_i^\top\big)^\top$, $i = 1, \ldots, n$, be i.i.d. random vectors such that $\boldsymbol{Y} \,\big|\, \mathbb{Z} \;\sim\; \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, where $\mathbb{X}$ is the $n \times k$ model matrix based on a known transformation $\boldsymbol{t} : \mathbb{R}^p \longrightarrow \mathbb{R}^k$ of the covariates $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. Let $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$. Finally, let for all $\boldsymbol{z} \in \mathbb{R}^p$ $\boldsymbol{t}(\boldsymbol{z}) \neq \mathbf{0}_k$. Then for any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$

$$\mathsf{P}\Big(\textit{for all } \boldsymbol{z} \in \mathbb{R}^p$$

$$\boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;\pm\; \sqrt{k\, \mathcal{F}_{k,\,n-k}(1-\alpha)\, \mathsf{MS}_e\, \boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})} \;\ni\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0;\quad \boldsymbol{\beta} = \boldsymbol{\beta}^0\Big)$$

$$= \; 1 - \alpha.$$

Note. Requirement $\boldsymbol{t}(\boldsymbol{z}) \neq \mathbf{0}_k$ for all $\boldsymbol{z} \in \mathbb{R}^p$ is not too restrictive from a practical point of view as it is satisfied, e.g., for all linear models with intercept.

Proof. Let (for $0 < \alpha < 1$)

$$\mathcal{K} = \Big\{\boldsymbol{\beta} \in \mathbb{R}^k : \ \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big)^\top \big(\mathbb{X}^\top \mathbb{X}\big) \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big) \ \leq \ k\, \mathsf{MS}_e\, \mathcal{F}_{k,n-k}(1-\alpha)\Big\}.$$

Section 3.2: $\mathcal{K}$ is a confidence *ellipsoid* for $\boldsymbol{\beta}$ with a coverage of $1-\alpha$, that is, for any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$

$$\mathsf{P}\big(\mathcal{K} \ni \boldsymbol{\beta}^0;\ \boldsymbol{\beta} = \boldsymbol{\beta}^0\big) = 1 - \alpha.$$

$\mathcal{K}$ is an ellipsoid in $\mathbb{R}^k$, that is, bounded, convex and with our definition also *closed* subset of $\mathbb{R}^k$.

Let for $\boldsymbol{z} \in \mathbb{R}^p$:

$$L(\boldsymbol{z}) = \inf_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}, \qquad U(\boldsymbol{z}) = \sup_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}.$$

From construction:

$$\boldsymbol{\beta} \in \mathcal{K} \ \ \Rightarrow \ \ \forall \boldsymbol{z} \in \mathbb{R}^p \quad L(\boldsymbol{z}) \ \leq \ \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} \ \leq \ U(\boldsymbol{z}).$$

Due to the fact that $\mathcal{K}$ is bounded, convex and closed, we also have

$$\forall \boldsymbol{z} \in \mathbb{R}^p \quad L(\boldsymbol{z}) \ \leq \ \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} \ \leq \ U(\boldsymbol{z}) \ \ \Rightarrow \ \ \boldsymbol{\beta} \in \mathcal{K}.$$

That is,

$$\boldsymbol{\beta} \in \mathcal{K} \ \ \Leftrightarrow \ \ \forall \boldsymbol{z} \in \mathbb{R}^p \quad L(\boldsymbol{z}) \ \leq \ \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} \ \leq \ U(\boldsymbol{z}).$$

and hence, for any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$,

$$1 - \alpha = \mathsf{P}\big(\mathcal{K} \ni \boldsymbol{\beta}^0;\ \boldsymbol{\beta} = \boldsymbol{\beta}^0\big) = \mathsf{P}\big(\text{for all } \boldsymbol{z} \in \mathbb{R}^p \quad L(\boldsymbol{z}) \ \leq \ \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0 \ \leq \ U(\boldsymbol{z});\ \boldsymbol{\beta} = \boldsymbol{\beta}^0\big). \tag{10.12}$$

Further, since $\boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}$ is a linear function (in $\boldsymbol{\beta}$) and $\mathcal{K}$ is bounded, convex and closed, we have

$$L(\boldsymbol{z}) = \inf_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} = \min_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}, \qquad U(\boldsymbol{z}) = \sup_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} = \max_{\boldsymbol{\beta} \in \mathcal{K}} \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta},$$

and both extremes must lie on a boundary of $\mathcal{K}$, that is, both extremes are reached for $\boldsymbol{\beta}$ satisfying $\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big)^\top \big(\mathbb{X}^\top \mathbb{X}\big) \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big) \ = \ k\, \mathsf{MS}_e\, \mathcal{F}_{k,n-k}(1-\alpha)$.

Method of Lagrange multipliers:

$$\varphi(\boldsymbol{\beta},\, \lambda) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} \ + \ \frac{1}{2}\, \lambda \Big\{ \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big)^\top \big(\mathbb{X}^\top \mathbb{X}\big) \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big) \ - \ k\, \mathsf{MS}_e\, \mathcal{F}_{k,n-k}(1-\alpha) \Big\}$$

($\frac{1}{2}$ is only included to simplify subsequent expressions).

Derivatives of φ:

$$\frac{\partial\varphi}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta},\,\lambda) = \boldsymbol{t}(\boldsymbol{z}) \,+\, \lambda\,\mathbb{X}^\top\mathbb{X}\,\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\big),$$

$$\frac{\partial\varphi}{\partial\lambda}(\boldsymbol{\beta},\,\lambda) = \frac{1}{2}\Big\{\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\big)^\top\,\big(\mathbb{X}^\top\mathbb{X}\big)\,\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\big) \,-\, k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)\Big\}.$$

With given λ, the first set of equations is solved (with respect to $\boldsymbol{\beta}$) for

$$\boldsymbol{\beta}(\lambda) = \widehat{\boldsymbol{\beta}} \,-\, \frac{1}{\lambda}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z}).$$

Use $\boldsymbol{\beta}(\lambda)$ in the second equation:

$$\frac{1}{\lambda^2}\boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z}) = k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha),$$

$$\lambda = \pm\sqrt{\frac{\boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})}{k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)}}.$$

Hence, $\boldsymbol{\beta}$ which minimizes/maximizes $\boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}$ subject to

$$\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\big)^\top\,\big(\mathbb{X}^\top\mathbb{X}\big)\,\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\big) \;=\; k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)$$

is given as

$$\boldsymbol{\beta}_{min} = \widehat{\boldsymbol{\beta}} - \sqrt{\frac{k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)}{\boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})}}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z}),$$

$$\boldsymbol{\beta}_{max} = \widehat{\boldsymbol{\beta}} + \sqrt{\frac{k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)}{\boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})}}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z}).$$

Note that with our assumptions of $\boldsymbol{t}(\boldsymbol{z}) \neq \mathbf{0}$, we never divide by zero since $\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$ is a positive definite matrix.

That is,

$$\begin{aligned} L(\boldsymbol{z}) &= \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}_{min} \\ &= \boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;-\; \sqrt{\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})\;k\,\mathcal{F}_{k,n-k}(1-\alpha)}, \\ U(\boldsymbol{z}) &= \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}_{max} \\ &= \boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;+\; \sqrt{\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})\;k\,\mathcal{F}_{k,n-k}(1-\alpha)}. \end{aligned}$$

The proof is finalized by looking back at expression (10.12) and realizing that, due to continuity,

$$\begin{aligned} 1-\alpha &= \mathsf{P}\big(\text{for all } \boldsymbol{z}\in\mathbb{R}^p \quad L(\boldsymbol{z}) \;\le\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0 \;\le\; U(\boldsymbol{z});\;\boldsymbol{\beta}=\boldsymbol{\beta}^0\big) \\ &= \mathsf{P}\big(\text{for all } \boldsymbol{z}\in\mathbb{R}^p \quad L(\boldsymbol{z}) \;<\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0 \;<\; U(\boldsymbol{z});\;\boldsymbol{\beta}=\boldsymbol{\beta}^0\big) \\ &= \mathsf{P}\Big(\text{for all } \boldsymbol{z}\in\mathbb{R}^p \\ &\qquad \boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;\pm\; \sqrt{k\,\mathcal{F}_{k,\,n-k}(1-\alpha)\,\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})} \;\ni\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0;\;\boldsymbol{\beta}=\boldsymbol{\beta}^0\Big). \end{aligned}$$

❑

Terminology *(Confidence band for the regression function).*

If the covariates $Z_1, \ldots, Z_n \in \mathbb{R}$, confidence intervals according to Theorem (10.6) are often calculated for an (equidistant) sequence of values $z_1, \ldots, z_N \in \mathbb{R}$ and then plotted together with the fitted regression function $\widehat{m}(z) = \boldsymbol{t}^\top(z)\widehat{\boldsymbol{\beta}}$, $z \in \mathbb{R}$. A band that is obtained in this way is called the *confidence band for the regression function* [6] as it covers jointly all true values of the regression function with a given probability of $1 - \alpha$.

Note *(Confidence band for and around the regression function).*

For given $\boldsymbol{z} \in \mathbb{R}$:

Half width of the confidence band **FOR** the regression function (overall coverage) is

$$\sqrt{k\, \mathcal{F}_{k,n-k}(1-\alpha)\ \ \mathsf{MS}_e\, \boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})}.$$

Half width of the confidence band **AROUND** the regression function (pointwise coverage) is

$$\begin{aligned} & \mathsf{t}_{n-k}\Big(1-\frac{\alpha}{2}\Big)\ \ \sqrt{\mathsf{MS}_e\, \boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})} \\ = & \sqrt{\mathcal{F}_{1,n-k}(1-\alpha)\ \ \mathsf{MS}_e\, \boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})}, \end{aligned}$$

since for any $\nu > 0$, $\mathsf{t}^2_\nu\Big(1-\frac{\alpha}{2}\Big) = \mathcal{F}_{1,\nu}(1-\alpha)$.

For $k \geq 2$, and any $\nu > 0$,

$$k\, \mathcal{F}_{k,\nu}(1-\alpha) > \mathcal{F}_{1,\nu}(1-\alpha)$$

and hence the confidence band for the regression function is indeed wider than the confidence band around the regression function. Their width is the same only if $k = 1$.

End of Lecture #20 (01/12/2016)

[6] *pás spolehlivosti pro regresní funkci*

Chapter 11

Checking Model Assumptions

Start of Lecture #21 (07/12/2016)

In Chapter 4, we introduced some basic, mostly graphical methods to check the model assumptions. Now, we introduce some additional methods, mostly based on statistical tests. As in Chapter 4, we assume that data are represented by n random vectors $\left(Y_i, \boldsymbol{Z}_i^\top\right)^\top$, $\boldsymbol{Z}_i = \left(Z_{i,1}, \ldots, Z_{i,p}\right)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$ $i = 1, \ldots, n$. Possibly two sets of regressors have been derived from the covariates:

(i) $\boldsymbol{X}_i$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = \boldsymbol{t}_X(\boldsymbol{Z}_i)$ for some transformation $\boldsymbol{t}_X : \mathbb{R}^p \longrightarrow \mathbb{R}^k$. They give rise to the model matrix

$$\mathbb{X}_{n\times k} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \left(\boldsymbol{X}^0, \ldots, \boldsymbol{X}^{k-1}\right).$$

For most practical problems, $\boldsymbol{X}^0 = \left(1, \ldots, 1\right)^\top$ (almost surely).

(ii) $\boldsymbol{V}_i$, $i = 1, \ldots, n$, where $\boldsymbol{V}_i = \boldsymbol{t}_V(\boldsymbol{Z}_i)$ for some transformation $\boldsymbol{t}_V : \mathbb{R}^p \longrightarrow \mathbb{R}^l$. They give rise to the model matrix

$$\mathbb{V}_{n\times l} = \begin{pmatrix} \boldsymbol{V}_1^\top \\ \vdots \\ \boldsymbol{V}_n^\top \end{pmatrix} = \left(\boldsymbol{V}^1, \ldots, \boldsymbol{V}^l\right).$$

Primarily, we will assume that the model matrix $\mathbb{X}$ is sufficient to be able to assume that $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right) = \mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{k-1}\right)^\top \in \mathbb{R}^k$. That is, we will arrive from assuming

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

or even from assuming normality, i.e.,

$$\boldsymbol{Y} \,\middle|\, \mathbb{Z} \sim \mathcal{N}_n\left(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right).$$

The task is now to verify appropriateness of those assumptions that, in principle, consist of four subassumptions outlined in Chapter 4, which can all be written while using the error terms $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \ldots, \varepsilon_n\right)^\top = \left(Y_1 - \boldsymbol{X}_1^\top\boldsymbol{\beta}, \ldots, Y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\right)^\top = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$:

(A1) Correct regression function $\equiv$ (Conditionally) errors with zero mean $\equiv$ $\mathbb{E}\left(\varepsilon_i \,\middle|\, \boldsymbol{Z}_i\right) = 0$, $i = 1, \ldots, n$.

(A2) (Conditional) homoscedasticity of errors $\equiv \mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\big) = \sigma^2 = \text{const}$, $i = 1, \ldots, n$.

(A3) (Conditionally) uncorrelated/independent errors $\equiv \mathsf{cov}\big(\varepsilon_i,\, \varepsilon_j \,\big|\, \mathbb{Z}\big) = 0$, $i \neq j$.

(A4) (Conditionally) normal errors $\equiv \varepsilon_i \,\big|\, \mathbb{Z} \overset{\text{indep.}}{\sim} \mathcal{N}$.

The four assumptions then gradually imply

(A1) Errors with (marginal) zero mean: $\mathbb{E}\big(\varepsilon_i\big) = 0$, $i = 1, \ldots, n$.

(A2) (Marginal) homoscedasticity of errors $\equiv \mathsf{var}\big(\varepsilon_i\big) = \sigma^2 = \text{const}$, $i = 1, \ldots, n$.

(A3) (Marginally) uncorrelated/independent errors $\equiv \mathsf{cov}\big(\varepsilon_i,\, \varepsilon_j\big) = 0$, $i \neq j$.

(A4) (Marginally) Normal errors $\equiv \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}$.

11.1 Model with added regressors

In this section, we technically derive some expressions that will be useful in latter sections of this chapter and also in Chapter 14. We will deal with two models:

(i) Model M: $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$.

(ii) Model M_g: $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big)$, where the model matrix is an $n \times (k+l)$ matrix $\mathbb{G}$,

$$\mathbb{G} = \big(\mathbb{X},\, \mathbb{V}\big).$$

***Notation** (Quantities derived under the two models).*

(i) Quantities derived while assuming model M will be denoted as it is usual. In particular:

- (Any) solution to normal equations: $\boldsymbol{b} = \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\boldsymbol{Y}$. In case of a full-rank model matrix $\mathbb{X}$:
$$\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}$$
is the LSE of a vector $\boldsymbol{\beta}$ in model M;
- Hat matrix (projection matrix into the regression space $\mathcal{M}\big(\mathbb{X}\big)$):
$$\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top = \big(h_{i,t}\big)_{i,t=1,\ldots,n};$$
- Fitted values $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \big(\widehat{Y}_1, \ldots, \widehat{Y}_n\big)^\top$;
- Projection matrix into the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$:
$$\mathbb{M} = \mathbf{I}_n - \mathbb{H} = \big(m_{i,t}\big)_{i,t=1,\ldots,n};$$
- Residuals: $\boldsymbol{U} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \mathbb{M}\boldsymbol{Y} = \big(U_1, \ldots, U_n\big)^\top$;
- Residual sum of squares: $\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2$.

(ii) Analogous quantities derived while assuming model M_g will be indicated by a subscript g:

- (Any) solution to normal equations: $\big(\boldsymbol{b}_g^\top,\, \boldsymbol{c}_g^\top\big)^\top = \big(\mathbb{G}^\top\mathbb{G}\big)^-\mathbb{G}^\top\boldsymbol{Y}$. In case of a full-rank model matrix $\mathbb{G}$:
$$\big(\widehat{\boldsymbol{\beta}}_g^\top,\, \widehat{\boldsymbol{\gamma}}_g^\top\big)^\top = \big(\mathbb{G}^\top\mathbb{G}\big)^{-1}\mathbb{G}^\top\boldsymbol{Y}$$
provides the LSE of vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in model M_g;
- Hat matrix (projection matrix into the regression space $\mathcal{M}\big(\mathbb{G}\big)$):
$$\mathbb{H}_g = \mathbb{G}\big(\mathbb{G}^\top\mathbb{G}\big)^-\mathbb{G}^\top = \big(h_{g,i,t}\big)_{i,t=1,\ldots,n};$$
- Fitted values $\widehat{\boldsymbol{Y}}_g = \mathbb{H}_g\boldsymbol{Y} = \big(\widehat{Y}_{g,1}, \ldots, \widehat{Y}_{g,n}\big)^\top$;
- Projection matrix into the residual space $\mathcal{M}\big(\mathbb{G}\big)^\perp$:
$$\mathbb{M}_g = \mathbf{I}_n - \mathbb{H}_g = \big(m_{g,i,t}\big)_{i,t=1,\ldots,n};$$
- Residuals: $\boldsymbol{U}_g = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_g = \mathbb{M}_g\boldsymbol{Y} = \big(U_{g,1}, \ldots, U_{g,n}\big)^\top$;
- Residual sum of squares: $\mathsf{SS}_{e,g} = \big\|\boldsymbol{U}_g\big\|^2$.

Lemma 11.1 Model with added regressors.

Quantities derived while assuming model $\mathsf{M} : \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *and quantities derived while assuming model* $\mathsf{M}_g : \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big)$ *are mutually in the following relationship.*

$$\begin{aligned}\widehat{\boldsymbol{Y}}_g &= \widehat{\boldsymbol{Y}} + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}\\ &= \mathbb{X}\boldsymbol{b}_g + \mathbb{V}\boldsymbol{c}_g, \qquad \textit{for some } \boldsymbol{b}_g \in \mathbb{R}^k,\ \boldsymbol{c}_g \in \mathbb{R}^l.\end{aligned}$$

Vector $\boldsymbol{b}_g$ *and* $\boldsymbol{c}_g$ *such that* $\widehat{\boldsymbol{Y}}_g = \mathbb{X}\boldsymbol{b}_g + \mathbb{V}\boldsymbol{c}_g$ *satisfy:*

$$\begin{aligned}\boldsymbol{c}_g &= \big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U},\\ \boldsymbol{b}_g &= \boldsymbol{b} - \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{V}\boldsymbol{c}_g \qquad \textit{for some } \boldsymbol{b} = \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\boldsymbol{Y}.\end{aligned}$$

Finally

$$\mathsf{SS}_e - \mathsf{SS}_{e,g} = \big\|\mathbb{M}\mathbb{V}\boldsymbol{c}_g\big\|^2.$$

Proof.

- $\widehat{\boldsymbol{Y}}_g$ is a projection of $\boldsymbol{Y}$ into $\mathcal{M}\big(\mathbb{X},\, \mathbb{V}\big) = \mathcal{M}\big(\mathbb{X},\, \mathbb{M}\mathbb{V}\big)$.
- Use "$\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top$":

$$\begin{aligned}\mathbb{H}_g &= \begin{pmatrix}\mathbb{X}, & \mathbb{M}\mathbb{V}\end{pmatrix}\begin{pmatrix}\mathbb{X}^\top\mathbb{X} & \underbrace{\mathbb{X}^\top\mathbb{M}\mathbb{V}}_{\mathbf{0}}\\ \underbrace{\mathbb{V}^\top\mathbb{M}\mathbb{X}}_{\mathbf{0}} & \mathbb{V}^\top\mathbb{M}\mathbb{V}\end{pmatrix}^-\begin{pmatrix}\mathbb{X}^\top\\ \mathbb{V}^\top\mathbb{M}\end{pmatrix}\\ &= \begin{pmatrix}\mathbb{X}, & \mathbb{M}\mathbb{V}\end{pmatrix}\begin{pmatrix}\big(\mathbb{X}^\top\mathbb{X}\big)^- & \mathbf{0}\\ \mathbf{0} & \big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\end{pmatrix}\begin{pmatrix}\mathbb{X}^\top\\ \mathbb{V}^\top\mathbb{M}\end{pmatrix}\\ &= \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\mathbb{M}.\end{aligned}$$

- So that,

$$\begin{aligned}\widehat{\boldsymbol{Y}}_g = \mathbb{H}_g\boldsymbol{Y} &= \underbrace{\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\boldsymbol{Y}}_{\widehat{\boldsymbol{Y}}} + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\underbrace{\mathbb{M}\boldsymbol{Y}}_{\boldsymbol{U}}\\ &= \widehat{\boldsymbol{Y}} + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U} \qquad \text{🚲}\end{aligned}$$

- Theorem 2.5: It must be possible to write $\widehat{\boldsymbol{Y}}_g$ as

$$\widehat{\boldsymbol{Y}}_g = \mathbb{X}\boldsymbol{b}_g + \mathbb{V}\boldsymbol{c}_g,$$

  where $\big(\boldsymbol{b}_g,\, \boldsymbol{c}_g\big)^\top$ solves normal equations based on a model matrix $\big(\mathbb{X},\, \mathbb{V}\big)$.
- We rewrite 🚲 to see what $\boldsymbol{b}_g$ and $\boldsymbol{c}_g$ could be.

- Remember that $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$ for any $\boldsymbol{b} = \big(\mathbb{X}^\top\mathbb{X}\big)^- \mathbb{X}^\top\boldsymbol{Y}$. Take now 🚲 and further calculate:

$$\begin{aligned}\widehat{\boldsymbol{Y}}_g &= \underbrace{\mathbb{X}\boldsymbol{b}}_{\widehat{\boldsymbol{Y}}} + \underbrace{\big\{\mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\big\}}_{\mathbb{M}}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}\\ &= \mathbb{X}\boldsymbol{b} + \mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U} - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}\\ &= \mathbb{X}\underbrace{\big\{\boldsymbol{b} - \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}\big\}}_{\boldsymbol{b}_g} + \mathbb{V}\underbrace{\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}}_{\boldsymbol{c}_g}.\end{aligned}$$

- That is, $\boldsymbol{c}_g = \big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}$,
  $\boldsymbol{b}_g = \boldsymbol{b} - \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{V}\boldsymbol{c}_g$.
- Finally
$$\mathsf{SS}_e - \mathsf{SS}_{e,g} = \big\|\widehat{\boldsymbol{Y}}_g - \widehat{\boldsymbol{Y}}\big\|^2 = \big\|\mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^-\mathbb{V}^\top\boldsymbol{U}\big\|^2 = \big\|\mathbb{M}\mathbb{V}\boldsymbol{c}_g\big\|^2.$$

❑

11.2 Correct regression function

We are now assuming a linear model

$$\mathsf{M}: \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big),$$

where the error terms $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$ satisfy (Lemma 1.2):

$$\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \mathbf{0}_n, \quad \mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \sigma^2 \mathbf{I}_n.$$

The assumption (A1) of a correct regression function is, in particular,

$$\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}(\mathbb{X}), \qquad \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \mathbb{X}\boldsymbol{\beta} \quad \text{for some } \boldsymbol{\beta} \in \mathbb{R}^k,$$

$$\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \mathbf{0}_n \qquad \big(\Longrightarrow \;\; \mathbb{E}\big(\boldsymbol{\varepsilon}\big) = \mathbf{0}_n\big).$$

As (also) explained in Section 4.1, assumption (A1) implies

$$\mathbb{E}\big(\boldsymbol{U} \,\big|\, \mathbb{Z}\big) = \mathbf{0}_n$$

and this property is exploited by a basic diagnostic tool which is a plot of residuals against possible factors derived from the covariates $\mathbb{Z}$ that may influence the residuals expectation. Factors traditionally considered are

(i) Fitted values $\widehat{\boldsymbol{Y}}$;

(ii) Regressors included in the model M (columns of the model matrix $\mathbb{X}$);

(iii) Regressors not included in the model M (columns of the model matrix $\mathbb{V}$).

Assumptions.

For the rest of this section, we assume that model M is a model of general rank r with intercept, that is,

$$\mathsf{rank}\big(\mathbb{X}\big) = r \leq k < n, \qquad \mathbb{X} = \Big(\boldsymbol{X}^0,\, \ldots,\, \boldsymbol{X}^{k-1}\Big),\; \boldsymbol{X}^0 = \mathbf{1}_n.$$

In the following, we develop methods to examine whether for given j ($j \in \big\{1,\, \ldots, k-1\big\}$) the jth regressor, i.e., the column $\boldsymbol{X}^j$, is correctly included in the model matrix $\mathbb{X}$. In other words, we will aim in examining whether the jth regressor is possibly responsible for violation of the assumption (A1).

11.2.1 Partial residuals

***Notation** (Model with a removed regressor).*

For $j \in \big\{1,\, \ldots, k-1\big\}$, let $\mathbb{X}^{(-j)}$ denote the model matrix $\mathbb{X}$ without the column $\boldsymbol{X}^j$ and let

$$\boldsymbol{\beta}^{(-j)} = \big(\beta_0,\, \ldots,\, \beta_{j-1},\, \beta_{j+1},\, \ldots,\, \beta_{k-1}\big)^\top$$

denote the regression coefficients vector without the jth element. *Model with a removed jth regressor* will be a linear model

$$\mathsf{M}^{(-j)}: \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}^{(-j)}\boldsymbol{\beta}^{(-j)},\, \sigma^2 \mathbf{I}_n\big).$$

All quantities related to the model $\mathsf{M}^{(-j)}$ will be indicated by a superscript $(-j)$. In particular,

$$\mathbb{M}^{(-j)} = \mathbf{I}_n - \mathbb{X}^{(-j)} \left(\mathbb{X}^{(-j)\top} \mathbb{X}^{(-j)} \right)^{-} \mathbb{X}^{(-j)\top}$$

is a projection matrix into the residual space $\mathcal{M}\left(\mathbb{X}^{(-j)}\right)^{\perp}$;

$$\boldsymbol{U}^{(-j)} = \mathbb{M}^{(-j)} \boldsymbol{Y}$$

is a vector of residuals of the model $\mathsf{M}^{(-j)}$.

Assumptions.

We will assume $\mathsf{rank}\left(\mathbb{X}^{(-j)}\right) = r - 1$ which implies that

(i) $\boldsymbol{X}^j \notin \mathcal{M}\left(\mathbb{X}^{(-j)}\right)$;

(ii) $\boldsymbol{X}^j \neq \mathbf{0}_n$;

(iii) $\boldsymbol{X}^j$ is not a multiple of a vector $\mathbf{1}_n$.

Derivations towards partial residuals

Model M is now a model with one added regressor to a model $\mathsf{M}^{(-j)}$ and the two models form a pair (model–submodel). Let

$$\boldsymbol{b} = \left(b_0, \ldots, b_{j-1}, b_j, b_{j+1}, \ldots, b_{k-1}\right)^{\top}$$

be (any) solution to normal equations in model M. Lemma 11.1 (Model with added regressors) provides

$$b_j = \left(\boldsymbol{X}^{j\top} \mathbb{M}^{(-j)} \boldsymbol{X}^j \right)^{-} \boldsymbol{X}^{j\top} \boldsymbol{U}^{(-j)}. \tag{11.1}$$

Further, since a matrix $\mathbb{M}^{(-j)}$ is idempotent, we have

$$\boldsymbol{X}^{j\top} \mathbb{M}^{(-j)} \boldsymbol{X}^j = \left\| \mathbb{M}^{(-j)} \boldsymbol{X}^j \right\|^2.$$

At the same time, $\mathbb{M}^{(-j)} \boldsymbol{X}^j \neq \mathbf{0}_n$ since $\boldsymbol{X}^j \notin \mathcal{M}\left(\mathbb{X}^{(-j)}\right)$, $\boldsymbol{X}^j \neq \mathbf{0}_n$. Hence, $\boldsymbol{X}^{j\top} \mathbb{M}^{(-j)} \boldsymbol{X}^j > 0$ and a pseudoinverse in (11.1) can be replaced by an inverse. That is,

$$b_j = \widehat{\beta}_j = \left(\boldsymbol{X}^{j\top} \mathbb{M}^{(-j)} \boldsymbol{X}^j \right)^{-1} \boldsymbol{X}^{j\top} \boldsymbol{U}^{(-j)} = \frac{\boldsymbol{X}^{j\top} \boldsymbol{U}^{(-j)}}{\boldsymbol{X}^{j\top} \mathbb{M}^{(-j)} \boldsymbol{X}^j}$$

is the LSE of the *estimable* parameter β_j of model M (which is its BLUE).

In summary, under the assumptions used to perform derivations above, i.e., while assuming that $\boldsymbol{X}^0 = \mathbf{1}_n$ and for chosen $j \in \{1, \ldots, k-1\}$, the regression coefficient β_j is estimable. Consequently, we define a vector of jth partial residuals of model M as follows.

Definition 11.1 Partial residuals.

A vector of jth partial residuals[1] *of model* M *is a vector*

$$\boldsymbol{U}^{part,j} = \boldsymbol{U} + \widehat{\beta}_j \, \boldsymbol{X}^j = \begin{pmatrix} U_1 + \widehat{\beta}_j X_{1,j} \\ \vdots \\ U_n + \widehat{\beta}_j X_{n,j} \end{pmatrix}.$$

Note. We have

$$\boldsymbol{U}^{part,j} = \boldsymbol{U} + \widehat{\beta}_j \boldsymbol{X}^j = \boldsymbol{Y} - \big(\mathbb{X}\boldsymbol{b} - \widehat{\beta}_j \boldsymbol{X}^j\big) = \boldsymbol{Y} - \big(\widehat{\boldsymbol{Y}} - \widehat{\beta}_j \boldsymbol{X}^j\big).$$

That is, the jth partial residuals are calculated as (classical) residuals where, however, the fitted values subtract a part that corresponds to the column $\boldsymbol{X}^j$ of the model matrix.

Theorem 11.2 Property of partial residuals.

Let $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \le k$, $\boldsymbol{X}^0 = \mathbf{1}_n$, $\boldsymbol{\beta} = \big(\beta_0, \ldots, \beta_{k-1}\big)^\top$. *Let* $j \in \big\{1, \ldots, k-1\big\}$ *be such that* $\mathsf{rank}\big(\mathbb{X}^{(-j)}\big) = r - 1$ *and let* $\widehat{\beta}_j$ *be the LSE of* β_j. *Let us consider a linear model (regression line with covariates* $\boldsymbol{X}^j$*) with*

- *the jth partial residuals* $\boldsymbol{U}^{part,j}$ *as response;*
- *a matrix* $\big(\mathbf{1}_n,\, \boldsymbol{X}^j\big)$ *as the model matrix;*
- *regression coefficients* $\boldsymbol{\gamma}_j = \big(\gamma_{j,0},\, \gamma_{j,1}\big)^\top$.

The least squares estimators of parameters $\gamma_{j,0}$ *and* $\gamma_{j,1}$ *are*

$$\widehat{\gamma}_{j,0} = 0, \quad \widehat{\gamma}_{j,1} = \widehat{\beta}_j.$$

Proof.

- $\boldsymbol{U}^{part,j} = \boldsymbol{U} + \widehat{\beta}_j \boldsymbol{X}^j$.
- Hence $\Big\|\boldsymbol{U}^{part,j} - \gamma_{j,0}\,\mathbf{1}_n - \gamma_{j,1}\,\boldsymbol{X}^j\Big\|^2 = \Big\|\boldsymbol{U} - \big\{\gamma_{j,0}\,\mathbf{1}_n + (\gamma_{j,1} - \widehat{\beta}_j)\boldsymbol{X}^j\big\}\Big\|^2 = 🚲$.
- Since $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$, $\boldsymbol{X}^j \in \mathcal{M}\big(\mathbb{X}\big)$, $\boldsymbol{U} \in \mathcal{M}\big(\mathbb{X}\big)^\perp$, we have

$$🚲 = \big\|\boldsymbol{U}\big\|^2 + \Big\|\gamma_{j,0}\,\mathbf{1}_n + \big(\gamma_{j,1} - \widehat{\beta}_j\big)\,\boldsymbol{X}^j\Big\|^2 \ge \big\|\boldsymbol{U}\big\|^2$$

  with equality if and only if $\gamma_{j,0} = 0 \quad \& \quad \gamma_{j,1} = \widehat{\beta}_j$.

❑

[1] *vektor $j$tých parciálních reziduí*

Shifted partial residuals

Notation *(Response, regressor and partial residuals means).*

Let

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \overline{X}^j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}, \qquad \overline{U}^{part,j} = \frac{1}{n}\sum_{i=1}^{n} U_i^{part,j}.$$

If $\boldsymbol{X}^0 = \mathbf{1}_n$ (model with intercept), we have

$$0 = \sum_{i=1}^{n} U_i = \sum_{i=1}^{n}\Big(U_i^{part,j} + \widehat{\beta}_j X_{i,j}\Big),$$

$$\frac{1}{n}\sum_{i=1}^{n} U_i^{part,j} = \widehat{\beta}_j\Big(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\Big),$$

$$\overline{U}^{part,j} = \widehat{\beta}_j\,\overline{X}^j.$$

Especially for purpose of visualization by plotting the partial residuals against the regressors a shifted partial residuals are sometimes used. Note that this only changes the estimated intercept of the regression line of dependence of partial residuals on the regressor.

Definition 11.2 Shifted partial residuals.

A vector of jth response-mean *partial residuals of model* M *is a vector*

$$\boldsymbol{U}^{part,j,Y} = \boldsymbol{U}^{part,j} + \Big(\overline{Y} - \widehat{\beta}_j\overline{X}^j\Big)\,\mathbf{1}_n.$$

A vector of jth zero-mean *partial residuals of model* M *is a vector*

$$\boldsymbol{U}^{part,j,0} = \boldsymbol{U}^{part,j} - \widehat{\beta}_j\overline{X}^j\,\mathbf{1}_n.$$

Notes.

- A mean of the response-mean partial residuals is the response sample mean $\overline{Y}$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} U_i^{part,j,Y} = \overline{Y}.$$

- A mean of the zero-mean partial residuals is zero, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} U_i^{part,j,0} = 0.$$

The zero-mean partial residuals are calculated by the R function `residuals` with its `type` argument being set to `"partial"`.

***Notes** (Use of partial residuals).*

A vector of partial residuals can be interpreted as a response vector from which we removed a possible effect of all remaining regressors. Hence, dependence of $\boldsymbol{U}^{part,j}$ on $\boldsymbol{X}^j$ shows

- a *net* effect of the jth regressor on the response;
- a *partial* effect of the jth regressor on the response which is *adjusted* for the effect of the remaining regressors.

The partial residuals are then mainly used twofold:

Diagnostic tool. As a (graphical) diagnostic tool, a scatterplot $\big(\boldsymbol{X}^j, \boldsymbol{U}^{part,j}\big)$ is used. In case, the jth regressor is correctly included in the original regression model M, i.e., if no transformation of the regressor $\boldsymbol{X}^j$ is required to achieve $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}\big)$, points in the scatterplot $\big(\boldsymbol{X}^j, \boldsymbol{U}^{part,j}\big)$ should lie along a line.

Visualization. Property that the estimated slope of the regression line in a model $\boldsymbol{U}^{part,j} \sim \boldsymbol{X}^j$ is the same as the jth estimated regression coeffient in the multiple regression model $\boldsymbol{Y} \sim \mathbb{X}$ is also used to visualize dependence of the response of the jth regressor by showing a scatterplot $\big(\boldsymbol{X}^j, \boldsymbol{U}^{part,j}\big)$ equipped by a line with zero intercept and slope equal to $\widehat{\beta}_j$.

11.2.2 Test for linearity of the effect

To examine appropriateness of the linearity of the effect of the jth regressor $\boldsymbol{X}^j$ on the response expectation $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ by a statistical test, we can use a test on submodel (per se, requires additional assumption of normality). Without loss of generality, assume that the jth regressor $\boldsymbol{X}^j$ is the last column of the model matrix $\mathbb{X}$ and denote the remaining non-intercept columns of matrix $\mathbb{X}$ as $\mathbb{X}^0$. That is, assume that

$$\mathbb{X} = \Big(\mathbf{1}_n, \mathbb{X}^0, \boldsymbol{X}^j\Big).$$

Two classical choices of a pair model–submodel being tested in this context are the following.

More general parameterization of the jth regressor

Submodel is the model M with the model matrix $\mathbb{X}$. The (larger) model is model M_g obtained by replacing column $\boldsymbol{X}^j$ in the model matrix $\mathbb{X}$ by a matrix $\mathbb{V}$ such that

$$\boldsymbol{X}^j \in \mathcal{M}\big(\mathbb{V}\big), \qquad \mathsf{rank}\big(\mathbb{V}\big) \geq 2.$$

That is, the model matrices of the submodel and the (larger) model are

$$\begin{aligned} &\text{Submodel } \mathsf{M}: && \Big(\mathbf{1}_n, \mathbb{X}^0, \boldsymbol{X}^j\Big) = \mathbb{X}; \\ &\text{(Larger) model } \mathsf{M}_g: && \Big(\mathbf{1}_n, \mathbb{X}^0, \mathbb{V}\Big). \end{aligned}$$

Classical choices of the matrix $\mathbb{V}$ are such that it corresponds to:

(i) polynomial of degree $d \geq 2$ based on the regressor $\boldsymbol{X}^j$;

(ii) regression spline of degree $d \geq 1$ based on the regressor $\boldsymbol{X}^j$. In this case, $\mathbf{1}_n \in \mathbb{V}$ and hence for practical calculations, the larger model M_g is usually estimated while using a model matrix

$$\big(\mathbb{X}^0, \mathbb{V}\big)$$

that does not explicitely include the intercept term which is included implicitely.

End of Lecture #21 (07/12/2016)

Start of Lecture #22 (08/12/2016)

Categorization of the jth regressor

Let $-\infty < x_j^{low} < x_j^{upp} < \infty$ be chosen such that interval $\big(x_j^{low}, x_j^{upp}\big)$ covers the values $X_{1,j}, \ldots, X_{n,j}$ of the jth regressor. That is,

$$x_j^{low} < \min_i X_{i,j}, \qquad \max_i X_{i,j} < x_j^{upp}.$$

Let $\mathcal{I}_1, \ldots, \mathcal{I}_H$ be $H > 1$ subintervals of $\big(x_j^{low}, x_j^{upp}\big]$ based on a grid

$$x_j^{low} < \lambda_1 < \cdots < \lambda_{H-1} < x_j^{upp}.$$

Let $x_h \in \mathcal{I}_h$, $h = 1, \ldots, H$, be chosen representative values for each of the subintervals $\mathcal{I}_1, \ldots, \mathcal{I}_H$ (e.g., their midpoints) and let

$$\boldsymbol{X}^{j,cut} = \big(X_1^{j,cut}, \ldots, X_n^{j,cut}\big)^\top$$

be obtained by categorization of the jth regressor using the division $\mathcal{I}_1, \ldots, \mathcal{I}_H$ and representatives $x_1, \ldots, x_H$, i.e., ($i = 1, \ldots, n$):

$$X_i^{j,cut} = x_h \quad \equiv \quad X_i^j \in \mathcal{I}_h, \qquad h = 1, \ldots, H.$$

In this way, we obtained a categorical ordinal regressor $\boldsymbol{X}^{j,cut}$ whose values $x_1, \ldots, x_H$, can be considered as collapsed values of the original regressor $\boldsymbol{X}^j$. Consequently, if linearity with respect to the original regressor $\boldsymbol{X}^j$ holds then it also does (approximately, depending on chosen division $\mathcal{I}_1, \ldots, \mathcal{I}_H$ and the representatives $x_1, \ldots, x_H$) with respect to the ordinal categorical regressor $\boldsymbol{X}^{j,cut}$ if this is viewed as numeric one.

Let $\mathbb{V}$ be an $n \times (H-1)$ model matrix corresponding to some (pseudo)contrast parameterization of the covariate $\boldsymbol{X}^{j,cut}$ if this is viewed as categorical with H levels. We have

$$\boldsymbol{X}^{j,cut} \in \mathcal{M}\big(\mathbb{V}\big),$$

and test for linearity of the jth regressor is obtained by considering the following model matrices in the submodel and the (larger) model:

$$\begin{aligned} &\text{Submodel } \mathsf{M}: && \Big(\mathbf{1}_n, \mathbb{X}^0, \boldsymbol{X}^{j,cut}\Big); \\ &\text{(Larger) model } \mathsf{M}_g: && \Big(\mathbf{1}_n, \mathbb{X}^0, \mathbb{V}\Big). \end{aligned}$$

Additional insight concerning the correct inclusion of the jth regressor can be obtained by using the orthonormal polynomial contrasts (Section 7.4.4) in place of the $\mathbb{V}$ matrix.

Drawback of tests for linearity of the effect

Remind that hypothesis of linearity of the effect of the jth regressor always forms the null hypothesis of the proposed submodel tests. Hence we are only able to confirm non-linearity of the effect (if the submodel is rejected) but are never able to confirm linearity.

11.3 Homoscedasticity

We are again assuming a linear model

$$\mathsf{M}\colon \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big),$$

where the error terms $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$ satisfy (Lemma 1.2):

$$\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \mathbb{E}\big(\boldsymbol{\varepsilon}\big) = \mathbf{0}_n, \quad \mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \mathsf{var}\big(\boldsymbol{\varepsilon}\big) = \sigma^2 \mathbf{I}_n.$$

The assumption (A2) of homoscedasticity is, in particular,

$$\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \sigma^2\, \mathbf{I}_n, \qquad \mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \sigma^2\, \mathbf{I}_n, \qquad \big(\Longrightarrow \quad \mathsf{var}\big(\boldsymbol{\varepsilon}\big) = \sigma^2 \mathbf{I}_n\big),$$

where σ^2 is unknown but most importantly constant.

11.3.1 Tests of homoscedasticity

Many tests of homoscedasticity can be found in literature. They mostly consider the following null and alternative hypotheses: H_0: $\mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\big) = \text{const}$,
H_1: $\mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\big) = \text{certain function of some factor(s)}$.

A particular test is then sensitive (powerful) to detect heteroscedasticity if this expresses itself such that the conditional variance $\mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\big)$ is the *certain* function of the factor(s) as specified by the alternative hypothesis. The test is possibly weak to detect heteroscedasticity (weak to reject the null hypothesis of homoscedasticity) if heteroscedasticity expresses itself in a different way compared to the considered alternative hypothesis.

11.3.2 Score tests of homoscedasticity

A wide range of tests of homoscedasticity can be derived by assuming a (full-rank) *normal* linear model, basing the alternative hypothesis on a further generalization of a general linear model and then using an (asymptotic) maximum-likelihood theory to derive a testing procedure.

Assumptions.

For the rest of this section, we assume that model M (model under the null hypothesis) is normal of full-rank, i.e.,

$$\mathsf{M}\colon \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big), \quad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k,$$

and an alternative model is a generalization of a general normal linear model

$$\mathsf{M}_{hetero}\colon \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbb{W}^{-1}\big),$$

where

$$\mathbb{W} = \mathsf{diag}(w_1, \ldots, w_n), \qquad w_i^{-1} = \tau(\boldsymbol{\lambda},\, \boldsymbol{\beta},\, \boldsymbol{Z}_i),\ i = 1, \ldots, n,$$

τ is a *known* function of $\boldsymbol{\lambda} \in \mathbb{R}^q$, $\boldsymbol{\beta} \in \mathbb{R}^k$ (regression coefficients), $\boldsymbol{z} \in \mathbb{R}^p$ (covariates) such that

$$\tau(\mathbf{0},\, \boldsymbol{\beta},\, \boldsymbol{z}) = 1, \qquad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^k,\, \boldsymbol{z} \in \mathbb{R}^p.$$

In particular, we have under model M_{hetero}:

$$\mathsf{var}\big(\boldsymbol{Y}_i \,\big|\, \boldsymbol{Z}_i\big) \;=\; \mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\big) \;=\; \sigma^2\, \tau(\boldsymbol{\lambda},\, \boldsymbol{\beta},\, \boldsymbol{Z}_i), \qquad i = 1, \ldots, n.$$

That is, the τ function models the assumed heteroscedasticity.

Model M_{hetero} is then a model with unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, σ^2 which with $\boldsymbol{\lambda} = \mathbf{0}$ simplifies into model M. In other words, model M is a *nested*[2] model of model M_{hetero} and a test of homoscedasticity corresponds to testing

$$\begin{aligned} \mathrm{H}_0 &: \quad \boldsymbol{\lambda} = \mathbf{0}, \\ \mathrm{H}_1 &: \quad \boldsymbol{\lambda} \neq \mathbf{0}. \end{aligned} \tag{11.2}$$

Having assumed normality, both models M and M_{hetero} are fully parametric models and a standard (asymptotic) maximum-likelihood theory can now be used to derive a test of (11.2). A family of *score* tests based on specific choices of the weight function τ is derived by Cook and Weisberg (1983).

Breusch-Pagan test

A particular score test of homoscedasticity was also derived by Breusch and Pagan (1979) who consider the following weight function ($\boldsymbol{x} = t_X(\boldsymbol{z})$ is a transformation of the original covariates that determines the regressors of model M).

$$\tau(\lambda, \boldsymbol{\beta}, \boldsymbol{z}) = \tau(\lambda, \boldsymbol{\beta}, \boldsymbol{x}) = \exp\bigl(\lambda\, \boldsymbol{x}^\top \boldsymbol{\beta}\bigr).$$

That is, under the heteroscedastic model, for $i = 1, \ldots, n$,

$$\mathsf{var}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr) = \mathsf{var}\bigl(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\bigr) = \sigma^2 \, \exp\bigl(\lambda\, \boldsymbol{X}_i^\top \boldsymbol{\beta}\bigr) = \sigma^2 \, \exp\Bigl(\lambda\, \mathbb{E}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr)\Bigr), \tag{11.3}$$

and the test of homoscedasticity is testing

$$\begin{aligned} \mathrm{H}_0 &: \quad \lambda = 0, \\ \mathrm{H}_1 &: \quad \lambda \neq 0. \end{aligned}$$

It is seen from the model (11.3) that the Breusch-Pagan test is sensitive (powerful to detect heteroscedasticity) if the residual variance is a *monotone* function of the response expectation.

Note *(One-sided tests of homoscedasticity).*

In practical situations, if it can be assumed that the residual variance is possibly a monotone function of the response expectation then it can mostly be also assumed that it is its *increasing* function. A more powerful test of homoscedasticity is then obtained by considering the one-sided alternative

$$\mathrm{H}_1 : \ \lambda > 0.$$

Analogously, a test that is sensitive towards alternative of a residual variance which *decreases* with the response expectation is obtained by considering the alternative $\mathrm{H}_1 : \ \lambda < 0$.

Note *(Koenker's studentized Breusch-Pagan test).*

The original Breusch-Pagan test is derived using standard maximum-likelihood theory while starting from assumption of a *normal* linear model. It has been shown in the literature that the test is not robust towards non-normality. For this reason, Koenker (1981) derived a slightly modified version of the Breusch-Pagan test which is robust towards non-normality. It is usually referred to as (Koenker's) studentized Breusch-Pagan test and its use is preferred to the original test.

[2] *vnořený*

Linear dependence on the regressors

Let $t_W : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ be a given transformation, $\boldsymbol{w} := t_W(\boldsymbol{z})$, $\boldsymbol{W}_i = t_W(\boldsymbol{Z}_i)$, $i = 1, \ldots, n$. The following choice of the weight function can be considered:

$$\tau(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{z}) = \tau(\boldsymbol{\lambda}, \boldsymbol{w}) = \exp\bigl(\boldsymbol{\lambda}^\top \boldsymbol{w}\bigr).$$

That is, under the heteroscedastic model, for $i = 1, \ldots, n$,

$$\mathsf{var}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr) \;=\; \mathsf{var}\bigl(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\bigr) \;=\; \sigma^2 \exp\bigl(\boldsymbol{\lambda}^\top \boldsymbol{W}_i\bigr).$$

On a log-scale:

$$\log\Bigl(\mathsf{var}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr)\Bigr) = \underbrace{\log(\sigma^2)}_{\lambda_0} + \boldsymbol{\lambda}^\top \boldsymbol{W}_i.$$

In other words, the residual variance follows on a log-scale a linear model with regressors given by vectors $\boldsymbol{W}_i$.

If t_W is a univariate transformation leading to $w = t_W(\boldsymbol{z})$, one-sided alternatives are again possible reflecting assumption that under heteroscedasticity, the residual variance increases/decreases with a value of $W = t_W(Z)$. The most common use is then such that $t_W(\boldsymbol{z})$ and related values of $W_1 = t_W(\boldsymbol{Z}_1)$, …, $W_n = t_W(\boldsymbol{Z}_n)$ correspond to one of the (non-intercept) regressors from either the model matrix $\mathbb{X}$ (regressors included in the model), or from the matrix $\mathbb{V}$ that contains regressors currently not included in the model. The corresponding score test of homoscedasticity then examines whether the residual variance changes/increases/decreases (depending on chosen alternative) with that regressor.

Note *(Score tests of homoscedasticity in the R software).*

In the R software, the score tests of homoscedasticity are provided by functions:

(i) `ncvTest` (abbreviation for a "non-constant variance test") from package `car`;

(ii) `bptest` from package `lmtest`.

The Koenker's studentized variant of the test is only possible with the `bptest` function.

11.3.3 Some other tests of homoscedasticity

Some other tests of homoscedasticity that can be encountered in practice include the following

Goldfeld-Quandt test is an adaptation of a classical F-test of equality of the variances of the two independent samples into a regression context proposed by Goldfeld and Quandt (1965). It is applicable in linear models with both numeric and categorical covariates and under the alternative, heteroscedasticity is expressed by a monotone dependence of the residual variance on a prespecified ordering of the observations.

G-sample tests of homoscedasticity are tests applicable for linear models with only categorical covariates (ANOVA models). They require repeated observations for each combination of values of the covariates and basically test equality of variances of G independent random samples. The most common tests of this type include:

Bartlett test by Bartlett (1937) which, however, is quite sensitive towards non-normality and hence its use is not recommended. It is implemented in the R function `bartlett.test`;

Levene test by Levene (1960), implemented in the R function `leveneTest` from package `car` or in the R function `levene.test` from package `lawstat`;

Brown-Forsythe test by Brown and Forsythe (1974) which is a robustified version of the Levene test and is implemented in the R function `levene.test` from package `lawstat`;

Fligner-Killeen test by Fligner and Killeen (1976) which is implemented in the R function `fligner.test`.

11.4 Normality

In this section, we are assuming a *normal* linear model

$$\mathsf{M}\colon \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}) = r,$$

where the error terms $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} = \big(\varepsilon_1,\, \ldots,\, \varepsilon_n\big)^\top$ satisfy (Lemma 3.1):

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma^2), \quad i = 1, \ldots, n. \tag{11.4}$$

Our interest now lies in verifying assumption (A4) of normality of the error terms ε_i, $i = 1, \ldots, n$. Let us remind our standard notation needed in this section:

(i) Hat matrix (projection matrix into the regression space $\mathcal{M}\big(\mathbb{X}\big)$):

$$\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top = \big(h_{i,t}\big)_{i,t=1,\ldots,n};$$

(ii) Projection matrix into the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$:

$$\mathbb{M} = \mathbf{I}_n - \mathbb{H} = \big(m_{i,t}\big)_{i,t=1,\ldots,n};$$

(iii) Residuals: $\boldsymbol{U} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \mathbb{M}\boldsymbol{Y} = \big(U_1, \ldots, U_n\big)^\top$;

(iv) Residual sum of squares: $\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2$;

(v) Residual mean square: $\mathsf{MS}_e = \frac{1}{n-r}\,\mathsf{SS}_e$.

(vi) Standardized residuals: $\boldsymbol{U}^{std} = \big(U_1^{std}, \ldots, U_n^{std}\big)^\top$, where

$$U_i^{std} = \frac{U_i}{\sqrt{\mathsf{MS}_e\, m_{i,i}}}, \qquad i = 1, \ldots, n \quad (\text{if } m_{i,i} > 0).$$

Notes. If the normal linear model (11.4) holds then Theorems 3.2 and 4.1 provide:

(i) For (raw) residuals:

$$\boldsymbol{U} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\,\mathbb{M}\big).$$

That is, the (raw) residuals follow also a normal distribution, nevertheless, the variances of the individual residuals $U_1, \ldots, U_n$ differ (a diagonal of the projection matrix $\mathbb{M}$ is not necessarily constant). On top of that, the residuals are not necessarily independent (the projection matrix $\mathbb{M}$ is not necessarily a diagonal matrix).

(ii) For standardized residuals (if $m_{i,i} > 0$ for all $i = 1, \ldots, n$, which is always the case in a full-rank model):

$$\mathbb{E}\big(U_i^{std} \,\big|\, \mathbb{Z}\big) = 0, \qquad \mathsf{var}\big(U_i^{std} \,\big|\, \mathbb{Z}\big) = 1, \qquad i = 1, \ldots, n.$$

That is, the standardized residuals have the same mean and also the variance but are neither necessarily normally distributed nor necessarily independent.

In summary, in a normal linear model, neither the raw residuals, nor standardized residuals form a random sample (a set of i.i.d. random variables) from a normal distribution.

11.4.1 Tests of normality

There exist formal tests of the null hypothesis on a normality of the error terms:

$$\mathrm{H}_0: \text{ distribution of } \varepsilon_1, \ldots, \varepsilon_n \text{ is normal,} \tag{11.5}$$

where a distribution of the test statistic is exactly known under the null hypothesis of normality. Nevertheless, those tests have quite a low power and hence are only rarely used in practice.

In practice, approximate approaches are used that apply standard tests of normality on either the raw residuals $\boldsymbol{U}$ or the standardized residuals $\boldsymbol{U}^{std}$ (both of them, under the null hypothesis (11.5), do not form a random sample from the normal distribution). Several empirical studies showed that such approaches maintain quite well a significance level of the test on a requested value. At the same time, they mostly recommend to use the raw residuals $\boldsymbol{U}$ rather than the standardized residuals $\boldsymbol{U}^{std}$.

Classical tests of normality include the following:

Shapiro-Wilk test implemented in the R function `shapiro.test`.

Lilliefors test implemented in the R function `lillie.test` from package `nortest`.

Anderson-Darling test implemented in the R function `ad.test` from package `nortest`.

11.5 Uncorrelated errors

In this section, we are again assuming a (not necessarily normal) linear model

$$\mathsf{M}: \quad \boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),$$

where the error terms $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$ satisfy (Lemma 1.2):

$$\mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) = \mathbb{E}\big(\boldsymbol{\varepsilon}\big) = \mathbf{0}_n, \quad \mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{X}\big) = \mathsf{var}\big(\boldsymbol{\varepsilon}\big) = \sigma^2\mathbf{I}_n.$$

The assumption (A3) is, in particular,

$$\mathsf{cov}\big(\varepsilon_i,\, \varepsilon_l \,\big|\, \mathbb{X}\big) = 0,\ i \neq l \qquad \big(\Longrightarrow \quad \mathsf{cov}\big(\varepsilon_i,\, \varepsilon_l\big) = 0,\ i \neq l\big). \tag{11.6}$$

Our interest now lies in verifying assumption (A3) of whether the error terms ε_i, $i = 1, \ldots, n$, are (conditionally) uncorrelated.

The fact that errors are (conditionally) uncorrelated often follows from a design of the study/data collection (measurements on independently behaving units, ...) and then there is no need to check this assumption. Situation when uncorrelated errors cannot be taken for granted is if the observations are obtained *sequentially*. Typical examples are

(i) time series (time does not have to be a covariate of the model) which may lead to so called *serial depedence* among the error terms of the linear model;

(ii) repeated measurements performed using one measurement unit or on one subject.

In the following, we introduce a classical procedure that is used to test a null hypothesis of uncorrelated errors against alternative of serial dependence expressed by the first order autoregressive process.

11.5.1 Durbin-Watson test

Assumptions.

It is assumed that the *ordering* of the observations expressed by their indeces $1, \ldots, n$, has a practical meaning and may induce depedence between the error terms $\varepsilon_1, \ldots, \varepsilon_n$ of the model.

Model M can also be written as

$$\begin{aligned}
\mathsf{M}: \quad & Y_i = \boldsymbol{X}_i^\top\boldsymbol{\beta} + \varepsilon_i, && i = 1, \ldots, n,\\
& \mathbb{E}\big(\varepsilon_i \,\big|\, \mathbb{X}\big) = 0, \quad \mathsf{var}\big(\varepsilon_i \,\big|\, \mathbb{X}\big) = \sigma^2, && i = 1, \ldots, n,\\
& \mathsf{cor}\big(\varepsilon_i,\, \varepsilon_l \,\big|\, \mathbb{X}\big) = 0, && i \neq l.
\end{aligned} \tag{11.7}$$

One of the simplest stochastic processes that capture a certain form of serial dependence is the first order autoregressive process AR(1). Assuming this for the error terms $\varepsilon_1, \ldots, \varepsilon_n$ of the linear model (11.7) leads to a more general model

$$\begin{aligned}
\mathsf{M}_{AR}: \quad & Y_i = \boldsymbol{X}_i^\top\boldsymbol{\beta} + \varepsilon_i, && i = 1, \ldots, n,\\
& \varepsilon_1 = \eta_1, \quad \varepsilon_i = \varrho\,\varepsilon_{i-1} + \eta_i, && i = 2, \ldots, n,\\
& \mathbb{E}\big(\eta_i \,\big|\, \mathbb{X}\big) = 0, \quad \mathsf{var}\big(\eta_i \,\big|\, \mathbb{X}\big) = \sigma^2, && i = 1, \ldots, n,\\
& \mathsf{cor}\big(\eta_i,\, \eta_l \,\big|\, \mathbb{X}\big) = 0, && i \neq l,
\end{aligned} \tag{11.8}$$

where $-1 < \varrho < 1$ is additional unknown parameter of the model.

Notes. It has been shown in the course *Stochastic Processes 2 (NMSA409)*:

- $\varepsilon_1, \ldots, \varepsilon_n$ is a stacionary process (given $\mathbb{X}$) if and only if $-1 < \varrho < 1$.
- For each $m \geq 0$: $\mathsf{cor}\big(\varepsilon_i, \varepsilon_{i-m} \,\big|\, \mathbb{X}\big) = \varrho^m$, $i = m+1, \ldots, n$. In particular
$$\varrho = \mathsf{cor}\big(\varepsilon_i, \varepsilon_{i-1} \,\big|\, \mathbb{X}\big), \qquad i = 2, \ldots, n.$$

Test of uncorrelated errors in model M can be now be based on testing

$$\begin{aligned} \mathrm{H}_0\text{:} &\quad \varrho = 0, \\ \mathrm{H}_1\text{:} &\quad \varrho \neq 0 \end{aligned}$$

in model M_{AR}. Since positive autocorrelation ($\varrho > 0$) is more common in practice, one-sided tests (with H_1: $\varrho > 0$) are used frequently as well.

Let $\boldsymbol{U} = \big(U_1, \ldots, U_n\big)^\top$ be residuals from model M which corresponds to the null hypothesis. A test statistic proposed by Durbin and Watson (1950, 1951, 1971) takes a form

$$DW = \frac{\sum_{i=2}^{n} (U_i - U_{i-1})^2}{\sum_{i=1}^{n} U_i^2}.$$

A testing procedure is based on observing that a statistic DW is approximately equal to $2\,(1 - \widehat{\varrho})$, where $\widehat{\varrho}$ is an estimator of the autoregression parameter ϱ from model M_{AR}.

Calculations.

First remember that

$$\mathbb{E}\big(U_i \,\big|\, \mathbb{X}\big) = 0, \qquad i = 1, \ldots, n,$$

and this property is maintained even if the error terms of the model are not uncorrelated (see process of the proof of Theorem 2.3).

As residuals can be considered as predictions of the error terms $\varepsilon_1, \ldots, \varepsilon_n$, a suitable estimator of their (conditional) covariance of lag 1 is

$$\widehat{\sigma}_{1,2} = \widehat{\mathsf{cov}}\big(\varepsilon_l, \varepsilon_{l-1} \,\big|\, \mathbb{X}\big) = \frac{1}{n-1} \sum_{i=2}^{n} U_i\, U_{i-1}.$$

Similarly, three possible estimators of the (conditional) variance σ^2 of the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are

$$\widehat{\sigma}^2 = \widehat{\mathsf{var}}\big(\varepsilon_l \,\big|\, \mathbb{X}\big) = \frac{1}{n-1} \sum_{i=1}^{n-1} U_i^2 \quad \text{or} \quad \frac{1}{n-1} \sum_{i=2}^{n} U_i^2 \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^{n} U_i^2.$$

Then,

$$\begin{aligned} DW &= \frac{\sum_{i=2}^{n} (U_i - U_{i-1})^2}{\sum_{i=1}^{n} U_i^2} = \frac{\sum_{i=2}^{n} U_i^2 \,+\, \sum_{i=2}^{n} U_{i-1}^2 \,-\, 2\,\sum_{i=2}^{n} U_i\, U_{i-1}}{\sum_{i=1}^{n} U_i^2} \\ &\approx \frac{\widehat{\sigma}^2 \,+\, \widehat{\sigma}^2 \,-\, 2\,\widehat{\sigma}_{1,2}}{\widehat{\sigma}^2} = 2\left(1 - \frac{\widehat{\sigma}_{1,2}}{\widehat{\sigma}^2}\right) \\ &= 2\,\big(1 - \widehat{\varrho}\big). \end{aligned}$$

Use of the test statistic DW for tests of $H_0: \varrho = 0$ is complicated by the fact that distribution of DW under the null hypothesis depends on the model matrix $\mathbb{X}$. It is hence not possible to derive (and tabulate) critical values in full generality. In practice, two approaches are used to calculate approximate critical values and p-values:

(i) Numerical algorithm of Farebrother (1980, 1984) which is implemented in the R function `dwtest` from package `lmtest`;

(ii) General simulation method *bootstrap* (introduced by Efron, 1979) whose use for the Durbin-Watson test is implemented in the R function `durbinWatsonTest` from package `car`. For general principles of the bootstrap method, see the course *Modern Statistical Methods (NMST434)*.

11.6 Transformation of response

Especially in situations when *homoscedasticity* and/or *normality* does not hold, it is often possible to achieve a linear model where both those assumptions are fulfilled by a suitable (non-linear) transformation $t : \mathbb{R} \longrightarrow \mathbb{R}$ of the response. That is, it is worked with a normal linear model

$$\begin{aligned} \boldsymbol{Y}^\star \,|\, \mathbb{X} &\sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\, \mathbf{I}_n\big), \\ \boldsymbol{Y}^\star &= \big(t(Y_1),\, \ldots,\, t(Y_n)\big)^\top, \end{aligned} \tag{11.9}$$

where it is already assumed that both homoscedasticity and normality hold. That is, the elements of the error terms vector

$$\big(\varepsilon_1,\, \ldots,\, \varepsilon_n\big)^\top = \boldsymbol{\varepsilon} \;=\; \boldsymbol{Y}^\star - \mathbb{X}\boldsymbol{\beta} = \big(t(Y_1) - \boldsymbol{X}_1^\top\boldsymbol{\beta},\; \ldots,\; t(Y_n) - \boldsymbol{X}_n^\top\boldsymbol{\beta}\big)^\top$$

are, given $\mathbb{X}$, independent and $\mathcal{N}(0,\, \sigma^2)$ distributed (marginally, they are i.i.d. $\mathcal{N}(0,\, \sigma^2)$ distributed). Disadvantage of a model with transformed response is that the corresponding regression function $m(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}$ provides a model for expectation of the transformed response and not of the original response, i.e., for $\boldsymbol{x} \in \mathcal{X}$ (sample space of the regressors):

$$m(\boldsymbol{x}) = \mathbb{E}\big(t(Y) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big) \quad \neq \quad t\Big(\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)\Big),$$

unless the transformation t is a linear function. Similarly, regression coefficients have now interpretation of an expected change of the *transformed* response $t(Y)$ related to a unity increase of the regressor.

11.6.1 Prediction based on a model with transformed response

Nevertheless, the above mentioned interpretational issue is not a problem in a situation when *prediction* of a new value of the response Y_{new}, given $\boldsymbol{X}_{new} = \boldsymbol{x}_{new}$, is of interest. If this is the case, we can base the prediction on the model (11.9) for the transormed response. In the following, we assume that t is strictly increasing, nevertheless, the procedure can be adjusted for decreasing or even non-monotone t as well:

- Construct a prediction $\widehat{Y}^\star_{new}$ and a $(1-\alpha)\,100\%$ prediction interval $\big(\widehat{Y}^{\star,L}_{new},\; \widehat{Y}^{\star,U}_{new}\big)$ for $Y^\star_{new} = t(Y_{new})$ based on the model (11.9).
- Trivially, an interval
$$\big(\widehat{Y}^{L}_{new},\; \widehat{Y}^{U}_{new}\big) = \Big(t^{-1}\big(\widehat{Y}^{\star,L}_{new}\big),\; t^{-1}\big(\widehat{Y}^{\star,U}_{new}\big)\Big) \tag{11.10}$$
covers a value of Y_{new} with a probability of $1-\alpha$.
- A value $\widehat{Y}_{new} = t^{-1}\big(\widehat{Y}^\star_{new}\big)$ lies inside the prediction interval (11.10) and can be considered as a point prediction of Y_{new}. Only note that the prediction interval $\big(\widehat{Y}^{L}_{new},\; \widehat{Y}^{U}_{new}\big)$ is not necessarily centered around a value of $\widehat{Y}_{new}$.

11.6.2 Log-normal model

Suitably interpretable model is obtained if the response is logarithmically transformed. Suppose that the following model (normal linear model for log-transformed response) holds:

$$\begin{aligned} \log(Y_i) = \boldsymbol{X}_i^\top\boldsymbol{\beta} + \varepsilon_i, \qquad & i = 1, \ldots, n, \\ \varepsilon_i \,|\, \mathbb{X} \overset{\text{indep.}}{\sim}\; & \mathcal{N}\big(0,\, \sigma^2\big), \end{aligned} \tag{11.11}$$

which also implies $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\, \sigma^2\big)$. We then have

$$Y_i = \exp\big(\boldsymbol{X}_i^\top \boldsymbol{\beta}\big)\, \eta_i, \qquad i = 1, \ldots, n,$$
$$\eta_i \,\big|\, \mathbb{X} \overset{\text{indep.}}{\sim} \mathcal{LN}\big(0,\, \sigma^2\big),$$

which also implies $\eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{LN}\big(0,\, \sigma^2\big)$, where $\mathcal{LN}(0,\, \sigma^2)$ denotes a log-normal distribution with location parameter 0 and a scale parameter σ. That is, under validity of the model (11.11) for the log-transformed response, errors in a model for the original response are combined *multiplicatively* with the regression function.

We can easily calculate the first two moments of the log-normal distribution which provides (for $i = 1, \ldots, n$),

$$M := \mathbb{E}\big(\eta_i\big) = \mathbb{E}\big(\eta_i \,\big|\, \mathbb{X}\big) = \exp\Big(\frac{\sigma^2}{2}\Big) \qquad > 1 \;\; \text{(with } \sigma^2 > 0\text{)},$$
$$V := \mathsf{var}\big(\eta_i\big) = \mathsf{var}\big(\eta_i \,\big|\, \mathbb{X}\big) = \big\{\exp(\sigma^2) - 1\big\} \exp(\sigma^2).$$

Hence, for $\boldsymbol{x} \in \mathcal{X}$:

$$\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big) = M\, \exp\big(\boldsymbol{x}^\top \boldsymbol{\beta}\big),$$
$$\mathsf{var}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big) = V\, \exp\big(2\, \boldsymbol{x}^\top \boldsymbol{\beta}\big) = V \cdot \left(\frac{\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)}{M}\right)^2. \tag{11.12}$$

A log-normal model (11.11) is thus suitable in two typical situations that cause non-normality and/or heteroscedasticity of a linear model for the original response Y:

(i) a conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ is *skewed*. If this is the case, the log-normal distribution which is skewed as well may provide a satisfactory model for this distribution.

(ii) a conditional variance $\mathsf{var}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)$ increases with a conditional expectation $\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)$. This feature is captured by the log-normal model as shown by (11.12). Indeed, under the log-normal model, $\mathsf{var}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)$ increases with $\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)$. It is then said that the logarithmic transformation *stabilizes* the variance.

Interpretation of regression coefficients

With a log-normal model (11.11), the (non-intercept) regression coefficients have the following interpretation. Let for $j \in \{1, \ldots, k-1\}$,

$$\boldsymbol{x} = \big(x_0, \ldots, x_j \ldots, x_{k-1}\big)^\top \in \mathcal{X}, \quad \text{and} \quad \boldsymbol{x}^{j(+1)} := \big(x_0, \ldots, x_j + 1 \ldots, x_{k-1}\big)^\top \in \mathcal{X},$$

and suppose that $\boldsymbol{\beta} = \big(\beta_0, \ldots, \beta_{k-1}\big)^\top$ We then have

$$\frac{\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}^{j(+1)}\big)}{\mathbb{E}\big(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big)} = \frac{M\, \exp\big(\boldsymbol{x}^{j(+1)^\top} \boldsymbol{\beta}\big)}{M\, \exp\big(\boldsymbol{x}^\top \boldsymbol{\beta}\big)} = \exp(\beta_j).$$

Notes.

- If ANOVA linear model with log-transformed response is fitted, estimated differences between the group means of the log-response are equal to estimated log-*ratios* between the group means of the original response.
- If a linear model with logarithmically transformed response if fitted, estimated regression coefficients, estimates of estimable parameters etc. and corresponding confidence intervals are often reported back-transformed (exponentiated) due to above interpretation.

Evaluation of impact of the regressors on response

Evaluation of impact of the regressors on response requires necessity to perform *statistical tests* on regression coefficients or estimable parameters of a linear model. *Homoscedasticity* and for small samples also *normality* are needed to be able to use standard t- or F-tests. Both homoscedasticity and normality can be achieved by a log transformation of the response. Consequently performed statistical tests still have a reasonable practical interpretation as tests on ratios of two expectations of the (original) response.

Chapter 12

Consequences of a Problematic Regression Space

As in Chapter 11, we assume that data are represented by n random vectors $\left(Y_i, \boldsymbol{Z}_i^\top\right)^\top$, $\boldsymbol{Z}_i = \left(Z_{i,1}, \ldots, Z_{i,p}\right)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$ $i = 1, \ldots, n$. As usual, let $\boldsymbol{Y} = \left(Y_1, \ldots, Y_n\right)^\top$ and let $\mathbb{Z}_{n\times p}$ denote a matrix with covariate vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ in its rows. Finally, let $\boldsymbol{X}_i$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = \boldsymbol{t}_X(\boldsymbol{Z}_i)$ for some transformation $\boldsymbol{t}_X : \mathbb{R}^p \longrightarrow \mathbb{R}^k$, be the regressors that give rise to the model matrix

$$\mathbb{X}_{n\times k} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \left(\boldsymbol{X}^0, \ldots, \boldsymbol{X}^{k-1}\right).$$

It will be assumed that $\boldsymbol{X}^0 = \left(1, \ldots, 1\right)^\top$ (almost surely) leading to the model matrix

$$\mathbb{X}_{n\times k} = \left(\mathbf{1}_n, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^{k-1}\right),$$

with explicitely included intercept term.

Primarily, we will assume that the model matrix $\mathbb{X}$ is sufficient to be able to assume that $\mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{Z}\right) = \mathbb{E}\left(\boldsymbol{Y} \,\middle|\, \mathbb{X}\right) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{k-1}\right)^\top \in \mathbb{R}^k$. That is, we will arrive from assuming

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

It will finally be assumed in the whole chapter that the model matrix $\mathbb{X}$ is of full rank, i.e.,

$$\mathsf{rank}\left(\mathbb{X}\right) = k < n.$$

12.1 Multicollinearity

A principal assumption of any regression model is correct specification of the regression function. While assuming a linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, this means that $\mathbb{E}\big(\boldsymbol{Y} \,|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}\big)$. To guarantee this, it seems to be optimal to choose the regression space $\mathcal{M}\big(\mathbb{X}\big)$ as rich as possible. In other words, if many covariates are available, it seems optimal to include a high number k of columns in the model matrix $\mathbb{X}$. Nevertheless, as we show in this section, this approach bears certain complications.

12.1.1 Singular value decomposition of a model matrix

We are assuming $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k < n$. As was shown in the course *Fundamentals of Numerical Mathematics (NMNM201),* the matrix $\mathbb{X}$ can be decomposed as

$$\mathbb{X} = \mathbb{U}\,\mathbb{D}\,\mathbb{V}^\top = \sum_{j=0}^{k-1} d_j\,\boldsymbol{u}_j\,\boldsymbol{v}_j^\top, \qquad \mathbb{D} = \mathsf{diag}(d_0,\,\ldots,\,d_{k-1}),$$

where

- $\mathbb{U}_{n\times k} = \big(\boldsymbol{u}_0,\,\ldots,\,\boldsymbol{u}_{k-1}\big)$ are the first k orthonormal eigenvectors of the $n \times n$ matrix $\mathbb{X}\mathbb{X}^\top$.
- $\mathbb{V}_{k\times k} = \big(\boldsymbol{v}_0,\,\ldots,\,\boldsymbol{v}_{k-1}\big)$ are (all) orthonormal eigenvectors of the $k \times k$ (invertible) matrix $\mathbb{X}^\top\mathbb{X}$.
- $d_j = \sqrt{\lambda_j}, \quad j = 0, \ldots, k-1$, where $\lambda_0 \geq \cdots \geq \lambda_{k-1} > 0$ are
  - the first k eigenvalues of the matrix $\mathbb{X}\mathbb{X}^\top$;
  - (all) eigenvalues of the matrix $\mathbb{X}^\top\mathbb{X}$, i.e.,

$$\begin{aligned}\mathbb{X}^\top\mathbb{X} &= \sum_{j=0}^{k-1} \lambda_j\,\boldsymbol{v}_j\,\boldsymbol{v}_j^\top = \mathbb{V}\,\boldsymbol{\Lambda}\,\mathbb{V}^\top, \qquad \boldsymbol{\Lambda} = \mathsf{diag}(\lambda_0,\,\ldots,\,\lambda_{k-1}) \\ &= \sum_{j=0}^{k-1} d_j^2\,\boldsymbol{v}_j\,\boldsymbol{v}_j^\top = \mathbb{V}\,\mathbb{D}^2\,\mathbb{V}^\top.\end{aligned}$$

The numbers $d_0 \geq \cdots \geq d_{k-1} > 0$ are called *singular values*[1] of the matrix $\mathbb{X}$. We then have

$$\begin{aligned}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1} &= \sum_{j=0}^{k-1} \frac{1}{d_j^2}\,\boldsymbol{v}_j\,\boldsymbol{v}_j^\top = \mathbb{V}\,\mathbb{D}^{-2}\,\mathbb{V}^\top, \\ \mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\} &= \sum_{j=0}^{k-1} \frac{1}{d_j^2}.\end{aligned} \tag{12.1}$$

Note *(Moore-Penrose pseudoinverse of the matrix $\mathbb{X}^\top\mathbb{X}$).*

The singular value decomposition of the model matrix $\mathbb{X}$ provides also a way to calculate the Moore-Penrose pseudoinverse of the matrix $\mathbb{X}^\top\mathbb{X}$ if $\mathbb{X}$ is of less-than-full rank. If $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r < k$, then $d_0 \geq \cdots \geq d_{r-1} > d_r = \cdots = d_{k-1} = 0$. The Moore-Penrose pseudoinverse of $\mathbb{X}^\top\mathbb{X}$ is obtained as

$$\big(\mathbb{X}^\top\mathbb{X}\big)^{+} = \sum_{j=0}^{r-1} \frac{1}{d_j^2}\,\boldsymbol{v}_j\,\boldsymbol{v}_j^\top.$$

[1] *singulární hodnoty*

12.1.2 Multicollinearity and its impact on precision of the LSE

It is seen from (12.1) that with $d_{k-1} \longrightarrow 0$:

(i) the matrix $\mathbb{X}^\top\mathbb{X}$ tends to a singular matrix, i.e., the columns of the model matrix $\mathbb{X}$ tend to being *linearly dependent*;

(ii) $\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\} \longrightarrow \infty$.

Situation when the columns of the (full-rank) model matrix $\mathbb{X}$ are close to being linearly dependent is referred to as *multicollinearity*.

If a linear model $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}_{n\times k}) = k$ is assumed, then we know from Gauss-Markov theorem that

(i) The fitted values $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1,\, \ldots,\, \widehat{Y}_n\big)^\top = \mathbb{H}\boldsymbol{Y}$, where $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top$, is the best linear unbiased estimator (BLUE) of a vector parameter $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ with

$$\mathsf{var}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{Z}\big) = \sigma^2\,\mathbb{H};$$

(ii) The least squares estimator $\widehat{\boldsymbol{\beta}} = \big(\widehat{\beta}_0,\, \ldots,\, \widehat{\beta}_{k-1}\big)^\top = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\,\mathbb{X}^\top\boldsymbol{Y}$ is the BLUE of a vector of regression coefficients $\boldsymbol{\beta}$ with

$$\mathsf{var}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\big) = \sigma^2\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}.$$

It then follows

$$\sum_{i=1}^{n}\mathsf{var}\big(\widehat{Y}_i \,\big|\, \mathbb{Z}\big) = \mathsf{tr}\Big\{\mathsf{var}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{Z}\big)\Big\} = \mathsf{tr}\big(\sigma^2\,\mathbb{H}\big) = \sigma^2\,\mathsf{tr}(\mathbb{H}) = \sigma^2\,k,$$

$$\sum_{j=0}^{k-1}\mathsf{var}\big(\widehat{\beta}_j \,\big|\, \mathbb{Z}\big) = \mathsf{tr}\Big\{\mathsf{var}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\big)\Big\} = \mathsf{tr}\Big\{\sigma^2\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\} = \sigma^2\,\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\}.$$

This shows that multicollinearity

(i) does not have any impact on precision of the LSE of the response expectation $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$;

(ii) may have a serious impact on precision of the LSE of the regression coefficients $\boldsymbol{\beta}$. At the same time, since LSE is BLUE, there exist no better linear unbiased estimator of $\boldsymbol{\beta}$. If additionally normality is assumed there even exist no better unbiased estimator at all.

End of Lecture #22 (08/12/2016)

Start of Lecture #23 (14/12/2016)

An impact of multicollinearity can also be expressed by considering a problem of estimating the squared Euclidean norm of $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$ and $\boldsymbol{\beta}$, respectively. As natural estimators of those squared norms are the squared norms of the corresponding LSE's, i.e., $\big\|\widehat{\boldsymbol{Y}}\big\|^2$ and $\big\|\widehat{\boldsymbol{\beta}}\big\|^2$, respectively. As we show, those estimators are biased, nevertheless, the amount of bias does not depend on a degree of multicollinearity in case of $\big\|\widehat{\boldsymbol{Y}}\big\|^2$ but depends on it in case of $\big\|\widehat{\boldsymbol{\beta}}\big\|^2$.

Lemma 12.1 Bias in estimation of the squared norms.
Let $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}_{n\times k}) = k$. *The following then holds.*

$$\mathbb{E}\Big(\big\|\widehat{\boldsymbol{Y}}\big\|^2 - \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 \,\Big|\, \mathbb{Z}\Big) = \sigma^2\, k,$$

$$\mathbb{E}\Big(\big\|\widehat{\boldsymbol{\beta}}\big\|^2 - \big\|\boldsymbol{\beta}\big\|^2 \,\Big|\, \mathbb{Z}\Big) = \sigma^2\, \mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\}.$$

Proof. For clarity of notation, condition will be omitted from notation of most expectations and variances. Nevertheless, all are still understood as conditional expectations and variances given the covariate values $\mathbb{Z}$.

$\underline{\mathbb{E}\Big(\big\|\widehat{\boldsymbol{Y}}\big\|^2 - \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 \,\Big|\, \mathbb{Z}\Big)}$

- Let us calculate:

$$\mathbb{E}\big\|\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big\|^2 = \mathbb{E}\Big\{\sum_{i=1}^{n}\big(\widehat{Y}_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\big)^2\Big\} = \sum_{i=1}^{n}\mathsf{var}\big(\widehat{Y}_i\big)$$
$$= \mathsf{tr}\Big\{\mathsf{var}\big(\widehat{\boldsymbol{Y}}\big)\Big\} = \mathsf{tr}\big(\sigma^2\,\mathbb{H}\big) = \sigma^2\,\mathsf{tr}(\mathbb{H}) = \sigma^2\,k.$$

- At the same time:

$$\mathbb{E}\big\|\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big\|^2 = \mathbb{E}\big(\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big)^\top\big(\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big)$$
$$= \mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 + \mathbb{E}\big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 - 2\,\boldsymbol{\beta}^\top\mathbb{X}^\top\underbrace{\mathbb{E}\widehat{\boldsymbol{Y}}}_{\mathbb{X}\boldsymbol{\beta}}$$
$$= \mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 + \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 - 2\,\big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 = \mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 - \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2.$$

- So that, $\mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 - \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 = \sigma^2\,k,$
$\mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 = \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 + \sigma^2\,k.$

$\underline{\mathbb{E}\Big(\big\|\widehat{\boldsymbol{\beta}}\big\|^2 - \big\|\boldsymbol{\beta}\big\|^2 \,\Big|\, \mathbb{Z}\Big)}$

- Let us start in a similar way:

$$\mathbb{E}\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big\|^2 = \mathbb{E}\Big\{\sum_{j=0}^{k-1}\big(\widehat{\beta}_j - \beta_j\big)^2\Big\} = \sum_{j=0}^{k-1}\mathsf{var}\big(\widehat{\beta}_j\big)$$
$$= \mathsf{tr}\Big\{\mathsf{var}\big(\widehat{\boldsymbol{\beta}}\big)\Big\} = \mathsf{tr}\Big\{\sigma^2\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\} = \sigma^2\,\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\}.$$

- At the same time:

$$\begin{aligned}\mathbb{E}\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big\|^2 &= \mathbb{E}\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)^\top\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)\\ &= \mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2+\mathbb{E}\big\|\boldsymbol{\beta}\big\|^2-2\,\boldsymbol{\beta}^\top\underbrace{\mathbb{E}\widehat{\boldsymbol{\beta}}}_{\boldsymbol{\beta}}\\ &= \mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2+\big\|\boldsymbol{\beta}\big\|^2-2\,\big\|\boldsymbol{\beta}\big\|^2=\mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2-\big\|\boldsymbol{\beta}\big\|^2.\end{aligned}$$

- So that, $\mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2-\big\|\boldsymbol{\beta}\big\|^2 = \sigma^2\,\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\}$,

$$\mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2 = \big\|\boldsymbol{\beta}\big\|^2+\underbrace{\sigma^2\,\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\}}_{\sum_{j=0}^{k-1}\mathsf{var}\big(\widehat{\beta}_j\big)}.$$

❑

12.1.3 Variance inflation factor and tolerance

Notation. For a given linear model $\boldsymbol{Y}\,\big|\,\mathbb{Z}\sim\big(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k})=k$, where

$$\boldsymbol{Y}=\big(Y_1,\,\ldots,\,Y_n\big)^\top,$$

$$\mathbb{X}=\big(\mathbf{1}_n,\,\boldsymbol{X}^1,\,\ldots,\,\boldsymbol{X}^{k-1}\big),\qquad \boldsymbol{X}^j=\big(X_{1,j},\,\ldots,\,X_{n,j}\big)^\top,\quad j=1,\ldots,k-1,$$

the following (partly standard) notation, will be used:

Response sample mean:	$\overline{Y}=\dfrac{1}{n}\sum_{i=1}^{n}Y_i$;
Square root of the total sum of squares:	$T_Y=\sqrt{\sum_{i=1}^{n}\big(Y_i-\overline{Y}\big)^2}=\big\|\boldsymbol{Y}-\overline{Y}\mathbf{1}_n\big\|$;
Fitted values:	$\widehat{\boldsymbol{Y}}=\big(\widehat{Y}_1,\,\ldots,\,\widehat{Y}_n\big)^\top$;
Coefficient of determination:	$R^2\;=\;1\;-\;\dfrac{\big\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\big\|^2}{\big\|\boldsymbol{Y}-\overline{Y}\mathbf{1}_n\big\|^2}\;=\;1\;-\;\dfrac{\big\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\big\|^2}{T_Y^2}$.
Residual mean square:	$\mathsf{MS}_e=\dfrac{1}{n-k}\big\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\big\|^2$.

Further, for each $j=1,\ldots,k-1$, consider a linear model M_j, where the vector $\boldsymbol{X}^j$ acts as a response and the model matrix is

$$\mathbb{X}^{(-j)}=\big(\mathbf{1}_n,\,\boldsymbol{X}^1,\,\ldots,\,\boldsymbol{X}^{j-1},\,\boldsymbol{X}^{j+1},\,\ldots,\,\boldsymbol{X}^{k-1}\big).$$

The following notation will be used:

Column sample mean: $\overline{X}^j = \dfrac{1}{n}\sum_{i=1}^{n} X_{i,j}$;

Square root of the total sum of squares from model M_j:

$$T_j = \sqrt{\sum_{i=1}^{n}\left(X_{i,j} - \overline{X}^j\right)^2} = \left\|\boldsymbol{X}^j - \overline{X}^j\mathbf{1}_n\right\|;$$

Fitted values from model M_j: $\widehat{\boldsymbol{X}}^j = \left(\widehat{X}_{1,j}, \ldots, \widehat{X}_{n,j}\right)^\top$;

Coefficient of determination from model M_j:

$$R_j^2 \;=\; 1 \;-\; \frac{\left\|\boldsymbol{X}^j - \widehat{\boldsymbol{X}}^j\right\|^2}{\left\|\boldsymbol{X}^j - \overline{X}^j\mathbf{1}_n\right\|^2} \;=\; 1 \;-\; \frac{\left\|\boldsymbol{X}^j - \widehat{\boldsymbol{X}}^j\right\|^2}{T_j^2}.$$

Notes.

(i) If data (response random variables and non-intercept covariates) $\left(Y_i, X_{i,1}, \ldots, X_{i,k-1}\right)^\top$, $i = 1, \ldots, n$ are a random sample from a distribution of a generic random vector $\left(Y, X_1, \ldots, X_{k-1}\right)^\top$ then

- The coefficient of determination R^2 is also a squared value of a sample coefficient of multiple correlation between Y and $\boldsymbol{X} := \left(X_1, \ldots, X_{k-1}\right)^\top$.
- For each $j = 1, \ldots, k-1$, the coefficient of determination R_j^2 is also a squared value of a sample coefficient of multiple correlation between X_j and $\boldsymbol{X}_{(-j)} := \left(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{k-1}\right)^\top$.

(ii) For given $j = 1, \ldots, k-1$:

- A value of R_j^2 close to 1 means that the jth column $\boldsymbol{X}^j$ is almost equal to some linear combination of the columns of the matrix $\mathbb{X}^{(-j)}$ (remaining columns of the model matrix). We then say that $\boldsymbol{X}^j$ is *collinear* with the remaining columns of the model matrix.
- A value of $R_j^2 = 0$ means that
  - the column $\boldsymbol{X}^j$ is orthognal to all remaining non-intercept regressors (non-intercept columns of the matrix $\mathbb{X}^{(-j)}$);
  - the jth regressor represented by the random variable X_j is multiply uncorrelated with the remaining regressors represented by the random vector $\boldsymbol{X}_{(-j)}$.

For a given linear model $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \left(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\right),\ \mathsf{rank}(\mathbb{X}_{n\times k}) = k$,

$$\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\right) = \mathsf{MS}_e\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}.$$

The following Theorem shows that diagonal elements of the matrix $\mathsf{MS}_e\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$, i.e., values $\widehat{\mathsf{var}}\left(\widehat{\beta}_j \,\big|\, \mathbb{Z}\right)$ can also be calculated, for $j = 1, \ldots, k-1$, using above defined quantities T_Y, T_j, R^2, R_j^2.

Theorem 12.2 Estimated variances of the LSE of the regression coefficients.
For a given dataset for which a linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$, $\boldsymbol{X} = \big(\mathbf{1}_n,\, \boldsymbol{X}^1,\, \ldots,\, \boldsymbol{X}^{k-1}\big)$ *is applied, diagonal elements of the matrix* $\widehat{\mathsf{var}}\big(\widehat{\boldsymbol{\beta}} \,|\, \mathbb{Z}\big) = \mathsf{MS}_e\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$, *can also be calculated, for* $j = 1,\, \ldots,\, k-1$, *as*

$$\widehat{\mathsf{var}}\big(\widehat{\beta}_j \,|\, \mathbb{Z}\big) = \left(\frac{T_Y}{T_j}\right)^2 \cdot \frac{1-R^2}{n-k} \cdot \frac{1}{1-R_j^2}.$$

Proof. **Proof/calculations were skipped and are not requested for the exam.**

Suppose that $T_Y,\, T_1,\, \ldots,\, T_{k-1}$ are real constants such that the vectors

$$\boldsymbol{Y}^\star = \frac{1}{T_Y}\big(\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\big),$$
$$\boldsymbol{X}^{j,\star} = \frac{1}{T_j}\big(\boldsymbol{X} - \overline{X}^j\mathbf{1}_n\big), \qquad j = 1,\, \ldots, k-1$$

have all unity Euclidean norm. For a given dataset, appropriate constants $T_Y,\, T_1,\, \ldots,\, T_{k-1}$ are indeed given as indicated at the beginning of Section 12.1.3. Note that since we now only want to find an expression on how to calculate, for a given dataset, diagonal elements of a certain matrix $\widehat{\mathsf{var}}\big(\widehat{\boldsymbol{\beta}} \,|\, \mathbb{Z}\big) = \mathsf{MS}_e\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$, randomness of $T_Y,\, T_1,\, \ldots,\, T_{k-1}$ will not be taken into account. In this context, the vector $\boldsymbol{Y}^\star$ is also called the *standardized* respose vector and the vectors $\boldsymbol{X}^{j,\star}$ the *standardized* regressors. Further, let

$$\mathbb{X}^\star = \big(\boldsymbol{X}^{1,\star},\, \ldots,\, \boldsymbol{X}^{k-1,\star}\big)$$

be the matrix with the standardized non-intercept regressors in columns.

We have

- Vector $\boldsymbol{Y}^\star$ and all columns of $\mathbb{X}^\star$ are of *unity* Euclidean norm.
- Vector $\boldsymbol{Y}^\star$ and all columns of $\mathbb{X}^\star$ are orthogonal to a vector $\mathbf{1}_n$, i.e.,
$$\big(\boldsymbol{Y}^\star\big)^\top\mathbf{1}_n = 0, \qquad \big(\mathbb{X}^\star\big)^\top\mathbf{1}_n = \mathbf{0}_{k-1}.$$

Let us now consider a linear model based on *standardized variables* (as if $T_Y,\, T_1,\, \ldots,\, T_{k-1}$ were pre-specified constants). Let $\big(\beta_0^\star,\, \beta_1^\star,\, \ldots,\, \beta_{k-1}^\star\big)^\top$ be the regression coefficients in a model

$$\mathsf{M}^\star\colon\ \boldsymbol{Y}^\star \,|\, \mathbb{Z} \sim \left(\big(\mathbf{1}_n,\, \mathbb{X}^\star\big)\begin{pmatrix}\beta_0^\star\\ \boldsymbol{\beta}^\star\end{pmatrix},\, (\sigma^\star)^2\mathbf{I}_n\right),$$

with the model matrix $\mathbb{X}_{st} = \big(\mathbf{1}_n,\, \mathbb{X}^\star\big)$. Let $\boldsymbol{\beta}^\star = \big(\beta_1^\star,\, \ldots,\, \beta_{k-1}^\star\big)^\top$ be the subvector of the regression coefficients related to the non-intercept columns of the model matrix.

As usually, let $\boldsymbol{\beta} = \big(\beta_0,\, \beta_1,\, \ldots,\, \beta_{k-1}\big)^\top$ be the regression coefficients in the original model

$$\mathsf{M}\colon\ \boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big).$$

Model M can be written as

$$\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \sum_{j=1}^{k-1} \boldsymbol{X}^j \beta_j + \boldsymbol{\varepsilon}, \tag{12.2}$$

where $\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z} \sim \big(\mathbf{0}_n \, \sigma^2 \, \mathbf{I}_n\big)$.

That is, data satisfying model M also satisfy

$$\boldsymbol{Y} - \overline{Y}\mathbf{1}_n = (\beta_0 - \overline{Y})\mathbf{1}_n + \sum_{j=1}^{k-1} (\boldsymbol{X}^j - \overline{X}^j \mathbf{1}_n)\beta_j + \sum_{j=1}^{k-1} \overline{X}^j \beta_j \mathbf{1}_n + \boldsymbol{\varepsilon},$$

$$\underbrace{\frac{1}{T_Y}\left(\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\right)}_{\boldsymbol{Y}^\star} = \underbrace{\frac{\left(\beta_0 - \overline{Y} + \sum_{j=1}^{k-1} \overline{X}^j \beta_j\right)}{T_Y}}_{\beta_0^\star} \mathbf{1}_n + \sum_{j=1}^{k-1} \underbrace{\frac{1}{T_j}\left(\boldsymbol{X}^j - \overline{X}^j \mathbf{1}_n\right)}_{\boldsymbol{X}^{j,\star}} \underbrace{\frac{T_j}{T_Y}\beta_j}_{\beta_j^\star} + \underbrace{\frac{1}{T_Y}\boldsymbol{\varepsilon}}_{\boldsymbol{\varepsilon}^\star}.$$

In other words, if data satisfy model M then the standardized data satisfy the model M* with the error terms $\boldsymbol{\varepsilon}^\star = \boldsymbol{Y}^\star - \beta_0^\star \mathbf{1}_n - \mathbb{X}^\star \boldsymbol{\beta}^\star$ having $\boldsymbol{\varepsilon}^\star \,\big|\, \mathbb{Z} \sim \big(\mathbf{0}_n \, (\sigma^\star)^2 \, \mathbf{I}_n\big)$ and parameters of the two models are in mutual relationships

$$\begin{aligned}
\beta_0^\star &= \frac{\beta_0 - \overline{Y} + \sum_{j=1}^{k-1} \overline{X}_j \beta_j}{T_Y}, \\
\beta_j^\star &= \frac{T_j}{T_Y}\beta_j, \qquad\qquad j = 1, \ldots, k-1, \\
\sigma^\star &= \frac{\sigma}{T_Y}.
\end{aligned}$$

That is,

- $\beta_0^\star$ is only shifted-scaled β_0.
- $\beta_j^\star$ is only scaled β_j, $j = 1, \ldots, k-1$.
- $\sigma^\star$ is only scaled σ.

Due to linearity, the same relationships hold also for the LSE in both models. That is (now written in the opposite direction):

$$\begin{aligned}
\widehat{\beta}_0 &= T_Y \, \widehat{\beta}_0^\star + \overline{Y} - \sum_{j=1}^{k-1} \overline{X}_j \frac{T_Y}{T_j} \widehat{\beta}_j^\star, \\
\widehat{\beta}_j &= \frac{T_Y}{T_j} \widehat{\beta}_j^\star, \qquad\qquad j = 1, \ldots, k-1.
\end{aligned}$$

Moreover, the fitted values in both models must also be linked by the same (linear) relationship as the standardized and original response variables. That is,

$$\widehat{\boldsymbol{Y}}^\star = \frac{1}{T_Y}\big(\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}_n\big), \qquad \widehat{Y}_i^\star = \frac{1}{T_Y}(\widehat{Y}_i - \overline{Y}), \quad i = 1, \ldots, n,$$

$$\widehat{\boldsymbol{Y}} = T_Y \widehat{\boldsymbol{Y}}^\star + \overline{Y}\mathbf{1}_n, \qquad \widehat{Y}_i = T_Y \, \widehat{Y}_i^\star + \overline{Y}, \quad i = 1, \ldots, n.$$

The residual sum of squares in model $\mathsf{M}^\star$ is then:

$$\begin{aligned}\mathsf{SS}_e^\star &= \left\|\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\right\|^2 \\ &= \sum_{i=1}^n (Y_i^\star - \widehat{Y}_i^\star)^2 = \frac{1}{T_Y^2}\sum_{i=1}^n (T_Y\, Y_i^\star - T_Y\, \widehat{Y}_i^\star)^2 \\ &= \frac{1}{T_Y^2}\sum_{i=1}^n \left\{T_Y\, Y_i^\star + \overline{Y} - (T_Y\, \widehat{Y}_i^\star + \overline{Y})\right\}^2 \\ &= \frac{1}{T_Y^2}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \frac{1}{T_Y^2}\,\mathsf{SS}_e,\end{aligned}$$

where SS_e is the residual sum of squares in the original model M.

Moreover, note that $T_Y^2 = \left\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\right\|^2$ is also the total sum of squares SS_T for the original response vector $\boldsymbol{Y}$. That is,

$$\mathsf{SS}_e^\star = \frac{\mathsf{SS}_e}{\mathsf{SS}_T} = 1 - R^2, \tag{12.3}$$

where R^2 is the coefficient of determination of the original model M. The residual mean square in model $\mathsf{M}^\star$ can now be written as

$$\mathsf{MS}_e^\star = \frac{\mathsf{SS}_e^\star}{n-k} = \frac{1-R^2}{n-k}.$$

Let us now explicitely express the LSE of the regression coefficients vector $\left(\beta_0^\star,\, \boldsymbol{\beta}^{\star\top}\right)^\top$ in model $\mathsf{M}^\star$ which are given as

$$\begin{pmatrix}\widehat{\beta}_0^\star \\ \widehat{\boldsymbol{\beta}}^\star\end{pmatrix} = \left(\mathbb{X}_{st}^\top \mathbb{X}_{st}\right)^{-1}\mathbb{X}_{st}^\top \boldsymbol{Y}^\star.$$

First,

$$\mathbb{X}_{st}^\top \mathbb{X}_{st} \;=\; \left(\mathbf{1}_n,\, \mathbb{X}^\star\right)^\top\left(\mathbf{1}_n,\, \mathbb{X}^\star\right) \;=\; \begin{pmatrix} n & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \left(\mathbb{X}^\star\right)^\top\mathbb{X}^\star\end{pmatrix} \;=\; \begin{pmatrix} n & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}\end{pmatrix},$$

where $\mathbb{R}_{X,X} := \left(\mathbb{X}^\star\right)^\top\mathbb{X}^\star = \left(r_{X,X}^{j,l}\right)_{j,l=1,\ldots,k-1}$, has elements

$$\begin{aligned} r_{X,X}^{j,l} &= \frac{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)(X_{i,l} - \overline{X}^l)}{T_j\, T_l} \\ &= \frac{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)(X_{i,l} - \overline{X}^l)}{\sqrt{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)^2}\sqrt{\sum_{i=1}^n (X_{i,l} - \overline{X}^l)^2}}, \quad j,l = 1,\ldots,k-1.\end{aligned}$$

That is, $\mathbb{R}_{X,X} = \left(\mathbb{X}^\star\right)^\top\mathbb{X}^\star$ is a sample correlation matrix (with ones on a diagonal) of the non-intercept regressors from the original model matrix $\mathbb{X}$.

We then also have,

$$\left(\mathbb{X}_{st}^\top \mathbb{X}_{st}\right)^{-1} \;=\; \begin{pmatrix} n & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}\end{pmatrix} \;=\; \begin{pmatrix} \frac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}^{-1}\end{pmatrix}.$$

Second,

$$\mathbb{X}_{st}^\top \boldsymbol{Y}^\star \;=\; \big(\mathbf{1}_n,\, \mathbb{X}^\star\big)^\top \boldsymbol{Y}^\star \;=\; \begin{pmatrix} \sum_{i=1}^n Y_i^\star \\ \big(\mathbb{X}^\star\big)^\top \boldsymbol{Y}^\star \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ \boldsymbol{r}_{X,Y} \end{pmatrix},$$

where $\boldsymbol{r}_{X,Y} := \big(\mathbb{X}^\star\big)^\top \boldsymbol{Y}^\star = \big(r_{X,Y}^j\big)_{j=1,\ldots,k-1}$, has elements

$$\begin{aligned} r_{X,Y}^j &= \frac{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)(Y_i - \overline{Y})}{T_j\, T_Y} \\ &= \frac{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}, \quad j = 1, \ldots, k-1, \end{aligned}$$

That is, $\boldsymbol{r}_{X,Y} = \big(\mathbb{X}^\star\big)^\top \boldsymbol{Y}^\star$ is a vector of sample correlation coefficients between the regressors from the model matrix $\mathbb{X}$ and the response $\boldsymbol{Y}$.

Hence,

$$\begin{pmatrix} \widehat{\beta}_0^\star \\ \widehat{\boldsymbol{\beta}}^\star \end{pmatrix} = \begin{pmatrix} \frac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{r}_{X,Y} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbb{R}_{X,X}^{-1}\, \boldsymbol{r}_{X,Y} \end{pmatrix},$$

$$\mathsf{var}\left\{ \begin{pmatrix} \widehat{\beta}_0^\star \\ \widehat{\boldsymbol{\beta}}^\star \end{pmatrix} \,\middle|\, \mathbb{Z} \right\} = (\sigma^\star)^2 \left(\mathbb{X}_{st}^\top \mathbb{X}_{st}\right)^{-1} = (\sigma^\star)^2 \begin{pmatrix} \frac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}^{-1} \end{pmatrix}.$$

That is, we have

$$\begin{aligned} &\widehat{\beta}_0^\star = 0, && \mathsf{var}\big(\widehat{\beta}_0^\star \,\big|\, \mathbb{Z}\big) = \frac{(\sigma^\star)^2}{n}, \\ &\widehat{\boldsymbol{\beta}}^\star = \mathbb{R}_{X,X}^{-1}\, \boldsymbol{r}_{X,Y} && \mathsf{var}\big(\widehat{\boldsymbol{\beta}}^\star \,\big|\, \mathbb{Z}\big) = (\sigma^\star)^2\, \mathbb{R}_{X,X}^{-1}. \end{aligned} \tag{12.4}$$

Before we proceed, let us derive the hat matrix and the fitted values of model $\mathsf{M}^\star$. The hat matrix of model $\mathsf{M}^\star$ is calculated as

$$\begin{aligned} \mathbb{H}_{st} = \big(\mathbf{1}_n,\, \mathbb{X}^\star\big) \left(\mathbb{X}_{st}^\top \mathbb{X}_{st}\right)^{-1} \big(\mathbf{1}_n,\, \mathbb{X}^\star\big)^\top &= \big(\mathbf{1}_n,\, \mathbb{X}^\star\big) \begin{pmatrix} \frac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{x,x}^{-1} \end{pmatrix} \big(\mathbf{1}_n,\, \mathbb{X}^\star\big)^\top \\ &= \frac{1}{n}\, \mathbf{1}_n\, \mathbf{1}_n{}^\top + \underbrace{\mathbb{X}^\star \mathbb{R}_{X,X}^{-1} \big(\mathbb{X}^\star\big)^\top}_{=:\, \mathbb{H}^\star}. \end{aligned}$$

Observe that

- $\mathbb{H}_{st}$ is the projection matrix into $\mathcal{M}\Big(\big(\mathbf{1}_n,\, \mathbb{X}^\star\big)\Big)$.
- $\mathbb{H}^\star = \mathbb{X}^\star \mathbb{R}_{X,X}^{-1} \big(\mathbb{X}^\star\big)^\top$ is the projection matrix into $\mathcal{M}\big(\mathbb{X}^\star\big)$.

The fitted values of the model $\mathsf{M}^\star$ are then given by

$$\widehat{\boldsymbol{Y}}^\star = \mathbb{H}_{st} \boldsymbol{Y}^\star = \frac{1}{n} \mathbf{1}_n \underbrace{\mathbf{1}_n{}^\top \boldsymbol{Y}^\star}_{0} + \mathbb{H}^\star \boldsymbol{Y}^\star = \mathbb{H}^\star \boldsymbol{Y}^\star.$$

Finally, observe that (while remembering that any hat matrix is symmetric and idempotent)

$$\big(\widehat{\boldsymbol{Y}}^\star\big)^\top \boldsymbol{Y}^\star \;=\; \big(\boldsymbol{Y}^\star\big)^\top \mathbb{H}^\star \boldsymbol{Y}^\star \;=\; \big(\boldsymbol{Y}^\star\big)^\top \mathbb{H}^\star \mathbb{H}^\star \boldsymbol{Y}^\star \;=\; \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star.$$

Consequently,

$$\begin{aligned}\mathsf{SS}_e^\star \;=\; \big\|\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\big\|^2 \;&=\; \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star - \big(\boldsymbol{Y}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star - \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \boldsymbol{Y}^\star + \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star \\ &=\; \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star - \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star. \qquad (12.5)\end{aligned}$$

Let $d_{X,X}^{j,j}$, $j = 1, \ldots, k-1$ be diagonal elements of the matrix $\mathbb{R}_{X,X}^{-1}$. That is, from (12.4):

$$\mathsf{var}\big(\widehat{\beta}_j^\star \,\big|\, \mathbb{Z}\big) \;=\; (\sigma^\star)^2\, d_{X,X}^{j,j}, \quad j = 1, \ldots, k-1.$$

To derive the value of $d_{X,X}^{j,j}$, $j = 1, \ldots, k-1$, let us first consider the sample correlation matrix based on both the response vector and the non-intercept regressors:

$$\mathbb{R}_{(Y,X),(Y,X)} = \begin{pmatrix} 1 & \boldsymbol{r}_{X,Y}^\top \\ \boldsymbol{r}_{X,Y} & \mathbb{R}_{X,X} \end{pmatrix}.$$

Using Theorem A.4, we can express its inverse:

$$\mathbb{R}_{(Y,X),(Y,X)}^{-1} = \begin{pmatrix} \big(1 - \boldsymbol{r}_{X,Y}^\top \mathbb{R}_{X,X}^{-1} \boldsymbol{r}_{X,Y}\big)^{-1} & \circledast \\ \circledast & \circledast \end{pmatrix}.$$

Further (while also using Eqs. 12.3 and 12.5),

$$\begin{aligned}1 - \boldsymbol{r}_{X,Y}^\top \mathbb{R}_{X,X}^{-1} \boldsymbol{r}_{X,Y} \\ &= \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star \;-\; \big(\boldsymbol{Y}^\star\big)^\top \underbrace{\mathbb{X}^\star \big\{(\mathbb{X}^\star)^\top \mathbb{X}^\star\big\}^{-1} (\mathbb{X}^\star)^\top}_{\mathbb{H}^\star} \boldsymbol{Y}^\star \\ &= \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star \;-\; \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star = \mathsf{SS}_e^\star = 1 - R^2,\end{aligned}$$

where R^2 is coefficient of determination from the linear model $\mathsf{M}\colon\ \boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$.

That is, the $(Y - Y)$ diagonal element of matrix $\mathbb{R}_{(Y,X),(Y,X)}^{-1}$ equals to $(1 - R^2)^{-1}$, where R^2 is the coefficient of determination from a model with $\boldsymbol{Y}$ as response and the model matrix composed of the intercept column and the original regressors $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^{k-1}$, i.e., the model matrix

$$\mathbb{X} = \big(\mathbf{1}_n,\, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^{k-1}\big).$$

Now, consider for given $j = 1, \ldots, k-1$ a linear model where the response vector is equal to $\boldsymbol{X}^j$ (the jth regressor from the original model) and the model matrix is

$$\mathbb{X}^{(-j)} = \big(\mathbf{1}_n,\, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^{j-1}, \boldsymbol{X}^{j+1}, \ldots, \boldsymbol{X}^{k-1}\big).$$

The role of the matrix $\mathbb{R}_{(Y,X),(Y,X)}^{-1}$ would now be played by matrix $\mathbb{R}_{X,X}^{-1}$ whose rows and columns were reordered and its $(1-1)$ element is equal to $d_{X,X}^{j,j}$, i.e., to the jth diagonal element of the matrix $\mathbb{R}_{X,X}^{-1}$. By the same arguments as above, we arrive at

$$d_{X,X}^{j,j} = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination from a linear model with $\boldsymbol{X}^j$ as response and the model matrix $\mathbb{X}^{(-j)}$.

So we have,

$$\mathsf{var}\big(\widehat{\beta}_j^\star \,\big|\, \mathbb{Z}\big) \;=\; \frac{(\sigma^\star)^2}{1-R_j^2}, \qquad j = 1, \ldots, k-1.$$

The jth diagonal element ($j = 1, \ldots, k-1$) of the matrix $\mathsf{var}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\big)$ can now be expressed as

$$\mathsf{var}\big(\widehat{\beta}_j \,\big|\, \mathbb{Z}\big) \;=\; \mathsf{var}\left(\frac{T_Y}{T_j}\widehat{\beta}_j^\star \,\middle|\, \mathbb{Z}\right) \;=\; \left(\frac{T_Y}{T_j}\right)^2 \mathsf{var}\big(\widehat{\beta}_j^\star \,\big|\, \mathbb{Z}\big) \;=\; \left(\frac{T_Y}{T_j}\right)^2 \frac{(\sigma^\star)^2}{1-R_j^2}.$$

Let us now replace an unknown $(\sigma^\star)^2$ by its estimator $\mathsf{MS}_e^\star = \frac{\mathsf{SS}_e^\star}{n-k} = \frac{1-R^2}{n-k}$. We get

$$\widehat{\mathsf{var}}\big(\widehat{\beta}_j \,\big|\, \mathbb{Z}\big) \;=\; \left(\frac{T_Y}{T_j}\right)^2 \frac{1-R^2}{n-k}\,\frac{1}{1-R_j^2} \qquad j = 1, \ldots, k-1.$$

❑

Definition 12.1 Variance inflation factor and tolerance.

For given $j = 1, \ldots, k-1$, the variance inflation factor[2] *and the tolerance*[3] *of the jth regressor of the linear model $\boldsymbol{Y} \,|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$ are values VIF_j and $Toler_j$, respectively, defined as*

$$VIF_j = \frac{1}{1-R_j^2}, \qquad Toler_j = 1 - R_j^2 = \frac{1}{VIF_j}.$$

Notes.

- With $R_j = 0$ (the jth regressor orthogonal to all remaining regressors, the j regressor multiply uncorrelated with the remaining ones), $\mathrm{VIF}_j = 1$.
- With $R_j \longrightarrow 1$ (the jth regressor collinear with the remaining regressors, the jth regressor almost perfectly multiply correlated with the remaining ones), $\mathrm{VIF}_j \longrightarrow \infty$.

Interpretation and use of VIF

- If we take into account the statement of Theorem 12.2, the VIF of the jth regressor ($j = 1, \ldots, k-1$) can be interpreted as a factor by which the (estimated) variance of $\widehat{\beta}_j$ is multiplied (inflated) compared to an optimal situation when the jth regressor is orthogonal to (multiply uncorrelated with) the remaining regressors included in the model. Hence the term *variance inflation factor.*
- Under assumption of normality, the confidence interval for β_j with a coverage of $1-\alpha$ has the lower and the upper bounds given as

$$\widehat{\beta}_j \;\pm\; \mathsf{t}_{n-k}\Big(1-\frac{\alpha}{2}\Big)\,\sqrt{\widehat{\mathsf{var}}\big(\widehat{\beta}_j\big)}.$$

[2] *varianční inflační faktor* [3] *tolerance*

Using the statement of Theorem 12.2, the lower and the upper bounds of the confidence interval for β_j can also be written as

$$\widehat{\beta}_j \pm \mathsf{t}_{n-k}\Big(1-\frac{\alpha}{2}\Big)\,\frac{T_Y}{T_j}\,\sqrt{\frac{1-R^2}{n-k}}\,\sqrt{\mathrm{VIF}_j}.$$

That is, the (square root of) VIF also provides a factor by which the half-length (radius) of the confidence interval is inflated compared to an optimal situation when the jth regressor is orthogonal to (multiply uncorrelated with) the remaining regressors included in the model, namely,

$$\mathrm{VIF}_j = \left(\frac{\mathrm{Vol}_j}{\mathrm{Vol}_{0,j}}\right)^2, \tag{12.6}$$

where $\mathrm{Vol}_j =$ length (volume) of the confidence interval for β_j;
$\mathrm{Vol}_{0,j} =$ length (volume) of the confidence interval for β_j if it was $R_j^2 = 0$.

- Regressors with a high VIF are possibly responsible for multicollinearity. Nevertheless, the VIF does not reveal which regressors are mutually collinear.

Generalized variance inflation factor

Beginning of skipped part

A generalized variance inflation factor was derived by Fox and Monette (1992) to evaluate a degree of collinearity between a specified group of regressors and the remaining regressors. Let

- $\mathcal{J} \subset \{1,\dots,k-1\}$, $|\mathcal{J}| = m$;
- $\boldsymbol{\beta}_{[\mathcal{J}]}$ be a subvector of $\boldsymbol{\beta}$ having the elements indexed by $j \in \mathcal{J}$.

Under normality, a confidence ellipsoid for $\beta_{\mathcal{J}}$ with a coverage $1-\alpha$ is

$$\Big\{\boldsymbol{\beta}_{[\mathcal{J}]} \in \mathbb{R}^m :\ \big(\boldsymbol{\beta}_{[\mathcal{J}]} - \widehat{\boldsymbol{\beta}}_{[\mathcal{J}]}\big)^\top \Big(\mathsf{MS}_e\, \mathbb{V}_{[\mathcal{J}]}\Big)^{-1} \big(\boldsymbol{\beta}_{[\mathcal{J}]} - \widehat{\boldsymbol{\beta}}_{[\mathcal{J}]}\big) \ < \ m\,\mathcal{F}_{m,n-k}(1-\alpha)\Big\},$$

$$\mathbb{V}_{[\mathcal{J}]} = (\mathcal{J}-\mathcal{J}) \text{ block of the matrix } \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}. \tag{12.7}$$

Let $\mathrm{Vol}_{\mathcal{J}}$: volume of the confidence ellipsoid (12.7);
$\mathrm{Vol}_{0,\mathcal{J}}$: volume of the confidence ellipsoid (12.7) would all columns of $\mathbb{X}$ coresponding to $\boldsymbol{\beta}_{[\mathcal{J}]}$ be orthogonal to the remaining colums of $\mathbb{X}$.

A definition of the generalized variance inflation factor gVIF is motivated by (12.6) as it is given as

$$\mathrm{gVIF}_{\mathcal{J}} = \left(\frac{\mathrm{Vol}_{\mathcal{J}}}{\mathrm{Vol}_{0,\mathcal{J}}}\right)^2.$$

It is seen that with $\mathcal{J} = \{j\}$ for some $j = 1,\dots,k-1$, the generalized VIF simplifies into a standard VIF, i.e.,

$$\mathrm{gVIF}_j = \mathrm{VIF}_j.$$

Notes.

- The generalized VIF is especially useful if $\mathcal{J}$ relates to the regression coefficients corresponding to the reparameterizing (pseudo)contrasts of one categorical covariate. It can then be shown that $\mathrm{gVIF}_{\mathcal{J}}$ does not depend on a choice of the (pseudo)contrasts. $\mathrm{gVIF}_{\mathcal{J}}$ then evaluates the magnitude of the linear dependence of a categorical variable and the remaining regressors.

- When comparing $\text{gVIF}_{\mathcal{J}}$ for index sets $\mathcal{J}$, $\left|\mathcal{J}\right|$ of different cardinality m, quantities

$$\text{gVIF}_{\mathcal{J}}^{\frac{1}{2m}} = \left(\frac{\text{Vol}_{\mathcal{J}}}{\text{Vol}_{0,\mathcal{J}}}\right)^{\frac{1}{m}} \tag{12.8}$$

  should be compared which all relate to volume units in 1D.

- Generalized VIF's (and standard VIF's if $m = 1$) together with (12.8) are calculated by the R function `vif` from the package `car`.

End of skipped part

12.1.4 Basic treatment of multicollinearity

Especially in situations when inference on the regression coefficients is of interest, i.e., when the primary purpose of the regression modelling is to evaluate which variables influence significantly the response expectation and which not, multicollinearity is a serious problem. Basic treatment of multicollinearity consists of preliminary exploration of mutual relationships between all covariates and then choosing only suitable representatives of each group of mutually multiply correlated covariates. Very basic decision can be based on pairwise correlation coefficients. In some (especially "cook-book") literature, rules of thumb are applied like *"Covariates with a correlation (in absolute value) higher than 0.80 should not be included together in one model."* Nevertheless, such rules should never be applied in an automatic manner (why just 0.80 and not 0.79, ...?) Decision on which covariates cause multicollinearity can additionally be based on (generalized) variance inflation factors. Nevertheless, also those should be used comprehensively. In general, if a large set of covariates is available to relate it to the response expectation, a deep (and often timely) analyzis of mutual relationships and their understanding must preceed any regression modelling that is to lead to useful results.

12.2 Misspecified regression space

We are often in a situation when a large (potentially enormous) number p of candidate regressors is available. The question is then which of them should be included in a linear model. As shown in Section 12.1, inclusion of all possible regressors in the model is not necessarily optimal and may even have seriously negative impact on the statistical inference we would like to draw using the linear model. In this section, we explore some (additional) properties of the least squares estimators and of the related prediction in two situations:

(i) *Omitted* important regressors.

(ii) *Irrelevant* regressors included in a model.

12.2.1 Omitted and irrelevant regressors

We will assume that possibly two sets of regressors are available:

(i) $\boldsymbol{X}_i$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = \boldsymbol{t}_X(\boldsymbol{Z}_i)$ for some transformation $\boldsymbol{t}_X : \mathbb{R}^p \longrightarrow \mathbb{R}^k$. They give rise to the model matrix

$$\mathbb{X}_{n\times k} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \Big(\boldsymbol{X}^0, \ldots, \boldsymbol{X}^{k-1}\Big).$$

It will still be assumed that $\boldsymbol{X}^0 = \big(1, \ldots, 1\big)^\top$ (almost surely) leading to the model matrix

$$\mathbb{X}_{n\times k} = \Big(\mathbf{1}_n, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^{k-1}\Big),$$

with explicitely included intercept term.

(ii) $\boldsymbol{V}_i$, $i = 1, \ldots, n$, where $\boldsymbol{V}_i = \boldsymbol{t}_V(\boldsymbol{Z}_i)$ for some transformation $\boldsymbol{t}_V : \mathbb{R}^p \longrightarrow \mathbb{R}^l$. They give rise to the model matrix

$$\mathbb{V}_{n\times l} = \begin{pmatrix} \boldsymbol{V}_1^\top \\ \vdots \\ \boldsymbol{V}_n^\top \end{pmatrix} = \Big(\boldsymbol{V}^1, \ldots, \boldsymbol{V}^l\Big).$$

We will assume that both matrices $\mathbb{X}$ and $\mathbb{V}$ are of a full column rank and their columns are linearly independent, i.e., we assume

$$\begin{aligned} &\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k, \quad \mathsf{rank}\big(\mathbb{V}_{n\times l}\big) = l, \\ &\text{for} \quad \mathbb{G}_{n\times(k+l)} := \big(\mathbb{X}, \mathbb{V}\big), \quad \mathsf{rank}\big(\mathbb{G}\big) = k + l < n. \end{aligned}$$

The matrices $\mathbb{X}$ and $\mathbb{G}$ give rise to two nested linear models:

Model M_X $\quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$;

Model M_{XV} $\quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big)$.

Depending on which of the two models is a correct one and which model is used for inference, we face two situations:

Omitted important regressors mean that the larger model M_{XV} is correct (with $\boldsymbol{\gamma} \neq \mathbf{0}_m$) but we base inference on model M_X. In particular,

- $\boldsymbol{\beta}$ is estimated using model M_X;
- σ^2 is estimated using model M_X;
- prediction is based on the fitted model M_X.

Irrelevant regressors included in a model that the smaller model M_X is correct but we base inference on model M_{XV}. In particular,

- $\boldsymbol{\beta}$ is estimated (together with $\boldsymbol{\gamma}$) using model M_{XV};
- σ^2 is estimated using model M_{XV};
- prediction is based on the fitted model M_{XV}.

Note that if M_X is correct then M_{XV} is correct as well. Nevertheless, it includes redundant parameters $\boldsymbol{\gamma}$ which are known to be equal to zeros.

Notation *(Quantities derived under the two models).*

Quantities derived while assuming model M_X will be indicated by subscript X, quantities derived while assuming model M_{XV} will be indicated by subscript XV. Namely,

(i) Quantities derived while assuming model M_X:

- Least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_X = \big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \mathbb{X}^\top \boldsymbol{Y} = \big(\widehat{\beta}_{X,0}, \ldots, \widehat{\beta}_{X,k-1}\big)^\top;$$

- Projection matrices into the regression space $\mathcal{M}\big(\mathbb{X}\big)$ and into the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$:

$$\mathbb{H}_X = \mathbb{X}\big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \mathbb{X}^\top, \qquad \mathbb{M}_X = \mathbf{I}_n - \mathbb{H}_X;$$

- Fitted values (LSE of a vector $\mathbb{X}\boldsymbol{\beta}$):

$$\widehat{\boldsymbol{Y}}_X = \mathbb{H}_X \boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}_X = \big(\widehat{Y}_{X,1}, \ldots, \widehat{Y}_{X,n}\big)^\top;$$

- Residuals

$$\boldsymbol{U}_X = \mathbb{M}_X \boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_X = \big(U_{X,1}, \ldots, U_{X,n}\big)^\top;$$

- Residual sum of squares and residual mean square:

$$\mathsf{SS}_{e,X} = \big\|\boldsymbol{U}_X\big\|^2, \qquad \mathsf{MS}_{e,X} = \frac{\mathsf{SS}_{e,X}}{n-k}.$$

(ii) Quantities derived while assuming model M_{XV}:

- Least squares estimator of $\big(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top\big)^\top$:

$$\big(\widehat{\boldsymbol{\beta}}_{XV}^\top, \widehat{\boldsymbol{\gamma}}_{XV}^\top\big)^\top = \big(\mathbb{G}^\top \mathbb{G}\big)^{-1} \mathbb{G}^\top \boldsymbol{Y},$$

$$\widehat{\boldsymbol{\beta}}_{XV} = \big(\widehat{\beta}_{XV,0}, \ldots, \widehat{\beta}_{XV,k-1}\big)^\top, \qquad \widehat{\boldsymbol{\gamma}}_{XV} = \big(\widehat{\gamma}_{XV,1}, \ldots, \widehat{\gamma}_{XV,l}\big)^\top;$$

- Projection matrices into the regression space $\mathcal{M}\big(\mathbb{G}\big)$ and into the residual space $\mathcal{M}\big(\mathbb{G}\big)^\perp$:

$$\mathbb{H}_{XV} = \mathbb{G}\big(\mathbb{G}^\top \mathbb{G}\big)^{-1} \mathbb{G}^\top, \qquad \mathbb{M}_{XV} = \mathbf{I}_n - \mathbb{H}_{XV};$$

- Fitted values (LSE of a vector $\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma}$):

$$\widehat{\boldsymbol{Y}}_{XV} = \mathbb{H}_{XV}\boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}_{XV} + \mathbb{V}\boldsymbol{\gamma}_{XV} = \big(\widehat{Y}_{XV,1}, \ldots, \widehat{Y}_{XV,n}\big)^\top;$$

- Residuals

$$\boldsymbol{U}_{XV} = \mathbb{M}_{XV}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_{XV} = \big(U_{XV,1}, \ldots, U_{XV,n}\big)^\top;$$

- Residual sum of squares and residual mean square:

$$\mathsf{SS}_{e,XV} = \big\|\boldsymbol{U}_{XV}\big\|^2, \qquad \mathsf{MS}_{e,XV} = \frac{\mathsf{SS}_{e,XV}}{n-k-l}.$$

Consequence of Lemma 11.1: Relationship between the quantities derived while assuming the two models.

Quantities derived while assuming models M_X *and* M_{XV} *are mutually in the following relationships:*

$$\begin{aligned}
\widehat{\boldsymbol{Y}}_{XV} - \widehat{\boldsymbol{Y}}_X &= \mathbb{M}_X\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\big)^{-1}\mathbb{V}^\top\boldsymbol{U}_X, \\
&= \mathbb{X}\big(\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X\big) + \mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}, \\
\widehat{\boldsymbol{\gamma}}_{XV} &= \big(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\big)^{-1}\mathbb{V}^\top\boldsymbol{U}_X, \\
\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X &= -\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}, \\
\mathsf{SS}_{e,X} - \mathsf{SS}_{e,XV} &= \big\|\mathbb{M}_X\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}\big\|^2, \\
\mathbb{H}_{XV} &= \mathbb{H}_X + \mathbb{M}_X\mathbb{V}\left(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\right)^{-1}\mathbb{V}^\top\mathbb{M}_X.
\end{aligned}$$

Proof. Direct use of Lemma 11.1 while taking into account the fact that now, all involved model matrices are of full-rank.

Relationship $\mathbb{H}_{XV} = \mathbb{H}_X + \mathbb{M}_X\mathbb{V}\left(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\right)^{-1}\mathbb{V}^\top\mathbb{M}_X$ was shown inside the proof of Lemma 11.1. It easily follows from a general expression of the hat matrix if we realize that

$$\mathcal{M}\big(\mathbb{X}, \mathbb{V}\big) = \mathcal{M}\big(\mathbb{X}, \mathbb{M}_X\mathbb{V}\big),$$

and that $\mathbb{X}^\top\mathbb{M}_X\mathbb{V} = \mathbf{0}_{k\times l}$.

❑

Theorem 12.3 Variance of the LSE in the two models.

Irrespective of whether M_X *or* M_{XV} *holds, the covariance matrices of the fitted values and the LSE of the regression coefficients satisfy the following:*

$$\begin{aligned} \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) \;-\; \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) \quad &\geq \quad 0, \\ \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) \;-\; \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X} \,\big|\, \mathbb{Z}\big) \quad &\geq \quad 0. \end{aligned}$$

Proof.

$\underline{\mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) \;-\; \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) \geq 0}$

We have, $\mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) = \mathsf{var}\big(\mathbb{H}_X \boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \mathbb{H}_X(\sigma^2 \mathbf{I}_n)\mathbb{H}_X$

$= \sigma^2\, \mathbb{H}_X$ (even if M_X is not correct).

$$\begin{aligned} \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) &= \mathsf{var}\big(\mathbb{H}_{XV} \boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \sigma^2 \mathbb{H}_{XV} \\ &= \sigma^2 \left\{ \mathbb{H}_X \;+\; \mathbb{M}_X \mathbb{V} (\mathbb{V}^\top \mathbb{M}_X \mathbb{V})^{-1} \mathbb{V}^\top \mathbb{M}_X \right\} \\ &= \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) \;+\; \underbrace{\sigma^2\, \mathbb{M}_X \mathbb{V} (\mathbb{V}^\top \mathbb{M}_X \mathbb{V})^{-1} \mathbb{V}^\top \mathbb{M}_X}_{\text{positive semidefinite matrix}}. \end{aligned}$$

$\underline{\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) \;-\; \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X} \,\big|\, \mathbb{Z}\big) \geq 0}$

Proof/calculations for this part were skipped and are not requested for the exam. Proof/calculations below are shown only for those who are interested.

First, use a formula to calculate an inverse of a matrix divided into blocks (Theorem A.4):

$$\mathsf{var}\left\{ \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{XV} \\ \widehat{\boldsymbol{\gamma}}_{XV} \end{pmatrix} \,\middle|\, \mathbb{Z} \right\} = \sigma^2 \begin{pmatrix} \mathbb{X}^\top \mathbb{X} & \mathbb{X}^\top \mathbb{V} \\ \mathbb{V}^\top \mathbb{X} & \mathbb{V}^\top \mathbb{V} \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} \left\{ \mathbb{X}^\top \mathbb{X} - \mathbb{X}^\top \mathbb{V} \big(\mathbb{V}^\top \mathbb{V}\big)^{-1} \mathbb{V}^\top \mathbb{X} \right\}^{-1} & \circledast \\ \circledast & \circledast \end{pmatrix}.$$

Further,

$$\begin{aligned} \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X} \,\big|\, \mathbb{Z}\big) &= \mathsf{var}\Big(\big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \mathbb{X}^\top \boldsymbol{Y} \,\Big|\, \mathbb{Z} \Big) = \big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \mathbb{X}^\top (\sigma^2 \mathbf{I}_n) \mathbb{X} \big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \\ &= \sigma^2 \big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \qquad \text{(even if } \mathsf{M}_X \text{ is not correct).} \end{aligned}$$

$$\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) = \sigma^2 \left\{ \mathbb{X}^\top \mathbb{X} - \mathbb{X}^\top \mathbb{V} \big(\mathbb{V}^\top \mathbb{V}\big)^{-1} \mathbb{V}^\top \mathbb{X} \right\}^{-1}.$$

Property of positive definite matrices ("$\mathbb{A} - \mathbb{B} \geq 0 \;\Leftrightarrow\; \mathbb{B}^{-1} - \mathbb{A}^{-1} \geq 0$") finalizes the proof. ❑

Notes.

- Estimator of the response mean vector $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ based on a (smaller) model M_X is *always* (does not matter which model is correct) less or equally variable than the estimator based on the (richer) model M_{XV}.
- Estimators of the regression coefficients $\boldsymbol{\beta}$ based on a (smaller) model M_X have *always* lower (or equal if $\mathbb{X}^\top\mathbb{V} = \mathbf{0}_{k\times m}$) standard errors than the estimator based on the (richer) model M_{XV}.

12.2.2 Prediction quality of the fitted model

To evaluate a prediction quality of the *fitted* model, we will assume that data $\big(Y_i,\, \boldsymbol{Z}_i^\top\big)^\top$, $\boldsymbol{Z}_i = \big(Z_{i,1},\, \ldots,\, Z_{i,p}\big)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$, are a random sample from a distribution of a generic random vector $\big(Y,\, \boldsymbol{Z}^\top\big)^\top$, $\boldsymbol{Z} = \big(Z_1,\, \ldots,\, Z_p\big)^\top$. Let the conditional distribution $Y \,|\, \boldsymbol{Z}$ of Y given the covariates $\boldsymbol{Z}$ satisfies

$$\mathbb{E}\big(Y \,\big|\, \boldsymbol{Z}\big) = m(\boldsymbol{Z}), \qquad \mathsf{var}\big(Y \,\big|\, \boldsymbol{Z}\big) = \sigma^2, \tag{12.9}$$

for some (regression) function m and some $\sigma^2 > 0$.

Replicated response

Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be the values of the covariate vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ in the original data that are available to estimate the parameters of the model (12.9). Further, let $\big(Y_{n+i},\, \boldsymbol{Z}_{n+i}^\top\big)^\top$, $i = 1, \ldots, n$, be independent random vectors (new or future data) being distributed as a generic random vector $\big(Y,\, \boldsymbol{Z}\big)$ and being independent of the original data $\big(Y_i,\, \boldsymbol{Z}_i^\top\big)^\top$, $i = 1, \ldots, n$. Suppose that our aim is to *predict* values of Y_{n+i}, $i = 1, \ldots, n$, under the condition that the new covariate values are equal to the old ones. That is, we want to predict, for $i = 1, \ldots, n$, values of Y_{n+i} given $\boldsymbol{Z}_{n+i} = \boldsymbol{z}_i$.

***Terminology** (Replicated response).*

A random vector

$$\boldsymbol{Y}_{new} = \big(Y_{n+1},\, \ldots,\, Y_{n+n}\big)^\top,$$

where Y_{n+i} is supposed to come from the conditional distribution $Y \,|\, \boldsymbol{Z} = \boldsymbol{z}_i$, $i = 1, \ldots, n$, is called the *replicated response* vector or *replicated data*.

Notes.

- The original (old) response vector $\boldsymbol{Y}$ and the replicated response vector $\boldsymbol{Y}_{new}$ are assumed to be independent.
- Both $\boldsymbol{Y}$ and $\boldsymbol{Y}_{new}$ are assumed to be generated by the same conditional distribution (given $\mathbb{Z}$), where

$$\begin{aligned}
\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \boldsymbol{Z}_1 = \boldsymbol{z}_1,\, \ldots,\, \boldsymbol{Z}_n = \boldsymbol{z}_n\big) &= \quad \boldsymbol{\mu} \quad &&= \mathbb{E}\big(\boldsymbol{Y}_{new} \,\big|\, \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1,\, \ldots,\, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\big),\\
\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \boldsymbol{Z}_1 = \boldsymbol{z}_1,\, \ldots,\, \boldsymbol{Z}_n = \boldsymbol{z}_n\big) &= \sigma^2\mathbf{I}_n &&= \mathsf{var}\big(\boldsymbol{Y}_{new} \,\big|\, \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1,\, \ldots,\, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\big),\\
& & & \qquad \text{for some } \sigma^2 > 0,
\end{aligned}$$

and

$$\boldsymbol{\mu} = \big(m(\boldsymbol{z}_1),\, \ldots,\, m(\boldsymbol{z}_n)\big)^\top = \big(\mu_1,\, \ldots,\, \mu_n\big)^\top.$$

Prediction of replicated response

Let

$$\widehat{\boldsymbol{Y}}_{new} = \left(\widehat{Y}_{n+1}, \ldots, \widehat{Y}_{n+n}\right)^\top$$

be the prediction of a vector $\boldsymbol{Y}_{new}$ based on the assumed regression model (12.9) estimated using the original data $\boldsymbol{Y}$ with $\boldsymbol{Z}_1 = \boldsymbol{z}_1, \ldots, \boldsymbol{Z}_n = \boldsymbol{z}_n$. That is, $\widehat{\boldsymbol{Y}}_{new}$ is some statistic of $\boldsymbol{Y}$ (and $\mathbb{Z}$). Analogously to Section 5.4.3, we shall evaluate a quality of the prediction by the mean squared error of prediction (MSEP). Nevertheless, in contrast to Section 5.4.3, the following issues will be different:

(i) A value of a random vector rather than a value of a random variable (as in Section 5.4.3) is predicted now. Now, the MSEP will be given as a sum of the MSEPs of the elements of the random vector being predicted.

(ii) Since we are now interested in prediction of new response values given the covariate values being equal to the covariate values in the original data, the MSEP now will be based on a conditional distribution of the responses given $\mathbb{Z}$ (given $\boldsymbol{Z}_i = \boldsymbol{Z}_{n+i} = \boldsymbol{z}_i$, $i = 1, \ldots, n$). In contrast, variability of the covariates was taken into account in Section 5.4.3.

(iii) Variability of the prediction induced by estimation of the model parameters (estimation of the regression function) using the original data $\boldsymbol{Y}$ will also be taken into account now. In contrast, model parameters were assumed to be known when deriving the MSEP in Section 5.4.3.

Definition 12.2 Quantification of a prediction quality of the fitted regression model.

Prediction quality of the fitted regression model will be evaluated by the mean squared error of prediction (MSEP)[4] *defined as*

$$\mathsf{MSEP}\left(\widehat{\boldsymbol{Y}}_{new}\right) = \sum_{i=1}^{n} \mathbb{E}\left\{\left(\widehat{Y}_{n+i} - Y_{n+i}\right)^2 \,\middle|\, \mathbb{Z}\right\}, \tag{12.10}$$

where the expectation is with respect to the $(n+n)$*-dimensional conditional distribution of the vector* $\left(\boldsymbol{Y}^\top, \boldsymbol{Y}_{new}^\top\right)^\top$ *given*

$$\mathbb{Z} = \begin{pmatrix} \boldsymbol{Z}_1^\top \\ \vdots \\ \boldsymbol{Z}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_{n+1}^\top \\ \vdots \\ \boldsymbol{Z}_{n+n}^\top \end{pmatrix}.$$

Additionally, we define the averaged mean squared error of prediction (AMSEP)[5] *as*

$$\mathsf{AMSEP}\left(\widehat{\boldsymbol{Y}}_{new}\right) = \frac{1}{n}\,\mathsf{MSEP}\left(\widehat{\boldsymbol{Y}}_{new}\right).$$

---

[4] *střední čtvercová chyba predikce* [5] *průměrná střední čtvercová chyba predikce*

Prediction of replicated response in a linear model

With a linear model, it is assumed that $m(\boldsymbol{z}) = \boldsymbol{x}^\top\boldsymbol{\beta}$ for some (known) transformation $\boldsymbol{x} = \boldsymbol{t}_X(\boldsymbol{z})$ and a vector of (unknown) parameters $\boldsymbol{\beta}$. Hence, it is *assumed* that

$$\begin{aligned}\boldsymbol{\mu} &= \big(\mu_1, \dots, \mu_n\big)^\top \\ &= \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \boldsymbol{Z}_1 = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_n = \boldsymbol{z}_n\big) = \mathbb{E}\big(\boldsymbol{Y}_{new} \,\big|\, \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\big)\end{aligned}$$

satisfies

$$\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta} = \big(\boldsymbol{x}_1^\top\boldsymbol{\beta}, \dots, \boldsymbol{x}_n^\top\boldsymbol{\beta}\big)^\top,$$

for a model matrix $\mathbb{X}$ based on the (transformed) covariate values $\boldsymbol{x}_i = \boldsymbol{t}_X(\boldsymbol{z}_i)$, $i = 1, \dots, n$.

If we restrict our attention to *unbiased* and *linear* predictions of $\boldsymbol{Y}_{new}$, i.e., to predictions of the form $\widehat{\boldsymbol{Y}}_{new} = \boldsymbol{a} + \mathbb{A}\boldsymbol{Y}$ for some vector $\boldsymbol{a} \in \mathbb{R}^n$ and some $n \times n$ matrix $\mathbb{A}$ satisfying $\mathbb{E}\big(\widehat{\boldsymbol{Y}}_{new} \,\big|\, \mathbb{Z}\big) = \mathbb{E}\big(\boldsymbol{Y}_{new} \,\big|\, \mathbb{Z}\big) = \boldsymbol{\mu}$, a variant of the Gauss-Markov theorem would show that (12.10) is minimized for

$$\begin{aligned}\widehat{\boldsymbol{Y}}_{new} &= \widehat{\boldsymbol{Y}}, \qquad & \widehat{\boldsymbol{Y}} &= \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top\boldsymbol{Y}, \\ \widehat{Y}_{n+i} &= \widehat{Y}_i, \qquad & i &= 1, \dots, n.\end{aligned}$$

That is, for $\widehat{\boldsymbol{Y}}_{new}$ being equal to the fitted values of the model estimated using the original data. Note also that

$$\widehat{\boldsymbol{Y}}_{new} = \widehat{\boldsymbol{Y}} =: \widehat{\boldsymbol{\mu}},$$

where $\widehat{\boldsymbol{\mu}}$ is the LSE of a vector $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \boldsymbol{Z}_1 = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_n = \boldsymbol{z}_n\big) = \mathbb{E}\big(\boldsymbol{Y}_{new} \,\big|\, \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\big)$.

Lemma 12.4 Mean squared error of the BLUP in a linear model.

In a linear model, the mean squared error of the best linear unbiased prediction can be expressed as

$$\mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new}\big) = n\,\sigma^2 + \sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_i\big),$$

where

$$\mathsf{MSE}\big(\widehat{Y}_i\big) = \mathbb{E}\Big\{\big(\widehat{Y}_i - \mu_i\big)^2 \,\Big|\, \mathbb{Z}\Big\}, \qquad i = 1, \dots, n,$$

is the mean squared error[6] *of $\widehat{Y}_i$ if this is viewed as estimator of μ_i, $i = 1, \dots, n$.*

Proof. To simplify notation, condition will be omitted from notation of all expectations and variances. Nevertheless, all are still understood as conditional expectations and variances given the covariate values $\mathbb{Z}$.

We have for $i = 1, \dots, n$ (remember, $\widehat{Y}_{n+i} = \widehat{Y}_i$, $i = 1, \dots, n$),

[6] *střední čtvercová chyba*

$$\begin{aligned}
\mathbb{E}\big(\widehat{Y}_{n+i} - Y_{n+i}\big)^2 &= \mathbb{E}\big(\widehat{Y}_i - Y_{n+i}\big)^2 \\
&= \mathbb{E}\big\{\widehat{Y}_i - \mu_i - (Y_{n+i} - \mu_i)\big\}^2 \\
&= \mathbb{E}\big(\widehat{Y}_i - \mu_i\big)^2 + \mathbb{E}\big(Y_{n+i} - \mu_i\big)^2 - 2 \quad \underbrace{\mathbb{E}(\widehat{Y}_i - \mu_i)(Y_{n+i} - \mu_i)}_{\mathbb{E}(\widehat{Y}_i - \mu_i)\,\mathbb{E}(Y_{n+i} - \mu_i) = \mathbb{E}(\widehat{Y}_i - \mu_i)\cdot 0} \\
&= \mathbb{E}\big(\widehat{Y}_i - \mu_i\big)^2 + \mathbb{E}\big(Y_{n+i} - \mu_i\big)^2 \\
&= \mathsf{MSE}(\widehat{Y}_i) + \sigma^2.
\end{aligned}$$

So that

$$\mathsf{MSEP}(\widehat{\boldsymbol{Y}}_{new}) = \sum_{i=1}^{n} \mathbb{E}\big(\widehat{Y}_{n+i} - Y_{n+i}\big)^2 = n\,\sigma^2 \,+\, \sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_i\big).$$

❑

Notes.

- We can also write

$$\sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_i\big) = \mathbb{E}\Big\{ \big\|\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\big\|^2 \,\Big|\, \mathbb{Z}\Big\}.$$

  Hence,

$$\mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new}\big) \;=\; n\,\sigma^2 \,+\, \mathbb{E}\Big\{ \big\|\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\big\|^2 \,\Big|\, \mathbb{Z}\Big\}.$$

- If the assumed linear model is a correct model for data at hand, Gauss-Markov theorem states that $\widehat{\boldsymbol{Y}}$ is the BLUE of the vector $\boldsymbol{\mu}$ in which case

$$\mathsf{MSE}\big(\widehat{Y}_i\big) = \mathbb{E}\Big\{ \big(\widehat{Y}_i - \mu_i\big)^2 \,\Big|\, \mathbb{Z}\Big\} = \mathsf{var}\big(\widehat{Y}_i \,\big|\, \mathbb{Z}\big), \qquad i = 1, \ldots, n.$$

- Nevertheless, if the assumed linear model is not a correct model for data at hand, estimator $\widehat{\boldsymbol{Y}}$ might be a biased estimator of the vector $\boldsymbol{\mu}$, in which case

$$\begin{aligned}
\mathsf{MSE}\big(\widehat{Y}_i\big) &= \mathbb{E}\Big\{ \big(\widehat{Y}_i - \mu_i\big)^2 \,\Big|\, \mathbb{Z}\Big\} \\
&= \mathsf{var}\big(\widehat{Y}_i \,\big|\, \mathbb{Z}\big) + \Big\{\mathbb{E}\big(\widehat{Y}_i - \mu_i \,\big|\, \mathbb{Z}\big)\Big\}^2 = \mathsf{var}\big(\widehat{Y}_i \,\big|\, \mathbb{Z}\big) + \Big\{\mathsf{bias}\big(\widehat{Y}_i\big)\Big\}^2, \qquad i = 1, \ldots, n.
\end{aligned}$$

- Expression of the mean squared error of prediction is

$$\mathsf{MSEP}(\widehat{\boldsymbol{Y}}_{new}) = n\,\sigma^2 \,+\, \sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_i\big) = n\,\sigma^2 \,+\, \mathbb{E}\Big\{ \big\|\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\big\|^2 \,\Big|\, \mathbb{Z}\Big\}.$$

  By specification of a model for the conditional response expectation, i.e., by specification of a model for $\boldsymbol{\mu}$, we can influence only the second factor $\mathbb{E}\Big\{ \big\|\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\big\|^2 \,\Big|\, \mathbb{Z}\Big\}$. The first factor ($n\,\sigma^2$) reflects the true (conditional) variability of the response which does not depend on specification of the model for the expectation. Hence, if evaluating a prediction quality of a linear model with respect to ability to predict replicated data, the only term that matters is

$$\sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_i\big) = \mathbb{E}\Big\{ \big\|\widehat{\boldsymbol{Y}} - \boldsymbol{\mu}\big\|^2 \,\Big|\, \mathbb{Z}\Big\},$$

  that relates to the error of the fitted values being considered as an estimator of the vector $\boldsymbol{\mu}$.

12.2.3 Omitted regressors

In this section, we will assume that the correct model is model

$$\mathsf{M}_{XV}: \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big),$$

with $\boldsymbol{\gamma} \neq \mathbf{0}_l$. Hence all estimators derived under model M_{XV} are derived under the correct model and hence have usual properties of the LSE, namely,

$$\begin{aligned}
\mathbb{E}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) &= \boldsymbol{\beta}, \\
\mathbb{E}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) &= \mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma} =: \boldsymbol{\mu}, \\
\sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_{XV,i}\big) &= \sum_{i=1}^{n} \mathsf{var}\big(\widehat{Y}_{XV,i} \,\big|\, \mathbb{Z}\big) \;=\; \mathsf{tr}\Big(\mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big)\Big) \;=\; \mathsf{tr}\big(\sigma^2\,\mathbb{H}_{XV}\big) \\
&= \sigma^2\,(k+l), \\
\mathbb{E}\big(\mathsf{MS}_{e,XV} \,\big|\, \mathbb{Z}\big) &= \sigma^2.
\end{aligned} \tag{12.11}$$

Nevertheless, all estimators derived under model $\mathsf{M}_X: \; \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ are calculated while assuming a misspecified model with omitted important regressors and their properties do not coincide with properties of the LSE calculated under the correct model.

Theorem 12.5 Properties of the LSE in a model with omitted regressors.
Let $\mathsf{M}_{XV}: \; \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big)$ *hold, i.e.,* $\boldsymbol{\mu} := \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ *satisfies*

$$\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma}$$

for some $\boldsymbol{\beta} \in \mathbb{R}^k$, $\boldsymbol{\gamma} \in \mathbb{R}^l$.

Then the least squares estimators derived while assuming model $\mathsf{M}_X: \; \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *attain the following properties:*

$$\begin{aligned}
\mathbb{E}\big(\widehat{\boldsymbol{\beta}}_X \,\big|\, \mathbb{Z}\big) &= \boldsymbol{\beta} \;+\; \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma}, \\
\mathbb{E}\big(\widehat{\boldsymbol{Y}}_X \,\big|\, \mathbb{Z}\big) &= \boldsymbol{\mu} \;-\; \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}, \\
\sum_{i=1}^{n} \mathsf{MSE}\big(\widehat{Y}_{X,i}\big) &= k\,\sigma^2 \;+\; \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2, \\
\mathbb{E}\big(\mathsf{MS}_{e,X} \,\big|\, \mathbb{Z}\big) &= \sigma^2 \;+\; \frac{\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2}{n-k}.
\end{aligned}$$

Proof. As several times before, condition will be omitted from notation of all expectations and variances that appear in the proof. Nevertheless, all are still understood as conditional expectations and variances given the covariate values $\mathbb{Z}$.

$\underline{\mathbb{E}(\widehat{\boldsymbol{\beta}}_X \,|\, \mathbb{Z})}$

By Theorem 11.1: $\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X = -\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}$.

$$\begin{aligned}\text{Hence,}\quad \mathbb{E}(\widehat{\boldsymbol{\beta}}_X) &= \mathbb{E}\Big\{\widehat{\boldsymbol{\beta}}_{XV} + \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}\Big\}\\ &= \boldsymbol{\beta} + \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma},\\ \mathsf{bias}(\widehat{\boldsymbol{\beta}}_X) &= \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma}.\end{aligned}$$

$\underline{\mathbb{E}(\widehat{\boldsymbol{Y}}_X \,|\, \mathbb{Z})}$

By Theorem 11.1: $\widehat{\boldsymbol{Y}}_{XV} - \widehat{\boldsymbol{Y}}_X = \mathbb{X}(\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X) + \mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}$.

$$\begin{aligned}\text{Hence,}\quad \mathbb{E}(\widehat{\boldsymbol{Y}}_X) &= \mathbb{E}(\widehat{\boldsymbol{Y}}_{XV} - \mathbb{X}\widehat{\boldsymbol{\beta}}_{XV} + \mathbb{X}\widehat{\boldsymbol{\beta}}_X - \mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV})\\ &= \boldsymbol{\mu} - \mathbb{X}\boldsymbol{\beta} + \mathbb{X}\boldsymbol{\beta} + \mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma} - \mathbb{V}\boldsymbol{\gamma}\\ &= \boldsymbol{\mu} + \Big\{\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top - \mathbf{I}_n\Big\}\mathbb{V}\boldsymbol{\gamma}\\ &= \boldsymbol{\mu} - \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma},\\ \mathsf{bias}(\widehat{\boldsymbol{Y}}_X) &= -\,\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}.\end{aligned}$$

$\underline{\sum_{i=1}^n \mathsf{MSE}(\widehat{Y}_{X,i})}$

Let us first calculate $\mathsf{MSE}(\widehat{\boldsymbol{Y}}_X) = \mathbb{E}\Big\{(\widehat{\boldsymbol{Y}}_X - \boldsymbol{\mu})(\widehat{\boldsymbol{Y}}_X - \boldsymbol{\mu})^\top\Big\}$:

$$\begin{aligned}\mathsf{MSE}(\widehat{\boldsymbol{Y}}_X) &= \mathsf{var}(\widehat{\boldsymbol{Y}}_X) + \mathsf{bias}(\widehat{\boldsymbol{Y}}_X)\mathsf{bias}^\top(\widehat{\boldsymbol{Y}}_X)\\ &= \sigma^2\mathbb{H}_X + \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X.\end{aligned}$$

$$\begin{aligned}\text{Hence,}\quad \sum_{i=1}^n \mathsf{MSE}(\widehat{Y}_{X,i}) &= \mathsf{tr}\Big(\mathsf{MSE}(\widehat{\boldsymbol{Y}}_X)\Big)\\ &= \mathsf{tr}\big(\sigma^2\mathbb{H}_X + \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\big)\\ &= \mathsf{tr}\big(\sigma^2\mathbb{H}_X\big) + \mathsf{tr}\big(\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\big)\\ &= \sigma^2\,k + \mathsf{tr}\big(\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big)\\ &= \sigma^2\,k + \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2.\end{aligned}$$

$\underline{\mathbb{E}(\mathsf{MS}_{e,X} \,|\, \mathbb{Z})}$

Proof/calculations for this part were skipped and are not requested for the exam. Proof/calculations below are shown only for those who are interested.

Let us first calculate $\mathbb{E}(\mathsf{SS}_{e,X}) := \mathbb{E}(\mathsf{SS}_{e,X} \,|\, \mathbb{Z})$. To do that, write the linear model M_{XV} using the error terms as

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon} \,|\, \mathbb{Z}) = \mathbf{0}_n, \quad \mathsf{var}(\boldsymbol{\varepsilon} \,|\, \mathbb{Z}) = \sigma^2\mathbf{I}_n.$$

$$\begin{aligned}
\mathbb{E}\big(\mathsf{SS}_{e,X}\big) &= \mathbb{E}\big\|\mathbb{M}_X\boldsymbol{Y}\big\|^2 = \mathbb{E}\big\|\mathbb{M}_X(\mathbb{X}\boldsymbol{\beta}+\mathbb{V}\boldsymbol{\gamma}+\boldsymbol{\varepsilon})\big\|^2 \\
&= \mathbb{E}\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}+\mathbb{M}_X\boldsymbol{\varepsilon}\big\|^2 \\
&= \mathbb{E}\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2 + \mathbb{E}\big\|\mathbb{M}_X\boldsymbol{\varepsilon}\big\|^2 + 2\underbrace{\mathbb{E}\big(\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\mathbb{M}_X\boldsymbol{\varepsilon}\big)}_{\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\,\mathbb{E}\boldsymbol{\varepsilon}=0} \\
&= \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2 + \underbrace{\mathbb{E}\big(\boldsymbol{\varepsilon}^\top\mathbb{M}_X\boldsymbol{\varepsilon}\big)}_{\mathbb{E}\left(\mathsf{tr}(\boldsymbol{\varepsilon}^\top\mathbb{M}_X\boldsymbol{\varepsilon})\right)=\mathsf{tr}\left(\mathbb{E}(\mathbb{M}_X\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\right)=\mathsf{tr}\left(\sigma^2\,\mathbb{M}_X\right)=\sigma^2\,(n-k)} \\
&= \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2 + \sigma^2\,(n-k).
\end{aligned}$$

Hence,
$$\begin{aligned}
\mathbb{E}\big(\mathsf{MS}_{e,X}\big) &= \mathbb{E}\left(\frac{\mathsf{SS}_{e,X}}{n-k}\right) \\
&= \sigma^2 + \frac{\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2}{n-k},
\end{aligned}$$

$$\mathsf{bias}\big(\mathsf{MS}_{e,X}\big) = \frac{\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2}{n-k}.$$

❑

Least squares estimators

Theorem 12.5 shows that $\mathsf{bias}\big(\widehat{\boldsymbol{\beta}}_X\big) = \mathbb{E}\big(\widehat{\boldsymbol{\beta}}_X - \boldsymbol{\beta}\,\big|\,\mathbb{Z}\big) = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma}$, nevertheless, the estimator $\widehat{\boldsymbol{\beta}}_X$ is not necessarily biased. Let us consider two situations.

(i) $\mathbb{X}^\top\mathbb{V} = \mathbf{0}_{k\times l}$, which means that each column of $\mathbb{X}$ is *orthogonal* with each column in $\mathbb{V}$. In other words, regressors included in the matrix $\mathbb{X}$ are *uncorrelated* with regressors included in the matrix $\mathbb{V}$. Then
- $\widehat{\boldsymbol{\beta}}_X = \widehat{\boldsymbol{\beta}}_{XV}$ and $\mathsf{bias}\big(\widehat{\boldsymbol{\beta}}_X\big) = \mathbf{0}_k$.
- Hence $\boldsymbol{\beta}$ can be estimated using the smaller model M_X without any impact on a quality of the estimator.

(ii) $\mathbb{X}^\top\mathbb{V} \neq \mathbf{0}_{k\times l}$
- $\widehat{\boldsymbol{\beta}}_X$ is a *biased* estimator of $\boldsymbol{\beta}$.

Further, for the fitted values $\widehat{\boldsymbol{Y}}_X$ if those are considered as an estimator of the response vector expectation $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}+\mathbb{V}\boldsymbol{\gamma}$, we have

$$\mathsf{bias}\big(\widehat{\boldsymbol{Y}}_X\big) = -\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}.$$

In this case, all elements of the bias vector would be equal to zero if $\mathbb{M}_X\mathbb{V} = \mathbf{0}_{n\times l}$. Nevertheless, this would mean that $\mathcal{M}\big(\mathbb{V}\big) \subseteq \mathcal{M}\big(\mathbb{X}\big)$ which is in contradition with our assumption $\mathsf{rank}\big(\mathbb{X},\,\mathbb{V}\big) = k+l$. That is, if the omitted covariates (included in the matrix $\mathbb{V}$) are linearly independent (are not perfectly multiply correlated) with the covariates included in the model matrix $\mathbb{X}$, the fitted values $\widehat{\boldsymbol{Y}}_X$ always provide a biased estimator of the response expectation.

Prediction

Let us compare predictions $\widehat{\boldsymbol{Y}}_{new,X} = \widehat{\boldsymbol{Y}}_X$ based on a (misspecified) model M_X and predictions $\widehat{\boldsymbol{Y}}_{new,XV} = \widehat{\boldsymbol{Y}}_{XV}$ based on a (correct) model M_{XV}. Properties of the fitted values in a correct model (Expressions (12.11)) together with results of Lemma 12.4 and Theorem 12.5 give

$$\begin{aligned}\mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new,XV}\big) &= n\,\sigma^2 \,+\, k\,\sigma^2 \,+\, l\,\sigma^2,\\ \mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new,X}\big) &= n\,\sigma^2 \,+\, k\,\sigma^2 \,+\, \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2.\end{aligned}$$

That is, the average mean squared errors of prediction are

$$\begin{aligned}\mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new,XV}\big) &= \sigma^2 \,+\, \frac{k}{n}\,\sigma^2 \,+\, \frac{l}{n}\,\sigma^2,\\ \mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new,X}\big) &= \sigma^2 \,+\, \frac{k}{n}\,\sigma^2 \,+\, \frac{1}{n}\,\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2.\end{aligned}$$

We can now conclude the following.

- The term $\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2$ might be huge compared to $l\,\sigma^2$ in which case the prediction using the model with omitted important covariates is (much) worse than the prediction using the (correct) model.
- Additionally, $\frac{l}{n}\,\sigma^2 \to 0$ with $n \to \infty$ (while increasing the number of predictions).
- On the other hand, $\frac{1}{n}\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2$ does not necessarily tend to zero with $n \to \infty$.

Estimator of the residual variance

Theorem 12.5 shows that the mean residual square $\mathsf{MS}_{e,X}$ in a misspecified model M_X is a biased estimator of the residual variance σ^2 with the bias amounting to

$$\mathsf{bias}\big(\mathsf{MS}_{e,X}\big) \;=\; \mathbb{E}\big(\mathsf{MS}_{e,X} - \sigma^2 \,\big|\, \mathbb{Z}\big) \;=\; \frac{\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2}{n-k}.$$

Also in this case, bias does not necessarily tend to zero with $n \to \infty$.

12.2.4 Irrelevant regressors

In this section, we will assume that the correct model is model

$$\mathsf{M}_X\colon \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big).$$

This means, that also model

$$\mathsf{M}_{XV}\colon \quad \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big)$$

holds, nevertheless, $\boldsymbol{\gamma} = \mathbf{0}_l$ and hence the regressors from the matrix $\mathbb{V}$ are *irrelevant*.

Since both models M_X and M_{XV} hold, estimators derived under both models have usual properties of the LSE, namely,

$$\begin{aligned}
\mathbb{E}(\widehat{\boldsymbol{\beta}}_X \,|\, \mathbb{Z}) = \mathbb{E}(\widehat{\boldsymbol{\beta}}_{XV} \,|\, \mathbb{Z}) &= \boldsymbol{\beta}, \\
\mathbb{E}(\widehat{\boldsymbol{Y}}_X \,|\, \mathbb{Z}) = \mathbb{E}(\widehat{\boldsymbol{Y}}_{XV} \,|\, \mathbb{Z}) &= \mathbb{X}\boldsymbol{\beta} =: \boldsymbol{\mu}, \\
\sum_{i=1}^{n} \mathsf{MSE}(\widehat{Y}_{X,i}) &= \sum_{i=1}^{n} \mathsf{var}(\widehat{Y}_{X,i} \,|\, \mathbb{Z}) = \mathsf{tr}\Big(\mathsf{var}(\widehat{\boldsymbol{Y}}_X \,|\, \mathbb{Z})\Big) = \mathsf{tr}(\sigma^2\, \mathbb{H}_X) \\
&= \sigma^2\, k, \\
\sum_{i=1}^{n} \mathsf{MSE}(\widehat{Y}_{XV,i}) &= \sum_{i=1}^{n} \mathsf{var}(\widehat{Y}_{XV,i} \,|\, \mathbb{Z}) = \mathsf{tr}\Big(\mathsf{var}(\widehat{\boldsymbol{Y}}_{XV} \,|\, \mathbb{Z})\Big) = \mathsf{tr}(\sigma^2\, \mathbb{H}_{XV}) \\
&= \sigma^2\, (k + l), \\
\mathbb{E}(\mathsf{MS}_{e,X} \,|\, \mathbb{Z}) = \mathbb{E}(\mathsf{MS}_{e,XV} \,|\, \mathbb{Z}) &= \sigma^2.
\end{aligned}$$

Least squares estimators

Both estimators $\widehat{\boldsymbol{\beta}}_X$ and $\widehat{\boldsymbol{\beta}}_{XV}$ are unbiased estimators of a vector $\boldsymbol{\beta}$. Nevertheless, as stated in Theorem 12.3, their quality expressed by the mean squared error which in this case coincide with the covariance matrix (may) differ since

$$\begin{aligned}
\mathsf{MSE}(\widehat{\boldsymbol{\beta}}_{XV}) - \mathsf{MSE}(\widehat{\boldsymbol{\beta}}_{XV}) &= \mathbb{E}\Big\{(\widehat{\boldsymbol{\beta}}_{XV} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_{XV} - \boldsymbol{\beta})^\top \,\Big|\, \mathbb{Z}\Big\} - \mathbb{E}\Big\{(\widehat{\boldsymbol{\beta}}_X - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_X - \boldsymbol{\beta})^\top \,\Big|\, \mathbb{Z}\Big\} \\
&= \mathsf{var}(\widehat{\boldsymbol{\beta}}_{XV} \,|\, \mathbb{Z}) \;-\; \mathsf{var}(\widehat{\boldsymbol{\beta}}_X \,|\, \mathbb{Z}) \geq 0.
\end{aligned}$$

In particular, we derived during the proof of Theorem 12.3 that

$$\mathsf{var}(\widehat{\boldsymbol{\beta}}_{XV} \,|\, \mathbb{Z}) \;-\; \mathsf{var}(\widehat{\boldsymbol{\beta}}_X \,|\, \mathbb{Z}) = \sigma^2 \Big[\Big\{\mathbb{X}^\top\mathbb{X} - \mathbb{X}^\top\mathbb{V}(\mathbb{V}^\top\mathbb{V})^{-1}\mathbb{V}^\top\mathbb{X}\Big\}^{-1} - (\mathbb{X}^\top\mathbb{X})^{-1}\Big].$$

Let us again consider two situations.

(i) $\mathbb{X}^\top\mathbb{V} = \mathbf{0}_{k\times l}$, which means that each column of $\mathbb{X}$ is *orthogonal* with each column in $\mathbb{V}$. In other words, regressors included in the matrix $\mathbb{X}$ are *uncorrelated* with regressors included in the matrix $\mathbb{V}$. Then

- $\widehat{\boldsymbol{\beta}}_X = \widehat{\boldsymbol{\beta}}_{XV}$ and $\mathsf{var}(\widehat{\boldsymbol{\beta}}_X \,|\, \mathbb{Z}) = \mathsf{var}(\widehat{\boldsymbol{\beta}}_{XV} \,|\, \mathbb{Z})$.
- Hence $\boldsymbol{\beta}$ can be estimated using the model M_{XV} with irrelevant covariates included without any impact on a quality of the estimator.

(ii) $\mathbb{X}^\top\mathbb{V} \neq \mathbf{0}_{k\times l}$

- The estimator $\widehat{\boldsymbol{\beta}}_{XV}$ is worse than the estimator $\widehat{\boldsymbol{\beta}}_X$ in terms of its variability.
- If we take into account a fact that by including more regressors in the model, we are increasing a danger of multicollinearity, difference between variability of $\widehat{\boldsymbol{\beta}}_{XV}$ and that of $\widehat{\boldsymbol{\beta}}_X$ may become huge.

Prediction

Let us now compare predictions $\widehat{\boldsymbol{Y}}_{new,X} = \widehat{\boldsymbol{Y}}_X$ based on a correct model M_X and predictions $\widehat{\boldsymbol{Y}}_{new,XV} = \widehat{\boldsymbol{Y}}_{XV}$ based on also a correct model M_{XV}, where however, irrelevant covariates were included. Properties of the fitted values in a correct model together with results of Lemma 12.4 give

$$\begin{aligned}\mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new,XV}\big) &= n\,\sigma^2 + (k+l)\,\sigma^2,\\ \mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new,X}\big) &= n\,\sigma^2 + k\,\sigma^2.\end{aligned}$$

That is, the average mean squared errors of prediction are

$$\begin{aligned}\mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new,XV}\big) &= \sigma^2 + \frac{k+l}{n}\,\sigma^2,\\ \mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new,X}\big) &= \sigma^2 + \frac{k}{n}\,\sigma^2.\end{aligned}$$

The following can now be concluded.

- If $n \to \infty$, both $\mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new,XV}\big)$ and $\mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new,X}\big)$ tend to σ^2. Hence on average, if sufficiently large number of predictions is needed, both models provide predictions of practically the same quality.
- On the other hand, by using the richer model M_{XV} (which for a finite n provides worse predictions than the smaller model M_X), we are eliminating a possible problem of omitted important covariates that leads to biased predictions with possibly even worse MSEP and AMSEP than that of model M_{XV}.

12.2.5 Summary

Interest in estimation of the regression coefficients and inference on them

If interest lies in estimation of and inference on the regression coefficients $\boldsymbol{\beta}$ related to the regressors included in the model matrix $\mathbb{X}$, the following was derived in Sections 12.2.3 and 12.2.4.

(i) If we omit important regressors which are (multiply) *correlated* with the regressors of main interest included in the matrix $\mathbb{X}$, the LSE of the regression coefficients is *biased*.

(ii) If we include irrelevant regressors which are (multiply) *correlated* with the regressors of main interest in the matrix $\mathbb{X}$, we are facing a danger of multicollinearity and related inflation of the standard errors of the LSE of the regression coefficients.

(iii) Regressors which are (multiply) *uncorrelated* with regressors of main interest influence neither bias nor variability of $\widehat{\boldsymbol{\beta}}$ irrespective of whether they are omitted or irrelevantly included.

Consequently, if a primary task of the analysis is to evaluate whether and how much the primary regressors included in the model matrix $\mathbb{X}$ influence the response expectation, detailed exploration and understanding of mutual relationships among all potential regressors and also between the regressors and the response is needed. In particular, regressors which are (multiply) correlated with the regressors from the model matrix $\mathbb{X}$ and at the same time do not have any influence on the response expectation should not be included in the model. On the other hand, regressors which are (multiply) uncorrelated with the regressors of primary interest can, without any harm, be included in the model. In general, it is necessary to find a trade-off between too poor and too rich model.

Interest in prediction

If prediction is the primary purpose of the regression analysis, results derived in Sections 12.2.3 and 12.2.4 dictate to follow a strategy to include all available covariates in the model. The reasons are the following.

(i) If we omit important regressors, the predictions get biased and the averaged mean squared error of prediction is possibly not tending to the optimal value of σ^2 with $n \to \infty$.

(ii) If we include irrelevant regressors in the model, this has, especially with $n \to \infty$, a negligible effect on a quality of the prediction. The averaged mean squared error of prediction is still tending to the optimal value of σ^2.

Chapter 13

Asymptotic Properties of the LSE and Sandwich Estimator

13.1 Assumptions and setup

Assumption (A0).

(i) Let $\big(Y_1,\, \boldsymbol{X}_1^\top\big)^\top,\, \big(Y_2,\, \boldsymbol{X}_2^\top\big)^\top, \ldots$ be a *sequence* of $(1+k)$-dimensional *independent and identically distributed (i.i.d.)* random vectors being distributed as a generic random vector $\big(Y,\, \boldsymbol{X}^\top\big)^\top$, ($\boldsymbol{X} = \big(X_0,\, X_1,\, \ldots,\, X_{k-1}\big)^\top$, $\boldsymbol{X}_i = \big(X_{i,0},\, X_{i,1},\, \ldots,\, X_{i,k-1}\big)^\top$, $i = 1,\, 2,\, \ldots$);

(ii) Let $\boldsymbol{\beta} = \big(\beta_0,\, \ldots,\, \beta_{k-1}\big)^\top$ be an unknown k-dimensional real parameter;

(iii) Let $\mathbb{E}\big(Y \,\big|\, \boldsymbol{X}\big) = \boldsymbol{X}^\top\boldsymbol{\beta}$.

***Notation** (Error terms).*

We denote $\varepsilon = Y - \boldsymbol{X}^\top\boldsymbol{\beta}$,

$$\varepsilon_i = Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}, \quad i = 1, 2, \ldots.$$

Notes.

- In this chapter, all unconditional expectations must be understood as expectations with respect to the joint distribution of a random vector $\big(Y,\, \boldsymbol{X}^\top\big)^\top$ (which depends on the vector $\boldsymbol{\beta}$).
- From assumption (A0), the error terms $\varepsilon_1,\, \varepsilon_2,\, \ldots$ are *i.i.d.* with a distribution of a generic error term ε. The following can be concluded for their first two (conditional) moments:

$$\begin{aligned}
\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big) &= \mathbb{E}\big(Y - \boldsymbol{X}^\top\boldsymbol{\beta} \,\big|\, \boldsymbol{X}\big) &&= 0,\\
\mathsf{var}\big(\varepsilon \,\big|\, \boldsymbol{X}\big) &= \mathsf{var}\big(Y - \boldsymbol{X}^\top\boldsymbol{\beta} \,\big|\, \boldsymbol{X}\big) = \mathsf{var}\big(Y \,\big|\, \boldsymbol{X}\big) &&=: \sigma^2(\boldsymbol{X}),
\end{aligned}$$

$$\mathbb{E}(\varepsilon) = \mathbb{E}\Big(\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) = \mathbb{E}(0) \qquad\qquad = 0,$$

$$\mathsf{var}(\varepsilon) = \mathsf{var}\Big(\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) + \mathbb{E}\Big(\mathsf{var}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) = \mathsf{var}\big(0\big) + \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\big\} = \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\big\}.$$

Assumption (A1).

Let the covariate random vector $\boldsymbol{X} = \big(X_0, \ldots, X_{k-1}\big)^\top$ satisfy

(i) $\mathbb{E}\big|X_j\, X_l\big| < \infty, \quad j, l = 0, \ldots, k-1;$

(ii) $\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big) = \mathbb{W}$, where $\mathbb{W}$ is a positive definite matrix.

Notation *(Covariates second and first mixed moments).*

Let $\mathbb{W} = \big(w_{j,l}\big)_{j,l=0,\ldots,k-1}$. We have,

$$w_j^2 := w_{j,j} = \mathbb{E}\big(X_j^2\big), \qquad j = 0, \ldots, k-1,$$
$$w_{j,l} = \mathbb{E}\big(X_j\, X_l\big), \qquad j \neq l.$$

Let

$$\mathbb{V} := \mathbb{W}^{-1} = \big(v_{j,l}\big)_{j,l=0,\ldots,k-1}.$$

Notation *(Data of size n).*

For $n \geq 1$:

$$\boldsymbol{Y}_n := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{X}_n := \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}, \qquad \mathbb{W}_n := \mathbb{X}_n^\top \mathbb{X}_n = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

$$\mathbb{V}_n := \big(\mathbb{X}_n^\top \mathbb{X}_n\big)^{-1} \text{ (if it exists).}$$

Lemma 13.1 Consistent estimator of the second and first mixed moments of the covariates.

Let assumpions (A0) and (A1) hold. Then

$$\frac{1}{n}\,\mathbb{W}_n \xrightarrow{a.s.} \mathbb{W} \qquad \textit{as } n \to \infty,$$

$$n\,\mathbb{V}_n \xrightarrow{a.s.} \mathbb{V} \qquad \textit{as } n \to \infty.$$

Proof. The statement of Lemma follows from applying, for each $j = 0, \ldots, k-1$ and $l = 0, \ldots, k-1$, the strong law of large numbers for i.i.d. random variables (Theorem C.2) to a sequence

$$Z_{i,j,l} = X_{i,j}\, X_{i,l}, \qquad i = 1, 2, \ldots.$$

❑

LSE based on data of size n

Since $\frac{1}{n}\,\mathbb{X}_n^\top \mathbb{X}_n \xrightarrow{\text{a.s.}} \mathbb{W} > 0$ then

$$\mathsf{P}\big(\text{there exists } n_0 > k \text{ such that for all } n \geq n_0 \quad \mathsf{rank}\big(\mathbb{X}_n\big) = k\big) = 1$$

and we define (for $n \geq n_0$)

$$\widehat{\boldsymbol{\beta}}_n \;=\; \big(\mathbb{X}_n^\top \mathbb{X}_n\big)^{-1} \mathbb{X}_n^\top \boldsymbol{Y}_n = \Big(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top\Big)^{-1} \Big(\sum_{i=1}^n \boldsymbol{X}_i Y_i\Big),$$

$$\mathsf{MS}_{e,n} \;=\; \frac{1}{n-k} \big\|\boldsymbol{Y}_n - \mathbb{X}_n \widehat{\boldsymbol{\beta}}_n\big\|^2 = \frac{1}{n-k} \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n)^2,$$

which are the LSE of $\boldsymbol{\beta}$ and the residual mean square based on the assumed linear model for data of size n.

$$\mathsf{M}_n : \; \boldsymbol{Y}_n \,\big|\, \mathbb{X}_n \sim \big(\mathbb{X}_n \boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big).$$

Further, for $n \geq n_0$ any non-trivial linear combination of regression coefficients is estimable parameter of model M_n.

- For a given real vector $\mathbf{l} = \big(l_0,\, l_1,\, \ldots,\, l_{k-1}\big)^\top \neq \mathbf{0}_k$ we denote

$$\theta = \mathbf{l}^\top \boldsymbol{\beta}, \qquad \widehat{\theta}_n = \mathbf{l}^\top \widehat{\boldsymbol{\beta}}_n.$$

- For a given $m \times k$ matrix $\mathbb{L}$ with rows $\mathbf{l}_1^\top \neq \mathbf{0}_k^\top, \ldots, \mathbf{l}_m^\top \neq \mathbf{0}_k^\top$ we denote

$$\boldsymbol{\xi} = \mathbb{L}\boldsymbol{\beta}, \qquad \widehat{\boldsymbol{\xi}}_n = \mathbb{L}\widehat{\boldsymbol{\beta}}_n.$$

It will be assumed that $m \leq k$ and that the rows of $\mathbb{L}$ are lineary independent.

Interest will be in asymptotic (as $n \to \infty$) behavior of

(i) $\widehat{\boldsymbol{\beta}}_n$;

(ii) $\mathsf{MS}_{e,n}$;

(iii) $\widehat{\theta}_n = \mathbf{l}^\top \widehat{\boldsymbol{\beta}}_n$ for given $\mathbf{l} \neq \mathbf{0}_k$;

(iv) $\widehat{\boldsymbol{\xi}}_n = \mathbb{L}\widehat{\boldsymbol{\beta}}_n$ for given $m \times k$ matrix $\mathbb{L}$ with linearly independent rows;

under two different scenarios (two different truths)

(i) *homoscedastic* errors (i.e., model $\mathsf{M}_n : \; \boldsymbol{Y}_n \,\big|\, \mathbb{X}_n \sim \big(\mathbb{X}_n \boldsymbol{\beta},\, \sigma^2\, \mathbf{I}_n\big)$ is correct);

(ii) *heteroscedastic* errors where $\mathsf{var}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)$ is not necessarily constant and perhaps depends on the covariate values $\boldsymbol{X}$ (i.e., model M_n is not necessarily fully correct).

Normality of the errors will not be assumed.

Assumption (A2 homoscedastic).

Let the conditional variance of the response satisfy

$$\sigma^2(\boldsymbol{X}) := \mathsf{var}\big(Y \,\big|\, \boldsymbol{X}\big) = \sigma^2,$$

where $\infty > \sigma^2 > 0$ is an unknown parameter.

Assumption (A2 heteroscedastic).

Let $\sigma^2(\boldsymbol{X}) := \mathsf{var}\big(Y \,\big|\, \boldsymbol{X}\big)$ satisfy, for each $j, l = 0, \ldots, k-1$, the condition

$$\mathbb{E}\big\{\sigma^2(\boldsymbol{X}) X_j \, X_l\big\} < \infty.$$

Notes.

- Condition (A2 heteroscedastic) states that the matrix

  $$\mathbb{W}^{\bigstar} := \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\, \boldsymbol{X}\boldsymbol{X}^{\top}\big\}$$

  is a real matrix (with all elements being finite).

- If (A0) and (Al) are assumed then

  $$\text{(A2 homoscedastic)} \Longrightarrow \text{(A2 heteroscedastic)}.$$

  Hence everything that will be proved under (A2 heteroscedastic) holds also under (A2 homoscedastic).

- Under assumptions (A0) and (A2 homoscedastic), we have

  $$\mathbb{E}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \boldsymbol{X}_i^{\top}\boldsymbol{\beta}, \qquad \mathsf{var}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big) = \sigma^2, \qquad i = 1, 2, \ldots,$$

  and for each $n > 1$, $Y_1, \ldots, Y_n$ are, given $\mathbb{X}_n$, independent and satisfying a linear model

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

- Under assumptions (A0) and (A2 heteroscedastic), we have

  $$\mathbb{E}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \boldsymbol{X}_i^{\top}\boldsymbol{\beta}, \qquad \mathsf{var}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big) = \sigma^2(\boldsymbol{X}_i), \qquad i = 1, 2, \ldots,$$

  and for each $n > 1$, $Y_1, \ldots, Y_n$ are, given $\mathbb{X}_n$, independent with

  $$\mathbb{E}\big(\boldsymbol{Y}_n \,\big|\, \mathbb{X}_n\big) = \mathbb{X}_n\boldsymbol{\beta}, \qquad \mathsf{var}\big(\boldsymbol{Y}_n \,\big|\, \mathbb{X}_n\big) = \begin{pmatrix} \sigma^2(\boldsymbol{X}_1) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma^2(\boldsymbol{X}_n) \end{pmatrix}.$$

13.2 Consistency of LSE

We shall show in this section:

(i) Strong consistency of $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\theta}_n$, $\widehat{\boldsymbol{\xi}}_n$ (LSE's regression coefficients or their linear combinations).
- No need of normality;
- No need of homoscedasticity.

(ii) Strong consistency of $\mathsf{MS}_{e,n}$ (unbiased estinator of the residual variance).
- No need of normality.

Theorem 13.2 Strong consistency of LSE.

Let assumptions (A0), (A1) and (A2 heteroscedastic*) hold.*
Then

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n &\xrightarrow{a.s.} \boldsymbol{\beta} && \text{as } n \to \infty, \\ \mathbf{l}^\top \widehat{\boldsymbol{\beta}}_n = \widehat{\theta}_n &\xrightarrow{a.s.} \theta \quad = \mathbf{l}^\top \boldsymbol{\beta} && \text{as } n \to \infty, \\ \mathbb{L}\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\xi}}_n &\xrightarrow{a.s.} \boldsymbol{\xi} \quad = \mathbb{L}\boldsymbol{\beta} && \text{as } n \to \infty. \end{aligned}$$

Proof.

It is sufficient to show that $\widehat{\boldsymbol{\beta}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\beta}$. The remaining two statements follow from properties of convergence almost surely.

We have

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n &= \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \left(\mathbb{X}_n^\top \boldsymbol{Y}_n\right) \\ &= \underbrace{\left(\frac{1}{n}\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1}}_{\boldsymbol{A}_n} \underbrace{\left(\frac{1}{n}\mathbb{X}_n^\top \boldsymbol{Y}_n\right)}_{\boldsymbol{B}_n}, \end{aligned}$$

where $\boldsymbol{A}_n = \left(\frac{1}{n}\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \xrightarrow{\text{a.s.}} \mathbb{W}^{-1}$ by Lemma 13.1.

Further

$$\begin{aligned} \boldsymbol{B}_n = \frac{1}{n}\mathbb{X}_n^\top \boldsymbol{Y}_n &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} + \boldsymbol{X}_i^\top \boldsymbol{\beta}\right) \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \varepsilon_i}_{\boldsymbol{C}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\beta}}_{\boldsymbol{D}_n}. \end{aligned}$$

(a) $\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \varepsilon_i \xrightarrow{\text{a.s.}} \mathbf{0}_k$ due to the SLLN (i.i.d., Theorem C.2). This is justified as follows.

- The jth ($j = 0, \ldots, k-1$) element of the vector $\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \varepsilon_i$ is $\frac{1}{n}\sum_{i=1}^n X_{i,j}\varepsilon_i$.

- The random variables $X_{i,j}\varepsilon_i$, $i = 1, 2, \ldots$ are i.i.d. by (A0).
- $$\begin{aligned}\mathbb{E}\big(X_{i,j}\varepsilon_i\big) &= \mathbb{E}\Big(\mathbb{E}\big(X_{i,j}\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big)\\ &= \mathbb{E}\Big(X_{i,j}\mathbb{E}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big)\\ &= \mathbb{E}\Big(X_{i,j}\,0\Big)\\ &= 0.\end{aligned}$$
- $$\begin{aligned}\mathsf{var}\big(X_{i,j}\varepsilon_i\big) &= \mathbb{E}\Big(\mathsf{var}\big(X_{i,j}\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big) \;+\; \mathsf{var}\Big(\mathbb{E}\big(X_{i,j}\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big)\\ &= \mathbb{E}\Big(X_{i,j}^2\,\mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big) \;+\; \mathsf{var}\Big(X_{i,j}\,0\Big)\\ &= \mathbb{E}\big(X_{i,j}^2\,\sigma^2(\boldsymbol{X}_i)\big)\\ &< \infty \text{ by (A2 heteroscedastic)}\\ &\Longrightarrow \mathbb{E}\big|X_{i,j}\varepsilon_i\big| < \infty.\end{aligned}$$

(b) $\boldsymbol{D}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\beta} = \dfrac{1}{n}\mathbb{W}_n\boldsymbol{\beta} \xrightarrow{\text{a.s.}} \mathbb{W}\boldsymbol{\beta}$ by Lemma 13.1.

In summary: $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{A}_n\big(\boldsymbol{C}_n + \boldsymbol{D}_n\big)$, where $\boldsymbol{A}_n \xrightarrow{\text{a.s.}} \mathbb{W}^{-1}$,
$\boldsymbol{C}_n \xrightarrow{\text{a.s.}} \boldsymbol{0}_k$,
$\boldsymbol{D}_n \xrightarrow{\text{a.s.}} \mathbb{W}\boldsymbol{\beta}$.

Hence

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{\text{a.s.}} \mathbb{W}^{-1}\,\mathbb{W}\,\boldsymbol{\beta} = \boldsymbol{\beta}.$$

❑

Theorem 13.3 Strong consistency of the mean squared error.
Let assumptions (A0), (A1), (A2 homoscedastic*) hold.*

Then

$$\mathsf{MS}_{e,n} \xrightarrow{a.s.} \sigma^2 \quad \textit{as } n \to \infty.$$

Proof.

We have

$$\mathsf{MS}_{e,n} = \frac{1}{n-k}\,\mathsf{SS}_{e,n} = \frac{n}{n-k}\,\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}\big)^2.$$

Since $\lim_{n\to\infty} \frac{n}{n-k} = 1$, it is sufficient to show that

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n\big)^2 \xrightarrow{\text{a.s.}} \sigma^2 \qquad \text{as } n \to \infty.$$

We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n\big)^2 &= \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} + \boldsymbol{X}_i^\top \boldsymbol{\beta} - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n\big)^2 \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\big)^2}_{\boldsymbol{A}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\big\{\boldsymbol{X}_i^\top\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)\big\}^2}_{\boldsymbol{B}_n} + \underbrace{\frac{2}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\big)\,\boldsymbol{X}_i^\top\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)}_{\boldsymbol{C}_n}.\end{aligned}$$

(a) $\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\big)^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \xrightarrow{\text{a.s.}} \sigma^2$ due to the SLLN (i.i.d., Theorem C.2).
This is justified by noting the following.

- The random variables ε_i^2, $i = 1, 2, \ldots$ are i.i.d. by (A0).
- $\mathbb{E}\big(\varepsilon_i\big) = 0$

  $\Longrightarrow \mathbb{E}\big(\varepsilon_i^2\big) = \mathsf{var}\big(\varepsilon_i\big) = \mathbb{E}\big\{\sigma^2(\boldsymbol{X}_i)\big\} = \mathbb{E}\big(\sigma^2\big) = \sigma^2$ by assumption (A2 homoscedastic).
- $\mathbb{E}\big|\varepsilon_i^2\big| = \mathbb{E}\big(\varepsilon_i^2\big) = \sigma^2 < \infty$ by assumption (A2 homoscedastic).

(b) $\boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n}\big\{\boldsymbol{X}_i^\top\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)\big\}^2 \xrightarrow{\text{a.s.}} 0$, which is seen as follows.

$$\begin{aligned}\boldsymbol{B}_n &= \frac{1}{n}\sum_{i=1}^{n}\big\{\boldsymbol{X}_i^\top\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)\big\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)^\top \boldsymbol{X}_i\boldsymbol{X}_i^\top\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big) \\ &= \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)^\top\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top\Big)\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big) \\ &= \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)^\top\Big(\frac{1}{n}\mathbb{X}_n^\top\mathbb{X}_n\Big)\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big).\end{aligned}$$

Now $\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ due to Theorem 13.2.

$\frac{1}{n}\mathbb{X}_n^\top\mathbb{X}_n \xrightarrow{\text{a.s.}} \mathbb{W}$ due to Lemma 13.1.

Hence

$$\boldsymbol{B}_n \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top\,\mathbb{W}\,\mathbf{0}_k = 0.$$

(c) $\boldsymbol{C}_n = \frac{2}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\boldsymbol{X}_i^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\right) \xrightarrow{\text{a.s.}} 0$, which is justified by the following.

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

Now $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i^\top \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top$ as was shown in the proof of Theorem 13.2.
$\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\right) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ due to Theorem 13.2.

Hence

$$\boldsymbol{C}_n \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top\,\mathbf{0}_k = 0.$$

In summary: $\mathsf{MS}_{e,n} = \frac{n}{n-k}\left(\boldsymbol{A}_n + \boldsymbol{B}_n + \boldsymbol{C}_n\right)$, where $\frac{n}{n-k} \to 1$,
$\boldsymbol{A}_n \xrightarrow{\text{a.s.}} \sigma^2$,
$\boldsymbol{B}_n \xrightarrow{\text{a.s.}} 0$,
$\boldsymbol{C}_n \xrightarrow{\text{a.s.}} 0$.

Hence

$$\mathsf{MS}_{e,n} \xrightarrow{\text{a.s.}} 1\left(\sigma^2 + 0 + 0\right) = \sigma^2.$$

❑

13.3 Asymptotic normality of LSE under homoscedasticity

We shall show in this section: asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\theta}_n$, $\widehat{\boldsymbol{\xi}}_n$ (LSE's regression coefficients or their linear combinations) when homoscedasticity of the errors is assumed but not their normality.

Reminder. $\mathbb{V} = \left\{\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\right\}^{-1}$.

Theorem 13.4 Asymptotic normality of LSE in homoscedastic case.

Let assumptions (A0), (A1), (A2 homoscedastic*) hold.*
Then

$$\begin{aligned}
\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) &\xrightarrow{\mathcal{D}} \mathcal{N}_k(\mathbf{0}_k,\ \sigma^2\,\mathbb{V}) && \text{as } n \to \infty,\\
\sqrt{n}\big(\widehat{\theta}_n - \theta\big) &\xrightarrow{\mathcal{D}} \mathcal{N}_1(0,\ \sigma^2\,\mathbf{l}^\top\mathbb{V}\,\mathbf{l}) && \text{as } n \to \infty,\\
\sqrt{n}\big(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\big) &\xrightarrow{\mathcal{D}} \mathcal{N}_m(\mathbf{0}_m,\ \sigma^2\,\mathbb{L}\,\mathbb{V}\,\mathbb{L}^\top) && \text{as } n \to \infty.
\end{aligned}$$

Proof. Will be provided jointly with Theorem 13.5. ❑

13.3.1 Asymptotic validity of the classical inference under homoscedasticity but non-normality

For given $n \geq n_0 > k$, the following statistics are used to infer on estimable parameters of the linear model M_n based on the response vector $\boldsymbol{Y}_n$ and the model matrix $\mathbb{X}_n$ (see Chapter 3):

$$T_n := \frac{\widehat{\theta}_n - \theta}{\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}}}, \tag{13.1}$$

$$Q_n := \frac{1}{m}\,\frac{\big(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\big)^\top\left\{\mathbb{L}\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbb{L}^\top\right\}^{-1}\big(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\big)}{\mathsf{MS}_{e,n}}. \tag{13.2}$$

Reminder.

- $\mathbb{V}_n = \big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}$.
- Under assumptions (A0) and (A1): $n\,\mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$ as $n \to \infty$.

Consequence of Theorem 13.4: Asymptotic distribution of t- and F-statistics.

Under assumptions of Theorem 13.4:

$$T_n \xrightarrow{\mathcal{D}} \mathcal{N}_1(0,\,1) \qquad \text{as } n \to \infty,$$
$$m\,Q_n \xrightarrow{\mathcal{D}} \chi^2_m \qquad \text{as } n \to \infty.$$

Proof. It follows directly from Lemma 13.1, Theorem 13.4 and Cramér-Slutsky theorem (Theorem C.7) as follows.

$$T_n = \frac{\mathbf{l}^\top\widehat{\boldsymbol{\beta}}_n - \mathbf{l}^\top\boldsymbol{\beta}}{\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\left(\mathbb{X}_n^\top\mathbb{X}_n\right)^{-1}\mathbf{l}}} = \underbrace{\frac{\sqrt{n}(\mathbf{l}^\top\widehat{\boldsymbol{\beta}}_n - \mathbf{l}^\top\boldsymbol{\beta})}{\sqrt{\sigma^2\,\mathbf{l}^\top\mathbb{V}\mathbf{l}}}}_{\xrightarrow{\mathcal{D}}\ \mathcal{N}(0,\,1)} \underbrace{\sqrt{\frac{\sigma^2\,\mathbf{l}^\top\mathbb{V}\mathbf{l}}{\mathsf{MS}_{e,n}\mathbf{l}^\top\left\{n\left(\mathbb{X}_n^\top\mathbb{X}_n\right)^{-1}\right\}\mathbf{l}}}}_{\xrightarrow{\mathsf{P}}\ 1}.$$

$$\begin{aligned} m\,Q_m &= \left(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta}\right)^\top\left\{\mathsf{MS}_{e,n}\mathbb{L}\left(\mathbb{X}_n^\top\mathbb{X}_n\right)^{-1}\mathbb{L}^\top\right\}^{-1}\left(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta}\right) \\ &= \underbrace{\sqrt{n}\left(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta}\right)^\top}_{\xrightarrow{\mathcal{D}}\ \mathcal{N}_m\left(\mathbf{0}_m,\ \sigma^2\,\mathbb{L}\mathbb{V}\mathbb{L}^\top\right)} \underbrace{\left\{\mathsf{MS}_{e,n}\mathbb{L}\,n\left(\mathbb{X}_n^\top\mathbb{X}_n\right)^{-1}\mathbb{L}^\top\right\}^{-1}}_{\xrightarrow{\mathsf{P}}\ \sigma^2\mathbb{L}\mathbb{V}\mathbb{L}^\top} \underbrace{\left(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta}\right)\sqrt{n}}_{\xrightarrow{\mathcal{D}}\ \mathcal{N}_m\left(\mathbf{0}_m,\ \sigma^2\,\mathbb{L}\mathbb{V}\mathbb{L}^\top\right)}. \end{aligned}$$

Convergence to χ^2_m in distribution follows from a property of (multivariate) normal distribution concerning the distribution of a quadratic form.

❑

If additionaly normality is assumed, i.e., if it is assumed $\boldsymbol{Y}_n \,|\, \mathbb{X}_n \sim \mathcal{N}_n(\mathbb{X}_n\boldsymbol{\beta},\,\sigma^2\mathbf{I}_n)$ then Theorem 3.2 (LSE under the normality) provides

$$T_n \sim \mathsf{t}_{n-k},$$
$$Q_n \sim \mathcal{F}_{m,\,n-k}.$$

This is then used for inference (derivation of confidence intervals and regions, construction of tests) on the estimable parameters of a linear model under assumption of normality.

The following holds in general:

$$\begin{aligned} &T_\nu \sim \mathsf{t}_\nu && \text{then} && T_\nu \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,1) && \text{as } \nu \to \infty, \\ &Q_\nu \sim \mathcal{F}_{m,\nu} && \text{then} && m\,Q_\nu \xrightarrow{\mathcal{D}} \chi^2_m && \text{as } \nu \to \infty. \end{aligned} \tag{13.3}$$

This, together with Consequence of Theorem 13.4 then justify asymptotic validity of a classical inference based on statistics T_n (Eq. 13.1) and Q_n (Eq. 13.2), respectively and a Student t and F-distribution, respectively, even if normality of the error terms of the linear model does not hold. The only requirements are assumptions of Theorem 13.4.

That is, for example, both intervals

(i) $\mathcal{I}_n^{\mathcal{N}} := \Big(\widehat{\theta}_n - u(1-\alpha/2)\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}},\quad \widehat{\theta}_n + u(1-\alpha/2)\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}}\Big)$;

(ii) $\mathcal{I}_n^{\mathsf{t}} := \Big(\widehat{\theta}_n - \mathsf{t}_{n-k}(1-\alpha/2)\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}},\quad \widehat{\theta}_n + \mathsf{t}_{n-k}(1-\alpha/2)\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}}\Big)$,

satisfy, for any $\theta^0 \in \mathbb{R}$ (even without normality of the error terms)

$$\mathsf{P}\big(\mathcal{I}_n^{\mathcal{N}} \ni \theta^0;\ \theta = \theta^0\big) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\big(\mathcal{I}_n^{\mathsf{t}} \ni \theta^0;\ \theta = \theta^0\big) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

Analogously, due to a general asymptotic property of the F-distribution (Eq. 13.3), asymptotically valid inference on the estimable vector parameter $\boldsymbol{\xi} = \mathbb{L}\boldsymbol{\beta}$ of a linear model can be based either on the statistic $m\,Q_n$ and the χ^2_m distribution or on the statistic Q_n and the $\mathcal{F}_{m,\,n-k}$ distribution. For example, for both ellipsoids

(i) $\mathcal{K}_n^{\chi} := \Big\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \big(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\big)^\top \Big\{\mathsf{MS}_{e,n}\mathbb{L}\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\big) \,<\, \chi^2_m(1-\alpha)\Big\}$;

(ii) $\mathcal{K}_n^{\mathcal{F}} := \Big\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \big(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\big)^\top \Big\{\mathsf{MS}_{e,n}\mathbb{L}\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\big) \,<\, m\,\mathcal{F}_{m,n-k}(1-\alpha)\Big\}$,

we have for any $\boldsymbol{\xi}^0 \in \mathbb{R}^m$ (under assumptions of Theorems 13.4):

$$\mathsf{P}\big(\mathcal{K}_n^{\chi} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\big) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\big(\mathcal{K}_n^{\mathcal{F}} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\big) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

13.4 Asymptotic normality of LSE under heteroscedasticity

We shall show in this section: asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\theta}_n$, $\widehat{\boldsymbol{\xi}}_n$ (LSE's regression coefficients or their linear combinations) when even homoscedasticity of the errors is not assumed.

Reminder.

- $\mathbb{V} = \left\{\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\right\}^{-1}$.
- $\mathbb{W}^\star = \mathbb{E}\{\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\}$.

Theorem 13.5 Asymptotic normality of LSE in heteroscedastic case.

Let assumptions (A0), (A1), (A2 heteroscedastic*) hold.*
Then

$$\begin{aligned}
\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) &\xrightarrow{\mathcal{D}} \mathcal{N}_k(\mathbf{0}_k,\ \mathbb{V}\mathbb{W}^\star\mathbb{V}) && \text{as } n\to\infty,\\
\sqrt{n}\big(\widehat{\theta}_n - \theta\big) &\xrightarrow{\mathcal{D}} \mathcal{N}_1(0,\ \mathbf{l}^\top\mathbb{V}\mathbb{W}^\star\mathbb{V}\,\mathbf{l}) && \text{as } n\to\infty,\\
\sqrt{n}\big(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\big) &\xrightarrow{\mathcal{D}} \mathcal{N}_m(\mathbf{0}_m,\ \mathbb{L}\,\mathbb{V}\mathbb{W}^\star\mathbb{V}\,\mathbb{L}^\top) && \text{as } n\to\infty.
\end{aligned}$$

Proof. We will jointly prove also Theorem 13.4.

We have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_n &= \underbrace{\left(\mathbb{X}_n^\top\mathbb{X}_n\right)^{-1}}_{\mathbb{V}_n}\mathbb{X}_n^\top\boldsymbol{Y}_n\\
&= \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i Y_i\\
&= \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\left(\boldsymbol{X}_i^\top\boldsymbol{\beta}+\varepsilon_i\right)\\
&= \mathbb{V}_n\underbrace{\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\right)}_{\mathbb{V}_n^{-1}}\boldsymbol{\beta} + \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\\
&= \boldsymbol{\beta} + \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i.
\end{aligned}$$

That is,

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i = n\,\mathbb{V}_n\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i. \tag{13.4}$$

By Lemma 13.1, $n\,\mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$ which implies

$$n\,\mathbb{V}_n \xrightarrow{\mathsf{P}} \mathbb{V} \qquad \text{as } n\to\infty. \tag{13.5}$$

In the following, let us explore asymptotic behavior of the term $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i$.

From assumption (A0), the term $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i$ is a sample mean of i.i.d. random vector $\boldsymbol{X}_i\,\varepsilon_i$, $i=1,\ldots,n$. The mean and the covariance matrix of the distribution of those random vectors are

$$\begin{aligned}
\mathbb{E}\big(\boldsymbol{X}\varepsilon\big) &= \mathbf{0}_k \qquad \text{(was shown in the proof of Theorem 13.2)},\\
\mathsf{var}\big(\boldsymbol{X}\varepsilon\big) &= \mathbb{E}\Big(\mathsf{var}\big(\boldsymbol{X}\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) \;+\; \mathsf{var}\Big(\mathbb{E}\big(\boldsymbol{X}\varepsilon \,\big|\, \boldsymbol{X}\big)\Big)\\
&= \mathbb{E}\Big(\boldsymbol{X}\,\underbrace{\mathsf{var}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)}_{\sigma^2(\boldsymbol{X})}\,\boldsymbol{X}^\top\Big) \;+\; \mathsf{var}\big(\boldsymbol{X}\,\underbrace{\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)}_{0}\big)\\
&= \mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big).
\end{aligned}$$

Depending, on whether (A2 homoscedastic) or (A2 heteroscedastic) is assumed, we have

$$\mathsf{var}\big(\boldsymbol{X}\varepsilon\big) \;=\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big) \;=\; \begin{cases} \sigma^2\,\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big) = \sigma^2\mathbb{W}, & \text{(A2 homoscedastic)},\\ \mathbb{W}^{\bigstar}, & \text{(A2 heteroscedastic)}. \end{cases} \tag{13.6}$$

Under both (A2 homoscedastic) and (A2 heteroscedastic) all elements of the covariance matrix $\mathsf{var}\big(\boldsymbol{X}\varepsilon\big)$ are finite. Hence by Theorem C.5 (multivariate CLT for i.i.d. random vectors):

$$\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i \;\xrightarrow{\mathcal{D}}\; \mathcal{N}_k\Big(\mathbf{0}_k,\, \mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big)\Big) \qquad \text{as } n\to\infty.$$

From (13.4) and (13.5), we now have,

$$\big(\widehat{\boldsymbol{\beta}}_n \;-\; \boldsymbol{\beta}\big) \;=\; \underbrace{n\,\mathbb{V}_n}_{\xrightarrow{\mathsf{P}}\mathbb{V}} \qquad \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i}_{\xrightarrow{\mathcal{D}}\mathcal{N}_k\left(\mathbf{0}_k,\,\mathbb{E}\left(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\right)\right)} \qquad \frac{1}{\sqrt{n}}.$$

That is,

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n \;-\; \boldsymbol{\beta}\big) \;=\; \underbrace{n\,\mathbb{V}_n}_{\xrightarrow{\mathsf{P}}\mathbb{V}} \qquad \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i}_{\xrightarrow{\mathcal{D}}\mathcal{N}_k\left(\mathbf{0}_k,\,\mathbb{E}\left(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\right)\right)} \qquad .$$

Finally, by applying Theorem C.7 (Cramér–Slutsky):

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n \;-\; \boldsymbol{\beta}\big) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}_k\Big(\mathbf{0}_k,\, \mathbb{V}\,\mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big)\,\mathbb{V}^\top\Big) \qquad \text{as } n\to\infty.$$

By using (13.6) and realizing that $\mathbb{V}^\top = \mathbb{V}$, we get

Under (A2 homoscedastic)

$$\mathbb{V}\,\mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big)\,\mathbb{V}^\top = \mathbb{V}\,\sigma^2\,\mathbb{W}\,\mathbb{V} = \sigma^2\,\mathbb{V}\,\mathbb{V}^{-1}\,\mathbb{V} = \sigma^2\,\mathbb{V}$$

and hence

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n \;-\; \boldsymbol{\beta}\big) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}_k\big(\mathbf{0}_k,\, \sigma^2\,\mathbb{V}\big) \qquad \text{as } n\to\infty.$$

Under (A2 heteroscedastic)

$$\mathbb{V}\,\mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big)\,\mathbb{V}^\top = \mathbb{V}\,\mathbb{W}^\bigstar\,\mathbb{V}$$

and hence

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n \;-\; \boldsymbol{\beta}\big) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}_k\Big(\mathbf{0}_k,\, \mathbb{V}\,\mathbb{W}^\bigstar\,\mathbb{V}\Big) \qquad \text{as } n \to \infty.$$

Asymptotic normality of $\widehat{\theta}_n = \mathbf{l}^\top\widehat{\boldsymbol{\beta}}_n$ and of $\boldsymbol{\xi}_n = \mathbb{L}\widehat{\boldsymbol{\beta}}_n$ follows now from Theorem C.6 (Cramér–Wold).

❑

Notation *(Residuals and related quantities based on a model for data of size n).*

For $n \geq n_0 > k$, the following notation will be used for quantities based on the model

$$\mathsf{M}_n\colon\; \boldsymbol{Y}_n \,\big|\, \mathbb{X}_n \;\sim\; \big(\mathbb{X}_n\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big).$$

- Hat matrix: $\mathbb{H}_n = \mathbb{X}_n\,\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\,\mathbb{X}_n^\top$;
- Residual projection matrix: $\mathbb{M}_n = \mathbf{I}_n - \mathbb{H}_n$;
- Diagonal elements of matrix $\mathbb{H}_n$: $h_{n,1},\, \ldots,\, h_{n,n}$;
- Diagonal elements of matrix $\mathbb{M}_n$: $m_{n,1} = 1 - h_{n,1},\, \ldots,\, m_{n,n} = 1 - h_{n,n}$;
- Residuals: $\boldsymbol{U}_n = \mathbb{M}_n\boldsymbol{Y}_n = \big(U_{n,1},\, \ldots,\, U_{n,n}\big)^\top$.

Reminder.

- $\mathbb{V}_n = \Big(\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i{}^\top\Big)^{-1} = \big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}$.
- Under assumptions (A0) and (A1): $n\,\mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$ as $n \to \infty$.

Theorem 13.6 Sandwich estimator of the covariance matrix.

Let assumptions (A0), (A1), (A2 heteroscedastic*) hold. Let additionally, for each $s,\, t,\, j,\, l = 0, \ldots, k-1$*

$$\mathbb{E}\big|\varepsilon^2\,X_j\,X_l\big| < \infty, \qquad \mathbb{E}\big|\varepsilon\,X_s\,X_j\,X_l\big| < \infty, \qquad \mathbb{E}\big|X_s\,X_t\,X_j\,X_l\big| < \infty.$$

Then

$$n\,\mathbb{V}_n\,\mathbb{W}_n^\bigstar\,\mathbb{V}_n \;\xrightarrow{\text{a.s.}}\; \mathbb{V}\,\mathbb{W}^\bigstar\,\mathbb{V} \qquad \textit{as } n \to \infty,$$

where for $n = 1,\, 2,\, \ldots$,

$$\mathbb{W}_n^\bigstar = \sum_{i=1}^{n} U_{n,i}^2\,\boldsymbol{X}_i\boldsymbol{X}_i^\top = \mathbb{X}_n^\top\boldsymbol{\Omega}_n\mathbb{X}_n,$$

$$\boldsymbol{\Omega}_n = \mathsf{diag}\big(\omega_{n,1},\, \ldots,\, \omega_{n,n}\big), \qquad \omega_{n,i} = U_{n,i}^2, \quad i = 1, \ldots, n.$$

Proof.

First, remind that

$$\mathbb{V}\,\mathbb{W}^{\bigstar}\,\mathbb{V} \;=\; \Big\{\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\Big\}^{-1}\,\mathbb{E}\big(\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\big)\,\Big\{\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\Big\}^{-1},$$

and we know from Lemma 13.1 that

$$n\,\mathbb{V}_n \;=\; n\,\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1} \;\xrightarrow{\text{a.s.}}\; \Big\{\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\Big\}^{-1} = \mathbb{V} \qquad \text{as } n\to\infty.$$

Hence, if we show that

$$\frac{1}{n}\,\mathbb{W}_n^{\bigstar} \;=\; \frac{1}{n}\sum_{i=1}^{n} U_{n,i}^2\,\boldsymbol{X}_i\,\boldsymbol{X}_i^\top \;\xrightarrow{\text{a.s.}}\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\big) = \mathbb{W}^{\bigstar} \qquad \text{as } n\to\infty,$$

the statement of Theorem will be proven.

Remember,

$$\sigma^2(\boldsymbol{X}) \;=\; \mathsf{var}\big(\varepsilon\,\big|\,\boldsymbol{X}\big) \;=\; \mathbb{E}\big(\varepsilon^2\,\big|\,\boldsymbol{X}\big).$$

From here, for each $j,\, l = 0,\ldots,k-1$

$$\begin{aligned}\mathbb{E}\big(\varepsilon^2\,X_j\,X_l\big) &= \mathbb{E}\Big(\mathbb{E}\big(\varepsilon^2\,X_j\,X_l\,\big|\,\boldsymbol{X}\big)\Big)\\ &= \mathbb{E}\Big(X_j\,X_l\,\mathbb{E}\big(\varepsilon^2\,\big|\,\boldsymbol{X}\big)\Big)\\ &= \mathbb{E}\big(\sigma^2(\boldsymbol{X})\,X_j\,X_l\big).\end{aligned}$$

For each $j,\, l = 0,\ldots,k-1$,

$$\mathbb{E}\big|\varepsilon^2\,X_j\,X_l\big| < \infty$$

by assumptions of Theorem. By assumption (A0), $\varepsilon_i\,X_{i,j}\,X_{i,l}$, $i = 1,\,2,\,\ldots$, is a sequence of i.i.d. random variables. Hence by Theorem C.2 (SLLN, i.i.d.),

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\,X_{i,j}\,X_{i,l} \;\xrightarrow{\text{a.s.}}\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\,X_j\,X_l\big) \qquad \text{as } n\to\infty.$$

That is, in a matrix form,

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\,\boldsymbol{X}_i\boldsymbol{X}_i^\top \;\xrightarrow{\text{a.s.}}\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\big) \;=\; \mathbb{W}^{\bigstar} \qquad \text{as } n\to\infty. \tag{13.7}$$

In the following, we show that (unobservable) squared error terms ε_i^2 in (13.7) can be replaced by squared residuals $U_{n,i}^2 = \big(Y_i \,-\, \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n\big)^2$ while keeping the same limitting matrix $\mathbb{W}^{\bigstar}$ as in (13.7).

We have

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} U_{n,i}^2\,\boldsymbol{X}_i\boldsymbol{X}_i^\top}_{\mathbb{W}_n^{\bigstar}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(Y_i \,-\, \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n\big)^2\,\boldsymbol{X}_i\,\boldsymbol{X}_i^\top$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\big(\underbrace{Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}}_{\varepsilon_i} + \boldsymbol{X}_i^\top\boldsymbol{\beta} - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n\big)^2\,\boldsymbol{X}_i\,\boldsymbol{X}_i^\top \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\,\boldsymbol{X}_i\boldsymbol{X}_i^\top}_{\mathbb{A}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)\,\boldsymbol{X}_i\boldsymbol{X}_i^\top}_{\mathbb{B}_n} \\
&\quad + \underbrace{\frac{2}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\varepsilon_i\,\boldsymbol{X}_i\boldsymbol{X}_i^\top}_{\mathbb{C}_n}.
\end{aligned}
$$

(a) $\mathbb{A}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\varepsilon_i^2\,\boldsymbol{X}_i\boldsymbol{X}_i^\top \xrightarrow{\text{a.s.}} \mathbb{E}\big(\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\big) = \mathbb{W}^\bigstar$ due to (13.7).

(b) To work with $\mathbb{B}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)\,\boldsymbol{X}_i\boldsymbol{X}_i^\top$, we can realize that $\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i = \boldsymbol{X}_i^\top\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)$ is a scalar quantity. Hence

$$
\mathbb{B}_n = \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\,\big(\boldsymbol{X}_i\boldsymbol{X}_i^\top\big)\,\boldsymbol{X}_i^\top\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)
$$

and the $(j,\,l)$th element of matrix $\mathbb{B}_n$ $(j,\,l=0,\ldots,k-1)$ is

$$
\begin{aligned}
B_n(j,\,l) &= \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\,(X_{i,j}\,X_{i,l})\,\boldsymbol{X}_i^\top\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big) \\
&= \big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\left\{\frac{1}{n}\sum_{i=1}^{n}(X_{i,j}\,X_{i,l})\,\boldsymbol{X}_i\boldsymbol{X}_i^\top\right\}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big).
\end{aligned}
$$

- From Theorem 13.2: $\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ as $n\to\infty$.
- Due to assumption (A0) and assumption $\mathbb{E}\big|X_s\,X_t\,X_j\,X_l\big| < \infty$ for any $s,\,t,\,j,\,l = 0,\ldots,k-1$, by Theorem C.2 (SLLN, i.i.d.), for any $j,\,l=0,\ldots,k-1$:

$$
\frac{1}{n}\sum_{i=1}^{n}(X_{i,j}\,X_{i,l})\,\boldsymbol{X}_i\boldsymbol{X}_i^\top \xrightarrow{\text{a.s.}} \mathbb{E}\big(X_j\,X_l\,\boldsymbol{X}\boldsymbol{X}^\top\big).
$$

- Hence, for any $j,\,l=0,\ldots,k-1$, $B_n(j,\,l) \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top\,\mathbb{E}\big(X_j\,X_l\,\boldsymbol{X}\boldsymbol{X}^\top\big)\,\mathbf{0}_k = 0$ and finally,

$$
\mathbb{B}_n \xrightarrow{\text{a.s.}} \mathbf{0}_{k\times k} \qquad \text{as } n\to\infty.
$$

(c) $\mathbb{C}_n = \dfrac{2}{n}\displaystyle\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\varepsilon_i\,\boldsymbol{X}_i\boldsymbol{X}_i^\top$ and the $(j,\,l)$th element of matrix $\mathbb{C}_n$ $(j,\,l=0,\ldots,k-1)$ is

$$
\begin{aligned}
C_n(j,\,l) &= \frac{2}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\boldsymbol{X}_i\varepsilon_i\,X_{i,j}X_{i,l} \\
&= 2\,\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\,X_{i,j}X_{i,l}\right).
\end{aligned}
$$

- From Theorem 13.2: $\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\right) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ as $n \to \infty$.
- Due to assumption (A0) and assumption $\mathbb{E}\left|\varepsilon\, X_s\, X_j\, X_l\right| < \infty$ for any $s,\, j,\, l = 0, \ldots, k-1$, by Theorem C.2 (SLLN, i.i.d.), for any $j,\, l = 0, \ldots, k-1$:

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \varepsilon_i\, X_{i,j} X_{i,l} \xrightarrow{\text{a.s.}} \mathbb{E}\big(\boldsymbol{X} \varepsilon\, X_j X_l\big).$$

- Hence, for any $j,\, l = 0, \ldots, k-1$, $C_n(j,\, l) \xrightarrow{\text{a.s.}} 2\, \mathbf{0}_k^\top\, \mathbb{E}\big(\boldsymbol{X} \varepsilon\, X_j X_l\big) = 0$ and finally,

$$\mathbb{C}_n \xrightarrow{\text{a.s.}} \mathbf{0}_{k \times k} \qquad \text{as } n \to \infty.$$

In summary:

$$\begin{aligned} n\, \mathbb{V}_n \mathbb{W}_n^{\bigstar} \mathbb{V}_n &= n\, \mathbb{V}_n \left(\frac{1}{n} \mathbb{W}_n^{\bigstar}\right) n\, \mathbb{V}_n \\ &= n\, \mathbb{V}_n \left(\mathbb{A}_n + \mathbb{B}_n + \mathbb{C}_n\right) n\, \mathbb{V}_n, \end{aligned}$$

where $n\, \mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$,

$\mathbb{A}_n \xrightarrow{\text{a.s.}} \mathbb{W}^{\bigstar}$,

$\mathbb{B}_n \xrightarrow{\text{a.s.}} \mathbf{0}_{k \times k}$,

$\mathbb{C}_n \xrightarrow{\text{a.s.}} \mathbf{0}_{k \times k}$.

Hence

$$n\, \mathbb{V}_n \mathbb{W}_n^{\bigstar} \mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V} \left(\mathbb{W}^{\bigstar} + \mathbf{0}_{k \times k} + \mathbf{0}_{k \times k}\right) \mathbb{V} = \mathbb{V} \mathbb{W}^{\bigstar} \mathbb{V} \qquad \text{as } n \to \infty.$$

❑

Terminology *(Heteroscedasticity consistent (sandwich) estimator of the covariance matrix).*

Matrix

$$\mathbb{V}_n\, \mathbb{W}_n^{\bigstar}\, \mathbb{V}_n = \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \mathbb{X}_n^\top\, \boldsymbol{\Omega}_n\, \mathbb{X}_n \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \tag{13.8}$$

is called the *heteroscedasticity consistent (HC)* estimator of the covariance matrix of the LSE $\widehat{\boldsymbol{\beta}}_n$ of the regression coefficients. Due to its form, the matrix (13.8) is also called as the *sandwich* estimator composed of a bread $\left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \mathbb{X}_n^\top$ and a meat $\boldsymbol{\Omega}_n$.

Notes *(Alternative sorts of meat for the sandwich).*

- It is directly seen that the meat matrix $\boldsymbol{\Omega}_n$ can, for a chosen sequence ν_n, such that $\frac{n}{\nu_n} \to 1$ as $n \to \infty$, be replaced by a matrix

$$\frac{n}{\nu_n}\, \boldsymbol{\Omega}_n,$$

and the statement of Theorem 13.6 remains valid. A value ν_n is then called *degrees of freedom* of the sandwich.

- It can also be shown (see references below) that the meat matrix $\mathbf{\Omega}_n$ can, for a chosen sequence ν_n, such that $\frac{n}{\nu_n} \to 1$ as $n \to \infty$ and a suitable sequence $\boldsymbol{\delta}_n = \big(\delta_{n,1}, \ldots, \delta_{n,n}\big)$, $n = 1, 2, \ldots$, be replaced by a matrix

$$\mathbf{\Omega}_n^{HC} := \mathsf{diag}\big(\omega_{n,1}, \ldots, \omega_{n,n}\big),$$
$$\omega_{n,i} = \frac{n}{\nu_n} \frac{U_{n,i}^2}{m_{n,i}^{\delta_{n,i}}}, \qquad i = 1, \ldots, n.$$

- The following choices of sequences ν_n and $\boldsymbol{\delta}_n$ have appeared in the literature ($n = 1, 2, \ldots$, $i = 1, \ldots, n$):

  HC0: $\nu_n = n$, $\delta_{n,i} = 0$, that is,
  $$\omega_{n,i} = U_{n,i}^2.$$
  This is the choice due to White (1980) who was the first who proposed the sandwich estimator of the covariance matrix. This choice was also used in Theorem 13.6.

  HC1: $\nu_n = n - k$, $\delta_{n,i} = 0$, that is,
  $$\omega_{n,i} = \frac{n}{n-k} U_{n,i}^2.$$
  This choice was suggested by MacKinnon and White (1985).

  HC2: $\nu_n = n$, $\delta_{n,i} = 1$, that is,
  $$\omega_{n,i} = \frac{U_{n,i}^2}{m_{n,i}}.$$
  This is the second proposal of MacKinnon and White (1985).

  HC3: $\nu_n = n$, $\delta_{n,i} = 2$, that is,
  $$\omega_{n,i} = \frac{U_{n,i}^2}{m_{n,i}^2}.$$
  This is the third proposal of MacKinnon and White (1985).

  HC4: $\nu_n = n$, $\delta_{n,i} = \min\big\{4, n\, h_{n,i}/k\big\}$, that is,
  $$\omega_{n,i} = \frac{U_{n,i}^2}{m_{n,i}^{\delta_{n,i}}}.$$
  This was proposed relatively recently by Cribari-Neto (2004). Note that $k = \sum_{i=1}^n h_{n,i}$, and hence
  $$\delta_{n,i} = \min\Big\{4, \frac{h_{n,i}}{\overline{h}_n}\Big\}, \qquad \overline{h}_n = \frac{1}{n} \sum_{i=1}^n h_{n,i}.$$

- An extensive study towards small sample behavior of different sandwich estimators was carried out by Long and Ervin (2000) who recommended usage of the HC3 estimator. Even better small sample behavior, especially in presence of influential observations was later concluded by Cribari-Neto (2004) for the HC4 estimator.

- Labels HC0, HC1, HC2, HC3, HC4 for the above sandwich estimators are used by the R package `sandwich` (Zeileis, 2004) that enables for their easy calculation based on the fitted linear model.

13.4.1 Heteroscedasticity consistent asymptotic inference

Let for given sequences ν_n and $\boldsymbol{\delta}_n$, $n = 1, 2, \ldots$, $\boldsymbol{\Omega}_n^{HC}$ be a sequence of the meat matrices that lead to the heteroscedasticity consistent estimator of the covariance matrix of the LSE $\widehat{\boldsymbol{\beta}}_n$. Let for given $n \geq n_0 > k$,

$$\mathbb{V}_n^{HC} := \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \mathbb{X}_n^\top \, \boldsymbol{\Omega}_n^{HC} \, \mathbb{X}_n \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1}.$$

Finally, let the statistics T_n^{HC} and Q_n^{HC} be defined as

$$T_n^{HC} := \frac{\widehat{\theta}_n - \theta}{\sqrt{\mathbf{l}^\top \mathbb{V}_n^{HC} \mathbf{l}}},$$

$$Q_n^{HC} := \frac{1}{m} \left(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\right)^\top \left(\mathbb{L} \mathbb{V}_n^{HC} \mathbb{L}^\top\right)^{-1} \left(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\right).$$

Note that the statistics T_n^{HC} and Q_n^{HC}, respectively, are the usual statistics T_n (Eq. 13.1) and Q_n (13.2), respectively, in which the term $\mathsf{MS}_{e,n} \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1}$ is replaced by the sandwich estimator $\mathbb{V}_n^{HC}$.

Consequence of Theorems 13.5 and 13.6: Heteroscedasticity consistent asymptotic inference.

Under assumptions of Theorem 13.5 and 13.6:

$$T_n^{HC} \xrightarrow{\mathcal{D}} \mathcal{N}_1(0, 1) \qquad \text{as } n \to \infty,$$

$$m\, Q_n^{HC} \xrightarrow{\mathcal{D}} \chi^2_m \qquad \text{as } n \to \infty.$$

Proof. **Proof/calculations were available on the blackboard in Kl.** ❑

Due to a general asymptotic property of the Student t-distribution (Eq. 13.3), asymptotically valid inference on the estimable parameter $\theta = \mathbf{l}^\top \boldsymbol{\beta}$ of a linear model where neither normality, nor homoscedasticity is necessarily satisfied, can be based on the statistic T_n^{HC} and either a Student t_{n-k} or a standard normal distribution. Under assumptions of Theorems 13.5 and 13.6, both intervals

(i) $\mathcal{I}_n^{\mathcal{N}} := \left(\widehat{\theta}_n - u(1-\alpha/2) \sqrt{\mathbf{l}^\top \mathbb{V}_n^{HC} \mathbf{l}}, \quad \widehat{\theta}_n + u(1-\alpha/2) \sqrt{\mathbf{l}^\top \mathbb{V}_n^{HC} \mathbf{l}}\right)$;

(ii) $\mathcal{I}_n^{\mathsf{t}} := \left(\widehat{\theta}_n - \mathsf{t}_{n-k}(1-\alpha/2) \sqrt{\mathbf{l}^\top \mathbb{V}_n^{HC} \mathbf{l}}, \quad \widehat{\theta}_n + \mathsf{t}_{n-k}(1-\alpha/2) \sqrt{\mathbf{l}^\top \mathbb{V}_n^{HC} \mathbf{l}}\right)$,

satisfy, for any $\theta^0 \in \mathbb{R}$:

$$\mathsf{P}\left(\mathcal{I}_n^{\mathcal{N}} \ni \theta^0;\ \theta = \theta^0\right) \longrightarrow 1 - \alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\left(\mathcal{I}_n^{\mathsf{t}} \ni \theta^0;\ \theta = \theta^0\right) \longrightarrow 1 - \alpha \qquad \text{as } n \to \infty.$$

Analogously, due to a general asymptotic property of the F-distribution (Eq. 13.3), asymptotically valid inference on the estimable vector parameter $\boldsymbol{\xi} = \mathbb{L}\boldsymbol{\beta}$ of a linear model can be based either on the statistic $m\,Q_n^{HC}$ and the χ^2_m distribution or on the statistic Q_n^{HC} and the $\mathcal{F}_{m,\,n-k}$ distribution. For example, for both ellipsoids

$$\text{(i)}\quad \mathcal{K}_n^{\chi} := \left\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \left(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\right)^{\top} \left(\mathbb{L}\,\mathbb{V}_n^{HC}\mathbb{L}^{\top}\right)^{-1} \left(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\right) \,<\, \chi^2_m(1-\alpha)\right\};$$

$$\text{(ii)}\quad \mathcal{K}_n^{\mathcal{F}} := \left\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \left(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\right)^{\top} \left(\mathbb{L}\,\mathbb{V}_n^{HC}\mathbb{L}^{\top}\right)^{-1} \left(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\right) \,<\, m\,\mathcal{F}_{m,n-k}(1-\alpha)\right\},$$

we have for any $\boldsymbol{\xi}^0 \in \mathbb{R}^m$ (under assumptions of Theorems 13.5 and 13.6):

$$\mathsf{P}\left(\mathcal{K}_n^{\chi} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\right) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\left(\mathcal{K}_n^{\mathcal{F}} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\right) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

Chapter 14

Unusual Observations

In the whole chapter, we assume a linear model

$$\mathsf{M}: \ \boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big), \quad \mathsf{rank}(\mathbb{X}_{n\times k}) = r \le k,$$

where standard notation is considered. That is,

- $\boldsymbol{b} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top\boldsymbol{Y} = \big(b_0,\, \ldots,\, b_{k-1}\big)^\top$: any solution to normal equations;
- $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top = \big(h_{i,t}\big)_{i,t=1,\ldots,n}$: the hat matrix;
- $\mathbb{M} = \mathbf{I}_n - \mathbb{H} = \big(m_{i,t}\big)_{i,t=1,\ldots,n}$: the residual projection matrix;
- $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \mathbb{X}\boldsymbol{b} = \big(\widehat{Y}_1,\, \ldots,\, \widehat{Y}_n\big)^\top$: the vector of fitted values;
- $\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \big(U_1,\, \ldots,\, U_n\big)^\top$: the residuals;
- $\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2$: the residual sum of squares;
- $\mathsf{MS}_e = \frac{1}{n-r}\,\mathsf{SS}_e$ is the residual mean square;
- $\boldsymbol{U}^{std} = \big(U_1^{std},\, \ldots,\, U_n^{std}\big)^\top$: vector of standardized residuals,
  $U_i^{std} = \frac{U_i}{\sqrt{\mathsf{MS}_e\, m_{i,i}}}$, $i = 1, \ldots, n$.

The whole chapter will deal with idntification of "unusual" observations in a particular dataset. Any probabilistic statements will hence be conditioned by the realized covariate values $\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n$. The same symbol $\mathbb{X}$ will be used for (in general random) model matrix and its realized counterpart, i.e.,

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}.$$

14.1 Leave-one-out and outlier model

Notation. For chosen $t \in \{1, \ldots, n\}$, we will use the following notation.

- $\boldsymbol{Y}_{(-t)}$: vector $\boldsymbol{Y}$ without the tth element;
- $\boldsymbol{x}_t$: the tth row (understood as a column vector) of the matrix $\mathbb{X}$;
- $\mathbb{X}_{(-t)}$: matrix $\mathbb{X}$ without the tth row;
- $\boldsymbol{j}_t$: vector $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$ of length n with 1 on the tth place.

Definition 14.1 Leave-one-out model.

The tth leave-one-out model[1] *is a linear model*

$$\mathsf{M}_{(-t)}: \quad \boldsymbol{Y}_{(-t)} \,|\, \mathbb{X}_{(-t)} \sim \left(\mathbb{X}_{(-t)}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_{n-1}\right).$$

Definition 14.2 Outlier model.

The tth outlier model[2] *is a linear model*

$$\mathsf{M}_t^{out}: \quad \boldsymbol{Y} \,|\, \mathbb{X} \sim \left(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{j}_t\gamma_t^{out},\, \sigma^2 \mathbf{I}_n\right).$$

Notation *(Quantities related to the leave-one-out and outlier models).*

- Quantities related to model $\mathsf{M}_{(-t)}$ will be recognized by subscript $(-t)$, i.e.,

$$\boldsymbol{b}_{(-t)}, \widehat{\boldsymbol{Y}}_{(-t)}, \mathsf{SS}_{e,(-t)}, \mathsf{MS}_{e,(-t)}, \ldots$$

- Quantities related to model M_t^{out} will be recognized by subscript t and superscript out, i.e.,

$$\boldsymbol{b}_t^{out}, \widehat{\boldsymbol{Y}}_t^{out}, \mathsf{SS}_{e,t}^{out}, \mathsf{MS}_{e,t}^{out}, \ldots$$

- Solutions to normal equations in model M_t^{out} will be denoted as

$$\left((\boldsymbol{b}_t^{out})^\top, (c_t^{out})^\top\right)^\top.$$

- If γ_t^{out} is an estimable parameter of model M_t^{out} then its LSE will be denoted as $\widehat{\gamma}_t^{out}$.

---

[1] *model vynechaného $t$tého pozorování* [2] *model $t$tého odlehlého pozorování*

Theorem 14.1 Four equivalent statements.

The following four statements are equivalent:

(i) $\mathsf{rank}(\mathbb{X}) = \mathsf{rank}\big(\mathbb{X}_{(-t)}\big)$, *i.e.,* $\boldsymbol{x}_t \in \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big)$;

(ii) $m_{t,t} > 0$;

(iii) γ_t^{out} *is an estimable parameter of model* M_t^{out};

(iv) $\mu_t := \mathbb{E}\big(Y_t \,\big|\, \boldsymbol{X}_t = \boldsymbol{x}_t\big) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$ *is an estimable parameter of model* $\mathsf{M}_{(-t)}$.

Proof.

(ii) $\Leftrightarrow$ (i)

- We will show this by showing non(i) $\Leftrightarrow$ non(ii).
- non(i) means that $\boldsymbol{x}_t \notin \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big) \subset \mathcal{M}\big(\mathbb{X}^\top\big)$.

$$\begin{aligned} & \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big) \subset \mathcal{M}\big(\mathbb{X}^\top\big) \quad \text{and } \mathcal{M}\big(\mathbb{X}_{(-t)}\big) \neq \mathcal{M}\big(\mathbb{X}^\top\big). \\ \Leftrightarrow\ & \mathcal{M}\big(\mathbb{X}^\top\big)^\perp \subset \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big)^\perp \quad \text{and } \mathcal{M}\big(\mathbb{X}^\top\big)^\perp \neq \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big)^\perp. \end{aligned}$$

- That is,
$$\begin{aligned} & \Leftrightarrow \exists \boldsymbol{a} \in \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big)^\perp \text{ such that } \boldsymbol{a} \notin \mathcal{M}\big(\mathbb{X}^\top\big)^\perp. \\ & \Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^k \text{ such that } \boldsymbol{a}^\top \mathbb{X}_{(-t)}^\top = \mathbf{0}^\top \ \&\ \boldsymbol{a}^\top \mathbb{X}^\top \neq \mathbf{0}^\top. \\ & \Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^k \text{ such that } \mathbb{X}_{(-t)}\boldsymbol{a} = \mathbf{0} \ \&\ \mathbb{X}\boldsymbol{a} \neq \mathbf{0}. \end{aligned}$$

  It must be
$$\mathbb{X}\boldsymbol{a} = \big(0, \ldots, 0, c, 0, \ldots, 0\big)^\top = c\,\boldsymbol{j}_t$$
  for some $c \neq 0$.

$\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^k$ such that $\mathbb{X}\boldsymbol{a} = c\boldsymbol{j}_t,\ c \neq 0$.

$\Leftrightarrow \boldsymbol{j}_t \in \mathcal{M}\big(\mathbb{X}\big)$.

$\Leftrightarrow \underbrace{\mathbb{M}\boldsymbol{j}_t}_{t\text{th column of } \mathbb{M}} = \mathbf{0}$.

$\Leftrightarrow \boldsymbol{m}_t = \mathbf{0}$.

$\Leftrightarrow \big\|\boldsymbol{m}_t\big\|^2 = m_{t,t} = 0$.

$\Leftrightarrow$ non(ii).

$\boldsymbol{m}_t$ denotes the tth row of $\mathbb{M}$ (and also its t column since $\mathbb{M}$ is symmetric).

(iii) $\Leftrightarrow$ (i)

- $\gamma_t^{out} = \underbrace{\big(\mathbf{0}_k^\top,\, 1\big)^\top}_{\mathbf{l}^\top} \begin{pmatrix} \boldsymbol{\beta} \\ \gamma_t^{out} \end{pmatrix}$.

- γ_t^{out} is estimable parameter of M_t^{out} $\Leftrightarrow \mathbf{l} \in \mathcal{M}\Big(\big(\mathbb{X},\, \boldsymbol{j}_t\big)^\top\Big)$.
  $\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^n$ such that $\big(\mathbf{0}^\top,\, 1\big) = \boldsymbol{a}^\top\big(\mathbb{X},\, \boldsymbol{j}_t\big)$.
  $\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^n$ such that $\mathbf{0}^\top = \boldsymbol{a}^\top\mathbb{X}\ \ \&\ \ 1 = \boldsymbol{a}^\top\boldsymbol{j}_t$.
  $\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^n$ such that $\mathbf{0}^\top = \boldsymbol{a}^\top\mathbb{X}\ \ \&\ \ a_t = 1$.
  $\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^n$ such that $\boldsymbol{x}_t^\top = -\,\boldsymbol{a}_{(-t)}^\top\mathbb{X}_{(-t)}$.
  $\Leftrightarrow \boldsymbol{x}_t \in \mathcal{M}\big(\mathbb{X}_{(-t)}^\top\big)$.
  $\Leftrightarrow$ (i).

(iv) $\Leftrightarrow$ (i)

- Follows directly from Theorem 2.7.

Theorem 14.2 Equivalence of the outlier model and the leave-one-out model.

1. *The residual sums of squares in models* $\mathsf{M}_{(-t)}$ *and* M_t^{out} *are the same, i.e.,*
$$\mathsf{SS}_{e,(-t)} = \mathsf{SS}_{e,t}^{out}.$$
2. *Vector* $\boldsymbol{b}_{(-t)}$ *solves the normal equations of model* $\mathsf{M}_{(-t)}$ *if and only if a vector* $\big((\boldsymbol{b}_t^{out})^\top,\, (c_t^{out})^\top\big)^\top$ *solves the normal equations of model* M_t^{out}*, where*
$$\begin{aligned} \boldsymbol{b}_t^{out} &= \boldsymbol{b}_{(-t)}, \\ c_t^{out} &= Y_t - \boldsymbol{x}_t{}^\top\boldsymbol{b}_{(-t)}. \end{aligned}$$

Proof.

Solution to normal equations minimizes the corresponding sum of squares.

The sum of squares to be minimized w.r.t. $\boldsymbol{\beta}$ and γ_t^{out} in the outlier model M_t^{out} is

$$\begin{aligned} \mathsf{SS}_t^{out}\big(\boldsymbol{\beta},\, \gamma_t^{out}\big) &= \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} - \boldsymbol{j}_t\gamma_t^{out}\big\|^2 && \text{separate the } t\text{th element of the sum} \\ &= \big\|\boldsymbol{Y}_{(-t)} - \mathbb{X}_{(-t)}\boldsymbol{\beta}\big\|^2 \;+\; \big(Y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta} - \gamma_t^{out}\big)^2 \\ &= \mathsf{SS}_{(-t)}(\boldsymbol{\beta}) \;+\; \big(Y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta} - \gamma_t^{out}\big)^2, \end{aligned}$$

where $\mathsf{SS}_{(-t)}(\boldsymbol{\beta})$ is the sum of squares to be minimized w.r.t. $\boldsymbol{\beta}$ in the leave-one-out model $\mathsf{M}_{(-t)}$.

The term $\big(Y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta} - \gamma_t^{out}\big)^2$ can for any $\boldsymbol{\beta} \in \mathbb{R}^k$ be equal to zero if we, for given $\boldsymbol{\beta} \in \mathbb{R}^k$, take

$$\gamma_t^{out} = Y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta}.$$

That is

(i) $\underbrace{\min_{\boldsymbol{\beta},\,\gamma_t^{out}} \mathsf{SS}_t^{out}(\boldsymbol{\beta},\,\gamma_t^{out})}_{\mathsf{SS}_{e,t}^{out}} \;=\; \underbrace{\min_{\boldsymbol{\beta}} \mathsf{SS}_{(-t)}(\boldsymbol{\beta})}_{\mathsf{SS}_{e,(-t)}};$

(ii) A vector $\boldsymbol{b}_{(-t)} \in \mathbb{R}^k$ minimizes $\mathsf{SS}_{(-t)}(\boldsymbol{\beta})$ if and only if a vector

$$\big(\underbrace{\boldsymbol{b}_{(-t)}^\top}_{\boldsymbol{b}_t^{out}},\; \underbrace{Y_t - \boldsymbol{x}_t^\top \boldsymbol{b}_{(-t)}}_{c_t^{out}}\big)^\top \in \mathbb{R}^{k+1}$$

minimizes $\mathsf{SS}_t^{out}(\boldsymbol{\beta},\,\gamma_t^{out})$.

❑

Notation *(Leave-one-out least squares estimators of the response expectations).*

If $m_{t,t} > 0$ for all $t = 1, \ldots, n$, we will use the following notation:

$$\widehat{Y}_{[t]} := \boldsymbol{x}_t^\top \boldsymbol{b}_{(-t)}, \quad t = 1, \ldots, n,$$

which is the LSE of the parameter $\mu_t = \mathbb{E}\big(Y_t \,\big|\, \boldsymbol{X}_t = \boldsymbol{x}_t\big) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$ based on the *leave-one-out* model $\mathsf{M}_{(-t)}$;

$$\widehat{\boldsymbol{Y}}_{[\bullet]} := \big(\widehat{Y}_{[1]}, \ldots, \widehat{Y}_{[n]}\big)^\top,$$

which is an estimator of the parameter $\boldsymbol{\mu} = \big(\mu_1, \ldots, \mu_n\big)^\top = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)$, where each element is estimated using the linear model based on data with the corresponding observation being left out.

Calculation of quantities of the outlier and the leave-one-out models

Model M_t^{out} is a model with added regressor for model M. Suppose that $m_{t,t} > 0$ for given $t = 1, \ldots, n$. By applying Lemma 11.1, we can express the LSE of the parameter γ_t^{out} as

$$\widehat{\gamma}_t^{out} \;=\; \big(\boldsymbol{j}_t^\top \mathbb{M} \boldsymbol{j}_t\big)^- \boldsymbol{j}_t^\top \boldsymbol{U} \;=\; (m_{t,t})^- U_t \;=\; (m_{t,t})^{-1} U_t \;=\; \frac{U_t}{m_{t,t}}.$$

Analogously, other quantities of the outlier model can be expressed using the quantities of model M. Namely,

$$\begin{aligned}
\boldsymbol{b}_t^{out} &= \boldsymbol{b} - \frac{U_t}{m_{t,t}} \big(\mathbb{X}^\top \mathbb{X}\big)^- \boldsymbol{x}_t, \\
\widehat{\boldsymbol{Y}}_t^{out} &= \widehat{\boldsymbol{Y}} + \frac{U_t}{m_{t,t}} \boldsymbol{m}_t, \\
\mathsf{SS}_e - \mathsf{SS}_{e,t}^{out} &= \frac{U_t^2}{m_{t,t}} = \mathsf{MS}_e \big(U_t^{std}\big)^2,
\end{aligned}$$

where $\boldsymbol{m}_t$ denotes the tth column (and row as well) of the residual projection matrix $\mathbb{M}$.

Lemma 14.3 Quantities of the outlier and leave-one-out model expressed using quantities of the original model.

Suppose that for given $t \in \{1, \ldots, n\}$, $m_{t,t} > 0$. The following quantities of the outlier model M_t^{out} and the leave-one-out model $\mathsf{M}_{(-t)}$ are expressable using the quantities of the original model M as follows.

$$\begin{aligned} \widehat{\gamma}_t^{out} &= Y_t - \boldsymbol{x}_t^\top \boldsymbol{b}_{(-t)} = Y_t - \widehat{Y}_{[t]} = \frac{U_t}{m_{t,t}}, \\ \boldsymbol{b}_{(-t)} = \boldsymbol{b}_t^{out} &= \boldsymbol{b} - \frac{U_t}{m_{t,t}} \left(\mathbb{X}^\top \mathbb{X}\right)^- \boldsymbol{x}_t, \\ \mathsf{SS}_{e,(-t)} = \mathsf{SS}_{e,t}^{out} &= \mathsf{SS}_e - \frac{U_t^2}{m_{t,t}} = \mathsf{SS}_e - \mathsf{MS}_e \left(U_t^{std}\right)^2, \\ \frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e} = \frac{\mathsf{MS}_{e,t}^{out}}{\mathsf{MS}_e} &= \frac{n - r - \left(U_t^{std}\right)^2}{n - r - 1}. \end{aligned} \tag{14.1}$$

Proof. Equality between the quantities of the outlier and the leave-one-out model follows from Theorem 14.2. Remaining expressions follow from previously conducted calculations.

To see the last equality in (14.1), remember that the residual degrees of freedom of both the outlier and the leave-one-out models are equal to $n - r - 1$. That is, whereas in model M,

$$\mathsf{MS}_e = \frac{\mathsf{SS}_e}{n - r},$$

in the outlier and the leave-one-out model,

$$\mathsf{MS}_{e,(-t)} = \frac{\mathsf{SS}_{e,(-t)}}{n - r - 1} = \frac{\mathsf{SS}_{e,t}^{out}}{n - r - 1} = \mathsf{MS}_{e,t}^{out}.$$

❑

Notes.

- Expressions in Lemma 14.3 quantify the influence of the tth observation on
  - (i) the LSE of a vector $\boldsymbol{\beta}$ of the regression coefficients (in case they are estimable);
  - (ii) the estimate of the residual variance.
- Lemma 14.3 also shows that it is not necessary to fit n leave-one-out (or outlier models) to calculate their LSE-related quantities. All important quantities can be calculated directly from the LSE-related quantities of the original model M.

Definition 14.3 Deleted residual.

If $m_{t,t} > 0$*, then the quantity*

$$\widehat{\gamma}_t^{out} = Y_t - \widehat{Y}_{[t]} \; = \; \frac{U_t}{m_{t,t}}$$

is called the t*th* deleted residual *of the model* M.

14.2 Outliers

By *outliers*[3] of the model M, we shall understand observations for which the response expectation does not follow the assumed model, i.e., the tth observation ($t \in \{1, \ldots, n\}$) is an outlier if

$$\mathbb{E}\big(Y_t \,\big|\, \boldsymbol{X}_t = \boldsymbol{x}_t\big) \neq \boldsymbol{x}_t^\top \boldsymbol{\beta},$$

in which case we can write

$$\mathbb{E}\big(Y_t \,\big|\, \boldsymbol{X}_t = \boldsymbol{x}_t\big) = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \gamma_t^{out}.$$

As such, an outlier can be characterized as an observation with *unusual* response (y) value.

If $m_{t,t} > 0$, γ_t^{out} is an estimable parameter of the tth outlier model M_t^{out} (for which the model M is a submodel) and decision on whether the tth observation is an outlier can be transferred into a problem of testing

$$\mathrm{H}_0\colon\ \gamma_t^{out} = 0$$

in the tth outlier model M_t^{out}. Note that the above null hypothesis also expresses the fact that the submodel M of the model M_t^{out} holds.

If *normality* is assumed, this null hypothesis can be tested using a classical t-test on a value of the estimable parameter. The corresponding t-statistic has a standard form

$$T_t = \frac{\widehat{\gamma}_t^{out}}{\sqrt{\widehat{\mathsf{var}}\big(\widehat{\gamma}_t^{out}\big)}}$$

and under the null hypothesis follows the Student t distribution with $n - r - 1$ degrees of freedom (residual degrees of freedom of the outlier model).

From Section 14.1, we have

$$\widehat{\gamma}_t^{out} = \frac{U_t}{m_{t,t}} = Y_t - \widehat{Y}_{[t]}.$$

Hence (the variance is conditional given the covariate values),

$$\mathsf{var}\big(\widehat{\gamma}_t^{out} \,\big|\, \mathbb{X}\big) \;=\; \mathsf{var}\left(\frac{U_t}{m_{t,t}} \,\middle|\, \mathbb{X}\right) \;=\; \frac{1}{m_{t,t}^2}\,\mathsf{var}\big(U_t \,\big|\, \mathbb{X}\big) \;\overset{(\star)}{=}\; \frac{1}{m_{t,t}^2}\,\sigma^2\, m_{t,t} \;=\; \frac{\sigma^2}{m_{t,t}}.$$

The equality $\overset{(\star)}{=}$ holds irrespective of whether $\gamma_t^{out} = 0$ (and model M holds) or $\gamma_t^{out} \neq 0$ (and model M_t^{out} holds).

The estimator $\widehat{\gamma}_t^{out}$ is the LSE of a parameter of the outlier model and hence

$$\widehat{\mathsf{var}}\big(\widehat{\gamma}_t^{out} \,\big|\, \mathbb{X}\big) = \frac{\mathsf{MS}_{e,t}^{out}}{m_{t,t}},$$

and finally,

$$T_t = \frac{\widehat{\gamma}_t^{out}}{\sqrt{\dfrac{\mathsf{MS}_{e,t}^{out}}{m_{t,t}}}}.$$

Two useful expressions of the statistic T_t are obtained by remembering from Section 14.1 (a) $\mathsf{MS}_{e,t}^{out} = \mathsf{MS}_{e,(-t)}$ and (b) two expressions of $\widehat{\gamma}_t^{out} = Y_t - \widehat{Y}_{[t]} = \widehat{\gamma}_t^{out} = \frac{U_t}{m_{t,t}}$. This leads to

$$T_t = \frac{Y_t - \widehat{Y}_{[t]}}{\sqrt{\mathsf{MS}_{e,(-t)}}}\sqrt{m_{t,t}} = \frac{U_t}{\sqrt{\mathsf{MS}_{e,(-t)}\, m_{t,t}}}.$$

[3] *odlehlá pozorování*

Definition 14.4 Studentized residual.

If $m_{t,t} > 0$, then the quantity

$$T_t = \frac{Y_t - \widehat{Y}_{[t]}}{\sqrt{\mathsf{MS}_{e,(-t)}}}\sqrt{m_{t,t}} = \frac{U_t}{\sqrt{\mathsf{MS}_{e,(-t)}\, m_{t,t}}}$$

is called the tth studentized residual[4] *of the model* M.

Notes.

- Using the last equality in (14.1), we can derive one more expression of the studentized residual using the standardized residual

  $$U_t^{std} = \frac{U_t}{\sqrt{\mathsf{MS}_e\, m_{t,t}}}.$$

  Namely,

  $$T_t = \sqrt{\frac{n-r-1}{n-r-\left(U_t^{std}\right)^2}}\; U_t^{std}.$$

  This directly shows that it is not necessary to fit the leave-one-out or the outlier model to calculate the studentized residual of the initial model M.

Theorem 14.4 On studentized residuals.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, where $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \le k < n$. Let further $n > r+1$. Let for given $t \in \{1,\ldots,n\}$ $m_{t,t} > 0$. Then

1. *The tth studentized residual T_t follows the Student t-distribution with $n-r-1$ degrees of freedom.*
2. *If additionally $n > r+2$ then $\mathbb{E}\big(T_t\big) = 0$.*
3. *If additionally $n > r+3$ then $\mathsf{var}\big(T_t\big) = \dfrac{n-r-1}{n-r-3}$.*

Proof. Point (i) follows from preceeding derivations, points (ii) and (iii) follow from properties of the Student t distribution.

Test for outliers

The studentized residual T_t of the model M is the test statistic (with $\mathtt{t}_{n-r-1}$ distribution under the null hypothesis) of the test

[4] *studentizované reziduum*

H_0: $\gamma_t^{out} = 0$,

H_1: $\gamma_t^{out} \neq 0$

in the tth outlier model M_t^{out}: $\boldsymbol{Y} \mid \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{j}_t\gamma_t^{out},\, \sigma^2\mathbf{I}_n\big)$.

The above testing problem can also be interpreted as a test of

H_0: tth observations is not outlier of model M,

H_1: tth observations is outlier of model M,

where "outlier" means outlier with respect to model M: $\boldsymbol{Y} \mid \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$:

- The expected value of the tth observation is different from that given by model M;
- The observed value of Y_t is unusual under model M.

When performing the test for outliers for all observations in the dataset, we are in fact facing a multiple testing problem and hence adjustment of the P-values resulted from comparison of the values of the studentized residuals with the quantiles of the Student t_{n-r-1} distribution are needed to keep the rate of falsely identified outliers under the requested level of α. For example, Bonferroni adjustment can be used.

Notes.

- Two or more outliers next to each other can hide each other.
- A notion of outlier is always relative to considered model (also in other areas of statistics). Observation which is outlier with respect to one model is not necessarily an outlier with respect to some other model.
- Especially in large datasets, few outliers are not a problem provided they are not at the same time also influential for statistical inference (see next section).
- In a context of a normal linear model, presence of outliers may indicate that the error distribution is some distribution with heavier tails than the normal distribution.
- Outlier can also suggest that a particular observation is a data-error.
- If some observation is indicated to be an outlier, it should always be explored:
  - Is it a data-error? If yes, try to correct it, if this is impossible, no problem (under certain assumptions) to exclude it from the data.
  - Is the assumed model correct and it is possible to find a physical/practical explanation for occurrence of such unusual observation?
  - If an explanation is found, are we interested in capturing such artefacts by our model or not?
  - Do the outlier(s) show a serious deviation from the model that cannot be ignored (for the purposes of a particular modelling)?
  - ⋮
- ***NEVER, NEVER, NEVER*** exclude "outliers" from the analysis in an automatic manner.
- Often, identification of outliers with respect to some model is of primary interest:
  - Example: model for amount of credit card transactions over a certain period of time depending on some factors (age, gender, income, ...).
  - Model found to be correct for a "standard" population (of clients).
  - Outlier with respect to such model $\equiv$ potentially a fraudulent use of the credit card.
- If the closer analysis of "outliers" suggest that the assumed model is not satisfactory capturing the reality we want to capture (it is not useful), some other model (maybe not linear, maybe not normal) must be looked for.

14.3 Leverage points

By *leverage points*[5] of the model M, we shall understand observations with, in a certain sense, *unusual* regressor (x) values. As will be shown, the fact whether the regressor values of a certain observation are unusual is closely related to the diagonal elements $h_{1,1}, \ldots, h_{n,n}$ of the hat matrix $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top$ of the model.

Terminology *(Leverage).*

A diagonal element $h_{t,t}$ ($t = 1, \ldots, n$) of the hat matrix $\mathbb{H}$ is called the *leverage* of the tth observation.

Interpretation of the leverage

To show that the leverage expresses how unusual the regressor values of the tth observations are, let us consider a linear model with intercept, i.e., the realized model matrix is

$$\mathbb{X} = \big(\mathbf{1}_n,\, \boldsymbol{x}^1,\, \ldots,\, \boldsymbol{x}^{k-1}\big),$$

where

$$\boldsymbol{x}^1 = \begin{pmatrix} x_{1,1} \\ \vdots \\ x_{n,1} \end{pmatrix}, \quad \ldots, \quad \boldsymbol{x}^{k-1} = \begin{pmatrix} x_{1,k-1} \\ \vdots \\ x_{n,k-1} \end{pmatrix}.$$

Let

$$\overline{x}^1 = \frac{1}{n}\sum_{i=1}^{n} x_{i,1}, \quad \ldots, \quad \overline{x}^{k-1} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k-1}$$

be the means of the non-intercept columns of the model matrix. That is, a vector

$$\overline{\boldsymbol{x}} = \big(\overline{x}^1,\, \ldots,\, \overline{x}^{k-1}\big)^\top$$

provides the mean values of the non-intercept regressors included in the model matrix $\mathbb{X}$ and as such is a gravity centre of the rows of the model matrix $\mathbb{X}$ (with excluded intercept).

Further, let $\widetilde{\mathbb{X}}$ be the non-intercept part of the model matrix $\mathbb{X}$ with all columns being *centered*, i.e.,

$$\widetilde{\mathbb{X}} = \big(\boldsymbol{x}^1 - \overline{x}^1\mathbf{1}_n, \quad \ldots, \quad \boldsymbol{x}^{k-1} - \overline{x}^{k-1}\mathbf{1}_n\big) = \begin{pmatrix} x_{1,1} - \overline{x}^1 & \ldots & x_{1,k-1} - \overline{x}^{k-1} \\ \vdots & \vdots & \vdots \\ x_{n,1} - \overline{x}^1 & \ldots & x_{n,k-1} - \overline{x}^{k-1} \end{pmatrix}.$$

Clearly, $\mathcal{M}\big(\mathbb{X}\big) = \mathcal{M}\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)$. Hence the hat matrix $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top$ can also be calculated using the matrix $\big(\mathbf{1},\, \widetilde{\mathbb{X}}\big)$, where we can use additional property $\mathbf{1}_n^\top\widetilde{\mathbb{X}} = \mathbf{0}_{k-1}^\top$:

$$\mathbb{H} = \big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\left\{\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)^\top\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\right\}^{-}\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)^\top$$

$$= \big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\begin{pmatrix} \underbrace{\mathbf{1}_n^\top\mathbf{1}_n}_{n} & \underbrace{\mathbf{1}_n^\top\widetilde{\mathbb{X}}}_{\mathbf{0}_{k-1}^\top} \\ \underbrace{\widetilde{\mathbb{X}}^\top\mathbf{1}_n}_{\mathbf{0}_{k-1}} & \widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}} \end{pmatrix}^{-}\begin{pmatrix} \mathbf{1}_n^\top \\ \widetilde{\mathbb{X}}^\top \end{pmatrix}$$

[5] *vzdálená pozorování*

$$= \left(\mathbf{1}_n,\ \widetilde{\mathbb{X}}\right) \begin{pmatrix} \frac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \left(\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}}\right)^- \end{pmatrix} \begin{pmatrix} \mathbf{1}_n^\top \\ \widetilde{\mathbb{X}}^\top \end{pmatrix}$$

$$= \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top + \widetilde{\mathbb{X}}\left(\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}}\right)^- \widetilde{\mathbb{X}}^\top.$$

That is, the tth leverage equals

$$h_{t,t} \ = \ \frac{1}{n} + \left(x_{t,1} - \overline{x}^1,\ \ldots,\ x_{t,k-1} - \overline{x}^{k-1}\right)\left(\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}}\right)^-\left(x_{t,1} - \overline{x}^1,\ \ldots,\ x_{t,k-1} - \overline{x}^{k-1}\right)^\top.$$

The second term is then a square of the generalized distance between the non-intercept regressors $\left(x_{t,1},\ \ldots,\ x_{t,k-1}\right)^\top$ of the tth observation and the vector of mean regressors $\overline{\boldsymbol{x}}$. Hence the observations with a high value of the leverage $h_{t,t}$ are observations with the regressor values being far from the mean regressor values and in this sense have unusual regressor (x) values.

High value of a leverage

To evaluate which values of the leverage are high enough to call a particular observation as a leverage point, let us remind an expression of the hat matrix using the orthonormal basis $\mathbb{Q}$ of the regression space $\mathcal{M}(\mathbb{X})$, which is a vector space of dimension $r = \mathsf{rank}\mathbb{X}$. We know that $\mathbb{H} = \mathbb{Q}\mathbb{Q}^\top$ and hence

$$\sum_{i=1}^{n} h_{i,i} = \mathsf{tr}(\mathbb{H}) = \mathsf{tr}\left(\mathbb{Q}\mathbb{Q}^\top\right) = \mathsf{tr}\left(\mathbb{Q}^\top\mathbb{Q}\right) = \mathsf{tr}(\mathbf{I}_r) = r.$$

That is,

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{i,i} = \frac{r}{n}. \tag{14.2}$$

Several *rules of thumbs* can be found in the literature and software implementations concerning a lower bound for the leverage to call a particular observation as a leverage point. Owing to (14.2), a reasonable bound is a value higher than $\frac{r}{n}$. For example, the R function `influence.measures` marks the tth observation as a leverage point if

$$h_{t,t} > \frac{3\,r}{n}.$$

Influence of leverage points

The fact that the leverage points may constitute a problem for the least squares based statistical inference in a linear model comes from remembering an expression for the variance (conditional given the covariate values) of the residuals of a linear model:

$$\mathsf{var}\left(U_t \,\middle|\, \mathbb{X}\right) = \sigma^2\, m_{t,t} = \sigma^2\,(1 - h_{t,t}), \qquad t = 1, \ldots, n.$$

Remind that $U_t = Y_t - \widehat{Y}_t$ and hence also

$$\mathsf{var}\left(Y_t - \widehat{Y}_t \,\middle|\, \mathbb{X}\right) = \sigma^2\,(1 - h_{t,t}), \qquad t = 1, \ldots, n.$$

That is, $\mathsf{var}\left(U_t \,\middle|\, \mathbb{X}\right) = \mathsf{var}\left(Y_t - \widehat{Y}_t \,\middle|\, \mathbb{X}\right)$ is *low* for observations with a high leverage. In other words, the fitted values of high leverage observations are forced to be closer to the observed response values than those of low leverage observations. In this way, the high leverage observations have a higher impact on the fitted regression function than the low leverage observations.

14.4 Influential diagnostics

Both outliers and leverage points do not necessarily constitute a problem. This occurs if they too much influence statistical inference of primary interest. Also other observations (neither outliers nor leverage points) may harmfully influence the statistical inference. In this section, several methods of quantifying the influence of a particular, tth ($t = 1, \ldots, n$) observation on statistical inference will be introduced. In all cases, we will compare a quantity of primary interest based on the model at hand, i.e.,

$$\mathsf{M}\colon\ \boldsymbol{Y} \,\big|\, \mathbb{X} \ \sim\ \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big), \qquad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r,$$

and the quantity based on the leave-one-out model

$$\mathsf{M}_{(-t)}\colon\ \boldsymbol{Y}_{(-t)} \,\big|\, \mathbb{X}_{(-t)} \ \sim\ \big(\mathbb{X}_{(-t)}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_{n-1}\big).$$

It will overally be assumed, that $m_{t,t} > 0$ which implies (see Theorem 14.1) $\mathsf{rank}\big(\mathbb{X}_{(-t)}\big) = \mathsf{rank}\big(\mathbb{X}\big) = r$.

14.4.1 DFBETAS

Let $r = k$, i.e., both models M and $\mathsf{M}_{(-t)}$ are full-rank models. The LSE's of the vector of regression coefficients based on the two models are

$$\begin{aligned}
\mathsf{M}\colon \qquad \widehat{\boldsymbol{\beta}} &= \big(\widehat{\beta}_0,\, \ldots,\, \widehat{\beta}_{k-1}\big)^\top &&= \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y},\\
\mathsf{M}_{(-t)}\colon \qquad \widehat{\boldsymbol{\beta}}_{(-t)} &= \big(\widehat{\beta}_{(-t),0},\, \ldots,\, \widehat{\beta}_{(-t),k-1}\big)^\top &&= \big(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\big)^{-1}\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}.
\end{aligned}$$

Using (14.1):

$$\widehat{\boldsymbol{\beta}} \ -\ \widehat{\boldsymbol{\beta}}_{(-t)} \ =\ \frac{U_t}{m_{t,t}}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{x}_t, \tag{14.3}$$

which quantifies influence of the tth observation on the LSE of the regression coefficients. In the following, let $\boldsymbol{v}_0 = \big(v_{0,0},\, \ldots,\, v_{0,k-1}\big)^\top,\ \ldots,\ \boldsymbol{v}_{k-1} = \big(v_{k-1,0},\, \ldots,\, v_{k-1,k-1}\big)^\top$ be the rows of the matrix $\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$, i.e.,

$$\big(\mathbb{X}^\top\mathbb{X}\big)^{-1} = \begin{pmatrix} \boldsymbol{v}_0^\top \\ \vdots \\ \boldsymbol{v}_{k-1}^\top \end{pmatrix} = \begin{pmatrix} v_{0,0} & \cdots & v_{0,k-1} \\ \vdots & \vdots & \vdots \\ v_{k-1,0} & \cdots & v_{k-1,k-1} \end{pmatrix}.$$

Expression (14.3) written elementwise lead to a quantities called **DFBETA**:

$$\mathrm{DFBETA}_{t,j} \ :=\ \widehat{\beta}_j \ -\ \widehat{\beta}_{(-t),j} \ =\ \frac{U_t}{m_{t,t}}\,\boldsymbol{v}_t^\top\boldsymbol{x}_t, \qquad t = 1, \ldots, n,\ j = 0, \ldots, k-1.$$

Note that $\mathrm{DFBETA}_{t,j}$ has a scale of the jth regressor. To get a dimensionless quantity, we can divide it by the standard error of either $\widehat{\beta}_j$ or $\widehat{\beta}_{(-t),j}$. We have

$$\mathsf{S.E.}\big(\widehat{\beta}_j\big) \ =\ \sqrt{\mathsf{MS}_e\, v_{j,j}}, \qquad \mathsf{S.E.}\big(\widehat{\beta}_{(-t),j}\big) \ =\ \sqrt{\mathsf{MS}_{e,(-t)}\, v_{(-t),j,j}},$$

where $v_{(-t),j,j}$ is the jth diagonal element of matrix $\big(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\big)^{-1}$. In practice, a combined quantity, namely $\sqrt{\mathsf{MS}_{e,(-t)}\, v_{j,j}}$ is used leading to so called **DFBETAS** (the last "S" stands for

"scaled"):

$$\text{DFBETAS}_{t,j} := \frac{\widehat{\beta}_j - \widehat{\beta}_{(-t),j}}{\sqrt{\text{MS}_{e,(-t)}\, v_{j,j}}} = \frac{U_t}{m_{t,t}\sqrt{\text{MS}_{e,(-t)}\, v_{j,j}}}\, \boldsymbol{v}_t^\top \boldsymbol{x}_t,$$
$$t = 1, \ldots, n,\ j = 0, \ldots, k-1.$$

The reason for using $\sqrt{\text{MS}_{e,(-t)}\, v_{j,j}}$ as a scale factor is that $\text{MS}_{e,(-t)}$ is a safer estimator of the residual variance σ^2 not being based on the observation whose influence is examined but at the same time, it can still be calculated from quantities of the full model M (see Eq. 14.1). On the other hand, a value of $v_{(-t),j,j}$ (that fits with the leave-one-out residual mean square $\text{MS}_{e,(-t)}$) cannot, in general, be calculated from quantities of the full model M and hence (a close) value of $v_{j,j}$ is used. Consequently, all values of DFBETAS can be calculated from quantities of the full model M and there is no need to fit n leave-one-out models.

***Note** (Rule-of-thumb used by* R*).*

The R function `influence.measures` marks the tth observation as being influential with respect to the LSE of the jth regression coefficient if

$$\left|\text{DFBETAS}_{t,j}\right| > 1.$$

14.4.2 DFFITS

We are assuming $m_{t,t} > 0$ and hence by Theorem 14.1, parameter $\mu_t := \mathbb{E}\left(Y_t \,\middle|\, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$ is estimable in both models M and $\text{M}_{(-t)}$. Let as usual, $\boldsymbol{b} = \left(\mathbb{X}^\top\mathbb{X}\right)^-\mathbb{X}^\top\boldsymbol{Y}$ be any solution to normal equations in model M (which is now not necessarily of a full-rank) and let $\boldsymbol{b}_{(-t)} = \left(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\right)^-\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}$ be any solution to normal equations in the leave-one-out model $\text{M}_{(-t)}$. The LSE's of μ_t in the two models are

$$\text{M}: \quad \widehat{Y}_t = \boldsymbol{x}_t^\top \boldsymbol{b},$$
$$\text{M}_{(-t)}: \quad \widehat{Y}_{[t]} = \boldsymbol{x}_t^\top \boldsymbol{b}_{(-t)}.$$

Using (14.1):

$$\widehat{Y}_{[t]} = \boldsymbol{x}_t^\top\left\{\boldsymbol{b} - \frac{U_t}{m_{t,t}}\left(\mathbb{X}^\top\mathbb{X}\right)^-\boldsymbol{x}_t\right\} = \widehat{Y}_t - \frac{U_t}{m_{t,t}}\boldsymbol{x}_t^\top\left(\mathbb{X}^\top\mathbb{X}\right)^-\boldsymbol{x}_t = \widehat{Y}_t - U_t\,\frac{h_{t,t}}{m_{t,t}}.$$

Difference between $\widehat{Y}_t$ and $\widehat{Y}_{[t]}$ is called **DFFIT** and quantifies influence of the tth observation on the LSE of its own expectation:

$$\text{DFFIT}_t := \widehat{Y}_t - \widehat{Y}_{[t]} = U_t\,\frac{h_{t,t}}{m_{t,t}}, \qquad t = 1, \ldots, n.$$

Analogously to DFBETAS, also DFFIT is scaled by a quantity that resembles the standard error of either $\widehat{Y}_t$ or $\widehat{Y}_{[t]}$ (remember, $\text{S.E.}\left(\widehat{Y}_t\right) = \sqrt{\text{MS}_e\, h_{t,t}}$) leading to a quantity called **DFFITS**:

$$\text{DFFITS}_t := \frac{\widehat{Y}_t - \widehat{Y}_{[t]}}{\sqrt{\text{MS}_{e,(-t)}\, h_{t,t}}}$$
$$= \frac{h_{t,t}}{m_{t,t}}\,\frac{U_t}{\sqrt{\text{MS}_{e,(-t)}\, h_{t,t}}} = \sqrt{\frac{h_{t,t}}{m_{t,t}}}\,\frac{U_t}{\sqrt{\text{MS}_{e,(-t)}\, m_{t,t}}} = \sqrt{\frac{h_{t,t}}{m_{t,t}}}\, T_t, \qquad t = 1, \ldots, n,$$

where T_t is the tth studentized residual of the model M. Again, all values of DFFITS can be calculated from quantities of the full model M and there is no need to fit n leave-one-out models.

***Note** (Rule-of-thumb used by R).*

The R function `influence.measures` marks the tth observation as excessively influencing the LSE of its expectation if

$$\left|\text{DFFITS}_t\right| > 3\sqrt{\frac{r}{n-r}}.$$

14.4.3 Cook distance

In this Section, we concentrate on evaluation of the influence of the tth observation on the LSE of a vector parameter $\boldsymbol{\mu} := \mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{X}) = \mathbb{X}\boldsymbol{\beta}$. As in Section 14.4.2, let $\boldsymbol{b} = (\mathbb{X}^\top\mathbb{X})^-\mathbb{X}^\top\boldsymbol{Y}$ be any solution to normal equations in model M and let $\boldsymbol{b}_{(-t)} = (\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)})^-\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}$ be any solution to normal equations in the leave-one-out model $\mathsf{M}_{(-t)}$. The LSE's of $\boldsymbol{\mu}$ in the two models are

$$\begin{aligned}
&\mathsf{M}: && \widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b} \quad = \mathbb{H}\boldsymbol{Y},\\
&\mathsf{M}_{(-t)}: && \widehat{\boldsymbol{Y}}_{(-t\bullet)} := \mathbb{X}\boldsymbol{b}_{(-t)}.
\end{aligned}$$

Note. Remind that $\widehat{\boldsymbol{Y}}_{(-t\bullet)}$, $\widehat{\boldsymbol{Y}}_{[\bullet]}$ and $\widehat{\boldsymbol{Y}}_{(-t)}$ are three different quantities. Namely,

$$\widehat{\boldsymbol{Y}}_{(-t\bullet)} = \mathbb{X}\boldsymbol{b}_{(-t)} = \begin{pmatrix} \boldsymbol{x}_1^\top\boldsymbol{b}_{(-t)} \\ \vdots \\ \boldsymbol{x}_n^\top\boldsymbol{b}_{(-t)} \end{pmatrix}, \qquad \widehat{\boldsymbol{Y}}_{[\bullet]} = \begin{pmatrix} \widehat{Y}_{[1]} \\ \vdots \\ \widehat{Y}_{[n]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top\boldsymbol{b}_{(-1)} \\ \vdots \\ \boldsymbol{x}_n^\top\boldsymbol{b}_{(-n)} \end{pmatrix}.$$

Finally, $\widehat{\boldsymbol{Y}}_{(-t)} = \mathbb{X}_{(-t)}\boldsymbol{b}_{(-t)}$ is a subvector of length $n-1$ of a vector $\widehat{\boldsymbol{Y}}_{(-t\bullet)}$ of length n.

Possible quantification of influence of the tth observation on the LSE of a vector parameter $\boldsymbol{\mu}$ is obtained by considering a quantity

$$\left\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)}\right\|^2.$$

Let us remind from Lemma 14.3:

$$\boldsymbol{b} - \boldsymbol{b}_{(-t)} = \frac{U_t}{m_{t,t}}(\mathbb{X}^\top\mathbb{X})^-\boldsymbol{x}_t.$$

Hence,

$$\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)} = \mathbb{X}\left(\boldsymbol{b} - \boldsymbol{b}_{(-t)}\right) = \frac{U_t}{m_{t,t}}\mathbb{X}(\mathbb{X}^\top\mathbb{X})^-\boldsymbol{x}_t.$$

Then

$$\begin{aligned}
\left\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)}\right\|^2 &= \left\|\frac{U_t}{m_{t,t}}\mathbb{X}(\mathbb{X}^\top\mathbb{X})^-\boldsymbol{x}_t\right\|^2 \\
&= \frac{U_t^2}{m_{t,t}^2}\boldsymbol{x}_t^\top(\mathbb{X}^\top\mathbb{X})^-\mathbb{X}^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X})^-\boldsymbol{x}_t \\
&= \frac{U_t^2}{m_{t,t}^2}h_{t,t}. \qquad (14.4)
\end{aligned}$$

The equality (14.4) follows from noting that

(a) $\boldsymbol{x}_t^\top \big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\boldsymbol{x}_t$ is the tth diagonal element of matrix $\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top$;

(b) $\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top = \mathbb{H}$ by the five matrices rule (Theorem A.2).

The so called **Cook distance** of the tth observation is (14.4) modified to get a unit-free quantity. Namely, the Cook distance is defined as

$$D_t \;:=\; \frac{1}{r\,\mathsf{MS}_e}\big\|\widehat{\boldsymbol{Y}} \;-\; \widehat{\boldsymbol{Y}}_{(-t\bullet)}\big\|^2.$$

Expression (14.4) shows that it is again not necessary to fit the leave-one-out model to calculate the Cook distance. Moreover, we can express it as follows

$$D_t \;=\; \frac{1}{r}\,\frac{h_{t,t}}{m_{t,t}}\,\frac{U_t^2}{\mathsf{MS}_e\,m_{t,t}} \;=\; \frac{1}{r}\,\frac{h_{t,t}}{m_{t,t}}\,\big(U_t^{std}\big)^2.$$

Notes.

- We are assuming $m_{t,t} > 0$. Hence $h_{t,t} = 1 - m_{t,t} \in (0,\,1)$ and the term $h_{t,t}/m_{t,t}$ increases with the leverage $h_{t,t}$ (having a limit of ∞ with $h_{t,t} \to 1$). The "$h_{t,t}/m_{t,t}$" part of the Cook distance thus quantifies how much is the tth observation the leverage point.
- The "U_t^{std}" part of the Cook distance increases with the distance between the observed and fitted value which is high for outliers.
- The Cook distance is thus a combined measure being high for observations which are either leverage points or outliers or both.

Cook distance in a full-rank model

If $r = k$ and both M and $\mathsf{M}_{(-t)}$ are of full-rank, we have

$$\boldsymbol{b} \;=\; \widehat{\boldsymbol{\beta}} \;=\; \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y},$$

$$\boldsymbol{b}_{(-t)} \;=\; \widehat{\boldsymbol{\beta}}_{(-t)} \;=\; \big(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\big)^{-1}\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}.$$

Then, directly from definition,

$$\big\|\widehat{\boldsymbol{Y}} \;-\; \widehat{\boldsymbol{Y}}_{(-t\bullet)}\big\|^2 \;=\; \big\|\mathbb{X}\widehat{\boldsymbol{\beta}} \;-\; \mathbb{X}\widehat{\boldsymbol{\beta}}_{(-t)}\big\|^2 \;=\; \big(\widehat{\boldsymbol{\beta}}_{(-t)} \;-\; \widehat{\boldsymbol{\beta}}\big)^\top\mathbb{X}^\top\mathbb{X}\big(\widehat{\boldsymbol{\beta}}_{(-t)} \;-\; \widehat{\boldsymbol{\beta}}\big).$$

The Cook distance is then

$$D_t \;=\; \frac{\big(\widehat{\boldsymbol{\beta}}_{(-t)} \;-\; \widehat{\boldsymbol{\beta}}\big)^\top\mathbb{X}^\top\mathbb{X}\big(\widehat{\boldsymbol{\beta}}_{(-t)} \;-\; \widehat{\boldsymbol{\beta}}\big)}{k\,\mathsf{MS}_e},$$

which is a distance between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(-t)}$ in a certain metric.

Remember now that under normality, the confidence region for parameter $\boldsymbol{\beta}$ with a coverage of $1-\alpha$, derived while assuming model M is

$$\mathcal{C}(\alpha) = \big\{\boldsymbol{\beta}:\; \big(\boldsymbol{\beta} \;-\; \widehat{\boldsymbol{\beta}}\big)^\top\mathbb{X}^\top\mathbb{X}\big(\boldsymbol{\beta} \;-\; \widehat{\boldsymbol{\beta}}\big) \;<\; k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)\big\}.$$

That is

$$\widehat{\boldsymbol{\beta}}_{(-t)} \in \mathcal{C}(\alpha) \qquad \text{if and only if} \qquad D_t < \mathcal{F}_{k,n-k}(1-\alpha). \tag{14.5}$$

This motivates the following rule-of-thumb.

***Note** (Rule-of-thumb used by R).*

The R function `influence.measures` marks the tth observation as excessively influencing the LSE of the full response expectation $\boldsymbol{\mu}$ if

$$D_t \; > \; \mathcal{F}_{r,n-r}(0.50).$$

14.4.4 COVRATIO

In this Section, we will again assume full-rank models ($r = k$) and explore influence of the tth observation on precision of the LSE of the vector of regression coefficients. The LSE's of the vector of regression coefficients based on the two models are

$$\begin{aligned} \mathsf{M}: \qquad \widehat{\boldsymbol{\beta}} &= \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{Y}, \\ \mathsf{M}_{(-t)}: \qquad \widehat{\boldsymbol{\beta}}_{(-t)} &= \left(\mathbb{X}_{(-t)}{}^\top \mathbb{X}_{(-t)}\right)^{-1} \mathbb{X}_{(-t)}{}^\top \boldsymbol{Y}_{(-t)}. \end{aligned}$$

The estimated covariance matrices of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(-t)}$, respectively, are

$$\begin{aligned} \widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}} \,\middle|\, \mathbb{X}\right) &= \mathsf{MS}_e \left(\mathbb{X}^\top \mathbb{X}\right)^{-1}, \\ \widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}}_{(-t)} \,\middle|\, \mathbb{X}\right) &= \mathsf{MS}_{e,(-t)} \left(\mathbb{X}_{(-t)}^\top \mathbb{X}_{(-t)}\right)^{-1}. \end{aligned}$$

Influence of the tth observation on the precision of the LSE of the vector of regression coefficients is quantified by so called **COVRATIO** being defined as

$$\text{COVRATIO}_t \;=\; \frac{\det\left\{\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}}_{(-t)} \,\middle|\, \mathbb{X}\right)\right\}}{\det\left\{\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}} \,\middle|\, \mathbb{X}\right)\right\}}, \qquad t = 1, \ldots, n.$$

After some calculation (see below), it can be shown that

$$\text{COVRATIO}_t \;=\; \frac{1}{m_{t,t}} \left\{ \frac{n-k-\left(U_t^{std}\right)^2}{n-k-1} \right\}^k, \qquad t = 1, \ldots, n.$$

That is, it is again not necessary to fit n leave-one-out models to calculate the COVARTIO values for all observations in the dataset.

***Note** (Rule-of-thumb used by R).*

The R function `influence.measures` marks the tth observation as excessively influencing precision of the estimation of the regression coefficients if

$$\left|1 - \text{COVRATIO}_t\right| \; > \; 3\,\frac{k}{n-k}.$$

Calculation towards COVRATIO

First, remind a matrix identity (e.g., Anděl, 2007, Theorem A.4): If $\mathbb{A}$ and $\mathbb{D}$ are square invertible matrices then

$$\begin{vmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{C} & \mathbb{D} \end{vmatrix} = \left|\mathbb{A}\right| \cdot \left|\mathbb{D} - \mathbb{C}\mathbb{A}^{-1}\mathbb{B}\right| = \left|\mathbb{D}\right| \cdot \left|\mathbb{A} - \mathbb{B}\mathbb{D}^{-1}\mathbb{C}\right|.$$

Use twice the above identity:

$$\begin{vmatrix} \mathbb{X}^\top\mathbb{X} & \boldsymbol{x}_t \\ \boldsymbol{x}_t^\top & 1 \end{vmatrix} = \left|\mathbb{X}^\top\mathbb{X}\right| \cdot \underbrace{\left|1 - \boldsymbol{x}_t^\top\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\boldsymbol{x}_t\right|}_{1-h_{t,t}=m_{t,t}} = \left|\mathbb{X}^\top\mathbb{X}\right| m_{t,t},$$

$$= |1| \cdot \left|\mathbb{X}^\top\mathbb{X} - \boldsymbol{x}_t\boldsymbol{x}_t^\top\right| = \left|\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\right|.$$

So that, $\quad m_{t,t}\left|\mathbb{X}^\top\mathbb{X}\right| = \left|\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\right|.$

Then,

$$\frac{\det\left\{\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}}_{(-t)} \,|\, \mathbb{X}\right)\right\}}{\det\left\{\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}} \,|\, \mathbb{X}\right)\right\}} = \frac{\left|\mathsf{MS}_{e,(-t)}\left(\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\right)^{-1}\right|}{\left|\mathsf{MS}_e\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right|}$$

$$= \left(\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e}\right)^k \cdot \frac{\left|\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\right|^{-1}}{\left|\mathbb{X}^\top\mathbb{X}\right|^{-1}} = \left(\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e}\right)^k \cdot \frac{1}{m_{t,t}}.$$

Expression (14.1):

$$\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e} = \frac{n-k-\left(U_t^{std}\right)^2}{n-k-1}.$$

Hence, $\quad \dfrac{\det\left\{\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}}_{(-t)} \,|\, \mathbb{X}\right)\right\}}{\det\left\{\widehat{\mathsf{var}}\left(\widehat{\boldsymbol{\beta}} \,|\, \mathbb{X}\right)\right\}} = \dfrac{1}{m_{t,t}}\left(\dfrac{n-k-\left(U_t^{std}\right)^2}{n-k-1}\right)^k.$

14.4.5 Final remarks

- All presented influence measures should be used sensibly.
- Depending on what is the purpose of the modelling, different types of influence are differently harmful.
- There is certainly no need to panic if some observations are marked as “influential”!

Appendix A

Matrices

A.1 Pseudoinverse of a matrix

Definition A.1 Pseudoinverse of a matrix.

The pseudoinverse *of a real matrix $\mathbb{A}_{n\times k}$ is such a matrix $\mathbb{A}^{-}$ of dimension $k \times n$ that satisfies*

$$\mathbb{A}\mathbb{A}^{-}\mathbb{A} = \mathbb{A}.$$

Notes.

- The pseudoinverse always exists. Nevertheless, it is not necessarily unique.
- If $\mathbb{A}$ is invertible then $\mathbb{A}^{-} = \mathbb{A}^{-1}$ is the only pseudoinverse.

Definition A.2 Moore-Penrose pseudoinverse of a matrix.

The Moore-Penrose pseudoinverse *of a real matrix $\mathbb{A}_{n\times k}$ is such a matrix $\mathbb{A}^{+}$ of dimension $k \times n$ that satisfies the following conditions:*

(i) $\mathbb{A}\mathbb{A}^{+}\mathbb{A} = \mathbb{A}$;

(ii) $\mathbb{A}^{+}\mathbb{A}\mathbb{A}^{+} = \mathbb{A}^{+}$;

(iii) $\left(\mathbb{A}\mathbb{A}^{+}\right)^{\top} = \mathbb{A}\mathbb{A}^{+}$;

(iv) $\left(\mathbb{A}^{+}\mathbb{A}\right)^{\top} = \mathbb{A}^{+}\mathbb{A}$.

Notes.

- The Moore-Penrose pseudoinverse always exists and it is unique.
- The Moore-Penrose pseudoinverse can be calculated from the *singular value decomposition (SVD)* of the matrix $\mathbb{A}$.

Theorem A.1 Pseudoinverse of a matrix and a solution of a linear system.

Let $\mathbb{A}_{n\times k}$ be a real matrix and let $\boldsymbol{c}_{n\times 1}$ be a real vector. Let there exist a solution of a linear system $\mathbb{A}\boldsymbol{x} = \boldsymbol{c}$, i.e., the linear system $\mathbb{A}\boldsymbol{x} = \boldsymbol{c}$ is consistent. *Let $\mathbb{A}^-$ be the pseudoinverse of $\mathbb{A}$.*

A vector $\boldsymbol{x}_{k\times 1}$ solves the linear system $\mathbb{A}\boldsymbol{x} = \boldsymbol{c}$ if and only if

$$\boldsymbol{x} = \mathbb{A}^-\boldsymbol{c}.$$

Proof. See Anděl (2007, Appendix A.4). ❑

Theorem A.2 Five matrices rule.

For a real matrix $\mathbb{A}_{n\times k}$, it holds

$$\mathbb{A}\big(\mathbb{A}^\top\mathbb{A}\big)^-\mathbb{A}^\top\mathbb{A} \;=\; \mathbb{A}.$$

That is, a matrix $\big(\mathbb{A}^\top\mathbb{A}\big)^-\mathbb{A}^\top$ is a pseudoinverse of a matrix $\mathbb{A}$.

Proof. See Anděl (2007, Theorem A.19). ❑

A.2 Kronecker product

Definition A.3 Kronecker product.

Let $\mathbb{A}_{m\times n}$ and $\mathbb{C}_{p\times q}$ be real matrices. Their Kronecker product $\mathbb{A}\otimes\mathbb{C}$ *is a matrix $\mathbb{D}_{m\cdot p\times n\cdot q}$ such that*

$$\mathbb{D}=\mathbb{A}\otimes\mathbb{C}=\begin{pmatrix} a_{1,1}\mathbb{C} & \dots & a_{1,s}\mathbb{C} \\ \vdots & \vdots & \vdots \\ a_{r,1}\mathbb{C} & \dots & a_{r,s}\mathbb{C} \end{pmatrix}=\big(a_{i,j}\mathbb{C}\big)_{i=1,\dots,m,j=1,\dots,n}.$$

Note. For $\boldsymbol{a}\in\mathbb{R}^m$, $\boldsymbol{b}\in\mathbb{R}^p$, we can write

$$\boldsymbol{a}\,\boldsymbol{b}^\top=\boldsymbol{a}\otimes\boldsymbol{b}^\top.$$

Theorem A.3 Properties of a Kronecker product.

It holds for the Kronecker product:

(i) $\mathbf{0}\otimes\mathbb{A}=\mathbf{0}$, $\mathbb{A}\otimes\mathbf{0}=\mathbf{0}$.

(ii) $(\mathbb{A}_1+\mathbb{A}_2)\otimes\mathbb{C}=(\mathbb{A}_1\otimes\mathbb{C})+(\mathbb{A}_2\otimes\mathbb{C})$.

(iii) $\mathbb{A}\otimes(\mathbb{C}_1+\mathbb{C}_2)=(\mathbb{A}\otimes\mathbb{C}_1)+(\mathbb{A}\otimes\mathbb{C}_2)$.

(iv) $a\mathbb{A}\otimes c\mathbb{C}=a\,c\,(A\otimes\mathbb{C})$.

(v) $\mathbb{A}_1\mathbb{A}_2\otimes\mathbb{C}_1\mathbb{C}_2=(\mathbb{A}_1\otimes\mathbb{C}_1)(\mathbb{A}_2\otimes\mathbb{C}_2)$.

(vi) $\big(\mathbb{A}\otimes\mathbb{C}\big)^{-1}=\mathbb{A}^{-1}\otimes\mathbb{C}^{-1}$, *if the inversions exist.*

(vii) $\big(\mathbb{A}\otimes\mathbb{C}\big)^{-}=\mathbb{A}^{-}\otimes\mathbb{C}^{-}$, *for arbitrary pseudoinversions.*

(viii) $\big(\mathbb{A}\otimes\mathbb{C}\big)^{\top}=\mathbb{A}^{\top}\otimes\mathbb{C}^{\top}$.

(ix) $\big(\mathbb{A},\ \mathbb{C}\big)\otimes\mathbb{D}=\big(\mathbb{A}\otimes\mathbb{D},\ \ \mathbb{C}\otimes\mathbb{D}\big)$.

(x) Upon a suitable reordering of the columns, matrices $\big(\mathbb{A}\otimes\mathbb{C},\ \ \mathbb{A}\otimes\mathbb{D}\big)$ *and* $\mathbb{A}\otimes\big(\mathbb{C},\ \ \mathbb{D}\big)$ *are the same.*

(xi) $\mathsf{rank}\big(\mathbb{A}\otimes\mathbb{C}\big)=\mathsf{rank}\big(\mathbb{A}\big)\,\mathsf{rank}\big(\mathbb{C}\big)$.

Proof. See Rao (1973, Section 1b.8). ❑

Definition A.4 Elementwise product of two vectors.

Let $\boldsymbol{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, $\boldsymbol{c} = (b_1, \ldots, b_p)^\top \in \mathbb{R}^p$. *Their* elementwise product[1] *is a vector* $(a_1\, c_1,\ \ldots,\ a_p\, c_p)^\top$ *that will be denoted as* $\boldsymbol{a} : \boldsymbol{c}$. *That is,*

$$\boldsymbol{a} : \boldsymbol{c} = \begin{pmatrix} a_1\, c_1 \\ \vdots \\ a_p\, c_p \end{pmatrix}.$$

Definition A.5 Columnwise product of two matrices.

Let

$$\mathbb{A}_{n\times p} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^p) \quad \textit{and}\ \mathbb{C}_{n\times q} = (\boldsymbol{c}^1, \ldots, \boldsymbol{c}^q)$$

be real matrices. Their columnwise product[2] $\mathbb{A} : \mathbb{C}$ *is a matrix* $\mathbb{D}_{n\times p\cdot q}$ *such that*

$$\mathbb{D} = \mathbb{A} : \mathbb{C} = (\boldsymbol{a}^1 : \boldsymbol{c}^1, \ldots, \boldsymbol{a}^p : \boldsymbol{c}^1, \ldots, \boldsymbol{a}^1 : \boldsymbol{c}^q, \ldots, \boldsymbol{a}^p : \boldsymbol{c}^q).$$

Notes.

- If we write

$$\mathbb{A} = \begin{pmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_n^\top \end{pmatrix}, \qquad \mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_n^\top \end{pmatrix},$$

  the columnwise product of two matrices can also be written as a matrix rows of which are obtained as Kronecker products of the rows of the two matrices:

$$\mathbb{A} : \mathbb{C} = \begin{pmatrix} \boldsymbol{c}_1^\top \otimes \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{c}_n^\top \otimes \boldsymbol{a}_n^\top \end{pmatrix}. \tag{A.1}$$

- It perhaps looks more logical to define the columnwise product of two matrices as

$$\mathbb{A} : \mathbb{C} = \begin{pmatrix} \boldsymbol{a}_1^\top \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{a}_n^\top \otimes \boldsymbol{c}_n^\top \end{pmatrix} = (\boldsymbol{a}^1 : \boldsymbol{c}^1, \ldots, \boldsymbol{a}^1 : \boldsymbol{c}^q, \ldots, \boldsymbol{a}^p : \boldsymbol{c}^1, \ldots, \boldsymbol{a}^p : \boldsymbol{c}^q),$$

  which only differs by ordering of the columns of the resulting matrix. Our definition (A.1) is motivated by the way in which an operator : acts in the R software.

[1] *součin po složkách* [2] *součin po sloupcích*

A.3 Additional theorems on matrices

Theorem A.4 Inverse of a matrix divided into blocks.

Let

$$\mathbb{M} = \begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B}^\top & \mathbb{D} \end{pmatrix}$$

be a positive definite *matrix divided in blocks* $\mathbb{A}$, $\mathbb{B}$, $\mathbb{D}$.

Then the following holds:

(i) Matrix $\mathbb{Q} = \mathbb{A} - \mathbb{B}\mathbb{D}^{-1}\mathbb{B}^\top$ *is positive definite.*

(ii) Matrix $\mathbb{P} = \mathbb{D} - \mathbb{B}^\top\mathbb{A}^{-1}\mathbb{B}$ *is positive definite.*

(iii) The inverse to $\mathbb{M}$ *is*

$$\mathbb{M}^{-1} = \begin{pmatrix} \mathbb{Q}^{-1} & -\mathbb{Q}^{-1}\mathbb{B}\mathbb{D}^{-1} \\ -\mathbb{D}^{-1}\mathbb{B}^\top\mathbb{Q}^{-1} & \mathbb{D}^{-1} + \mathbb{D}^{-1}\mathbb{B}^\top\mathbb{Q}^{-1}\mathbb{B}\mathbb{D}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbb{A}^{-1} + \mathbb{A}^{-1}\mathbb{B}\mathbb{P}^{-1}\mathbb{B}^\top\mathbb{A}^{-1} & -\mathbb{A}^{-1}\mathbb{B}\mathbb{P}^{-1} \\ -\mathbb{P}^{-1}\mathbb{B}^\top\mathbb{A}^{-1} & \mathbb{P}^{-1} \end{pmatrix}.$$

Proof. See Anděl (2007, Theorem A.10 in Appendix A.2). ❑

Appendix **B**

Distributions

B.1 Non-central univariate distributions

Definition B.1 Non-central Student t-distribution.

Let $U \sim \mathcal{N}(0, 1)$, let $V \sim \chi^2_\nu$ for some $\nu > 0$ and let U and V be independent. Let $\lambda \in \mathbb{R}$. Then we say that a random variable

$$T = \frac{U + \lambda}{\sqrt{\frac{V}{\nu}}}$$

follows a non-central Student t-distribution[1] *with ν degrees of freedom*[2] *and a non-centrality parameter*[3] *λ. We shall write*

$$T \sim \mathsf{t}_\nu(\lambda).$$

Notes.

- Non-central t-distribution is different from simply a shifted (central) t-distribution.
- Directly seen from definition: $\mathsf{t}_\nu(0) \equiv \mathsf{t}_\nu$.
- Moments of a non-central Student t-distribution:

$$\mathbb{E}(T) = \begin{cases} \lambda\sqrt{\frac{\nu}{2}}\,\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, & \text{if } \nu > 1, \\ \text{does not exist}, & \text{if } \nu \leq 1. \end{cases}$$

$$\mathsf{var}(T) = \begin{cases} \frac{\nu(1+\lambda^2)}{\nu-2} - \frac{\nu\lambda^2}{2}\left\{\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right\}^2, & \text{if } \nu > 2, \\ \text{does not exist}, & \text{if } \nu \leq 2. \end{cases}$$

[1] *necentrální Studentovo t-rozdělení* [2] *stupně volnosti* [3] *parametr necentrality*

Definition B.2 Non-central χ^2 distribution.

Let $U_1, \ldots, U_k$ be independent *random variables. Let further $U_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, k$, for some $\mu_1, \ldots, \mu_k \in \mathbb{R}$. That is $\boldsymbol{U} = (U_1, \ldots, U_k)^\top \sim \mathcal{N}_k(\boldsymbol{\mu}, \mathbf{I}_k)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\top$. Then we say that a random variable*

$$X = \sum_{i=1}^{k} U_i^2 = \|\boldsymbol{U}\|^2$$

follows a non-central chi-squared distribution[4] *with k degrees of freedom and a non-centrality parameter*

$$\lambda = \sum_{i=1}^{k} \mu_i^2 = \|\boldsymbol{\mu}\|^2.$$

We shall write

$$X \sim \chi^2_k(\lambda).$$

Notes.

- It can easily be proved that the distribution of the random variable X from Definition B.2 indeed depends only on k and $\lambda = \sum_{i=1}^{k} \mu_i^2$ and not on the particular values of $\mu_1, \ldots, \mu_k$.
- As an exercise for the use of a convolution theorem, we can derive a density of the $\chi^2_k(\lambda)$ distribution which is

$$f(x) = \begin{cases} \dfrac{\mathrm{e}^{-\frac{x+\lambda}{2}}\, x^{\frac{k-2}{2}}}{2^{\frac{k}{2}}\, \Gamma\left(\frac{k-1}{2}\right) \Gamma\left(\frac{1}{2}\right)} \displaystyle\sum_{j=0}^{\infty} \frac{\lambda^j\, x^j}{(2j)!}\, B\left(\frac{k-1}{2}, \frac{1}{2} + j\right), & x > 0, \\ 0, & x \leq 0. \end{cases}$$

- The non-central χ^2 distribution with general degrees of freedom $\nu \in (0, \infty)$ is defined as a distribution with the density given by the above expression with k replaced by ν.
- $\chi^2_\nu(0) \equiv \chi^2_\nu$.
- Moments of a non-central χ^2 distribution:

$$\mathbb{E}(X) = \nu + \lambda,$$

$$\mathsf{var}(X) = 2\,(\nu + 2\lambda).$$

[4] *necentrální chí-kvadrát rozdělení*

Definition B.3 Non-central F-distribution.

Let $X \sim \chi^2_{\nu_1}(\lambda)$*, where* $\nu_1,\ \lambda > 0$*. Let* $Y \sim \chi^2_{\nu_2}$*, where* $\nu_2 > 0$*. Let further* X *and* Y *be* independent. *Then we say that a random variable*

$$Q = \frac{\frac{X}{\nu_1}}{\frac{Y}{\nu_2}}$$

follows a non-central F-distribution[5] *with* ν_1 *and* ν_2 *degrees of freedom and a noncentrality parameter* λ*. We shall write*

$$Q \sim \mathcal{F}_{\nu_1,\nu_2}(\lambda).$$

Notes.

- Directly seen from definition: $\mathcal{F}_{\nu_1,\nu_2}(0) \equiv \mathcal{F}_{\nu_1,\nu_2}$.
- Moments of a non-central F-distribution:

$$\mathbb{E}(Q) = \begin{cases} \dfrac{\nu_2\,(\nu_1+\lambda)}{\nu_1\,(\nu_2-2)}, & \text{if } \nu_2 > 2, \\ \text{does not exist}, & \text{if } \nu_2 \leq 2. \end{cases}$$

$$\mathsf{var}(Q) = \begin{cases} 2\,\dfrac{(\nu_1+\lambda)^2\ +\ (\nu_1+2\lambda)\,(\nu_2-2)}{(\nu_2-2)^2\,(\nu_2-4)}\left(\dfrac{\nu_2}{\nu_1}\right)^2, & \text{if } \nu_2 > 4, \\ \text{does not exist}, & \text{if } \nu_2 \leq 4. \end{cases}$$

[5] *necentrální F-rozdělení*

B.2 Multivariate distributions

Definition B.4 Multivariate Student t-distribution.
Let $\boldsymbol{U} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma}_{p\times p}$ *is a positive semidefinite matrix. Let further* $V \sim \chi^2_\nu$ *for some* $\nu > 0$ *and let* $\boldsymbol{U}$ *and* V *be* independent. *Then we say that a random vector*

$$\boldsymbol{T} = \boldsymbol{U}\sqrt{\frac{\nu}{V}}$$

follows a p*-dimensional* multivariate Student t-distribution[6] *with* ν *degrees of freedom and a scale matrix*[7] $\boldsymbol{\Sigma}$. *We shall write*

$$\boldsymbol{T} \sim \mathsf{mvt}_{p,\nu}(\boldsymbol{\Sigma}).$$

Notes.

- Directly seen from definition: $\mathsf{mvt}_{1,\nu}(1) \equiv \mathsf{t}_\nu$.
- If $\boldsymbol{\Sigma}$ is a regular (positive definite) matrix, then the density (with respect to the p-dimensional Lebesgue measure) of the $\mathsf{mvt}_{p,\nu}(\boldsymbol{\Sigma})$ distribution is

$$f(\boldsymbol{t}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{p}{2}}\pi^{\frac{p}{2}}}\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\left\{1 + \frac{\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}}{\nu}\right\}^{-\frac{\nu+p}{2}}, \qquad \boldsymbol{t} \in \mathbb{R}^p.$$

- Expectation and a covariance matrix of $\boldsymbol{T} \sim \mathsf{mvt}_{p,\nu}(\boldsymbol{\Sigma})$ are

$$\mathbb{E}(\boldsymbol{T}) = \begin{cases} \mathbf{0}_p, & \text{if } \nu > 1, \\ \text{does not exist}, & \text{if } \nu \leq 1. \end{cases}$$

$$\mathsf{var}(\boldsymbol{T}) = \begin{cases} \dfrac{\nu}{\nu-2}\boldsymbol{\Sigma}, & \text{if } \nu > 2, \\ \text{does not exist}, & \text{if } \nu \leq 2. \end{cases}$$

Lemma B.1 Marginals of the multivariate Student t-distribution.
Let $\boldsymbol{T} = \left(T_1, \ldots, T_p\right)^\top \sim \mathsf{mvt}_{p,\nu}(\boldsymbol{\Sigma})$, *where the scale matrix* $\boldsymbol{\Sigma}$ *has positive diagonal elements* $\sigma_1^2 > 0, \ldots, \sigma_p^2 > 0$. *Then*

$$\frac{T_j}{\sigma_j} \sim \mathsf{t}_\nu, \qquad j = 1, \ldots, p.$$

Proof.

- From definition of the multivariate t-distribution, $\boldsymbol{T}$ can be written as $\boldsymbol{T} = \boldsymbol{U}\sqrt{\frac{\nu}{V}}$, where $\boldsymbol{U} = \left(U_1, \ldots, U_p\right)^\top \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$ and $V \sim \chi^2_\nu$ are independent.

[6] *vícerozměrné Studentovo t-rozdělení* [7] *měřítková matice*

- Then for all $j = 1, \ldots, p$:

$$\frac{T_j}{\sigma_j} = \frac{U_j}{\sigma_j}\sqrt{\frac{\nu}{V}} = \frac{Z_j}{\sqrt{\frac{V}{\nu}}},$$

  where $Z_j \sim \mathcal{N}(0,\, 1)$ is independent of $V \sim \chi^2_\nu$.

B.3 Some distributional properties

Lemma B.2 Property of a normal distribution.
Let $\boldsymbol{Z} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. *Let* $T : \mathbb{R}^n \longrightarrow \mathbb{R}$ *be a measurable function satisfying* $T(c\boldsymbol{z}) = T(\boldsymbol{z})$ *for all* $c > 0$ *and* $\boldsymbol{z} \in \mathbb{R}^n$. *The random variables* $T(\boldsymbol{Z})$ *and* $\|\boldsymbol{Z}\|$ *are then independent.*

Proof.

- Consider spherical coordinates:

$$\begin{aligned}
Z_1 &= R\,\cos(\phi_1),\\
Z_2 &= R\,\sin(\phi_1)\,\cos(\phi_2),\\
Z_3 &= R\,\sin(\phi_1)\,\sin(\phi_2)\,\cos(\phi_3),\\
&\vdots\\
Z_{n-1} &= R\,\sin(\phi_1)\,\cdots\,\sin(\phi_{n-2})\,\cos(\phi_{n-1}),\\
Z_n &= R\,\sin(\phi_1)\,\cdots\,\sin(\phi_{n-2})\,\sin(\phi_{n-1}).
\end{aligned}$$

  - Distance from origin: $R = \|\boldsymbol{Z}\|$.
  - Direction: $\boldsymbol{\phi} = \big(\phi_1, \ldots, \phi_{n-1}\big)^\top$.
- Exercise for the 3rd year bachelor students:
  If $\boldsymbol{Z} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then distance R from the origin and direction $\boldsymbol{\phi}$ are independent.
- $R = \|\boldsymbol{Z}\|$ (distance from origin itself), $T(\boldsymbol{Z})$ depends on the direction only (since $T(\boldsymbol{Z}) = T(c\boldsymbol{Z})$ for all $c > 0$) and hence $\|\boldsymbol{Z}\|$ and $T(\boldsymbol{Z})$ are independent.

❑

Appendix C

Asymptotic Theorems

Theorem C.1 Strong law of large numbers (SLLN) for i.n.n.i.d. random variables.

Let $Z_1, Z_2, \ldots$ *be a sequence of* independent not necessarily identically distributed (i.n.n.i.d.) *random variables. Let* $\mathbb{E}(Z_i) = \mu_i$, $\mathsf{var}(Z_i) = \sigma_i^2$, $i = 1, 2, \ldots$. *Let*

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty.$$

Then

$$\frac{1}{n} \sum_{i=1}^{n} (Z_i - \mu_i) \xrightarrow{a.s.} 0 \qquad \text{as } n \to \infty.$$

Proof. See *Probability and Mathematical Statistics (NMSA202)* lecture (2nd year of the Bc. study programme).

❑

Theorem C.2 Strong law of large numbers (SLLN) for i.i.d. random variables.

Let $Z_1, Z_2, \ldots$ *be a sequence of* independent identically distributed (i.i.d.) *random variables.*

Then

$$\frac{1}{n} \sum_{i=1}^{n} Z_i \xrightarrow{a.s.} \mu \qquad \text{as } n \to \infty$$

for some $\mu \in \mathbb{R}$ *if and only if*

$$\mathbb{E}\big|Z_1\big| < \infty,$$

in which case $\mu = \mathbb{E}\big(Z_1\big)$.

Proof. See *Probability and Mathematical Statistics (NMSA202)* lecture (2nd year of the Bc. study programme).

❑

Theorem C.3 Central limit theorem (CLT), Lyapunov.

Let $Z_1, Z_2, \ldots$ be a sequence of i.n.n.i.d. random variables with

$$\mathbb{E}\big(Z_i\big) = \mu_i, \quad \infty > \mathsf{var}\big(Z_i\big) = \sigma_i^2 > 0, \qquad i = 1, 2, \ldots$$

Let for some $\delta > 0$

$$\frac{\sum_{i=1}^{n} \mathbb{E}\big|Z_i - \mu_i\big|^{2+\delta}}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{\frac{2+\delta}{2}}} \longrightarrow 0 \quad \text{as } n \to \infty.$$

Then

$$\frac{\sum_{i=1}^{n}(Z_i - \mu_i)}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad \text{as } n \to \infty.$$

Proof. See *Probability Theory 1 (NMSA333)* lecture (3rd year of the Bc. study programme).

❑

Theorem C.4 Central limit theorem (CLT), i.i.d..

Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. random variables with

$$\mathbb{E}\big(Z_i\big) = \mu, \quad \infty > \mathsf{var}\big(Z_i\big) = \sigma^2 > 0, \qquad i = 1, 2, \ldots.$$

Let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Z_i - \mu}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad \text{as } n \to \infty,$$

$$\sqrt{n}\left(\overline{Z}_n - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \quad \text{as } n \to \infty.$$

Proof. See *Probability Theory 1 (NMSA333)* lecture (3rd year of the Bc. study programme).

❑

Theorem C.5 Central limit theorem (CLT), i.i.d. multivariate.
Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ be a sequence of i.i.d. p-dimensional random vectors with

$$\mathbb{E}(\boldsymbol{Z}_i) = \boldsymbol{\mu}, \qquad \mathsf{var}(\boldsymbol{Z}_i) = \boldsymbol{\Sigma}, \qquad i = 1, 2, \ldots,$$

where $\boldsymbol{\Sigma}$ is a real *positive semidefinite matrix. Let $\overline{\boldsymbol{Z}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i$.*
Then

$$\sqrt{n}\left(\overline{\boldsymbol{Z}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}).$$

If $\boldsymbol{\Sigma}$ is positive definite then also

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{Z}_i - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p).$$

Proof. See *Probability Theory 1 (NMSA333)* lecture (3rd year of the Bc. study programme). ❑

Theorem C.6 Cramér-Wold.
Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ be a sequence of p-dimensional random vectors. Let $\boldsymbol{Z}$ be a p-dimensional random vector.

$$\boldsymbol{Z}_n \xrightarrow{\mathcal{D}} \boldsymbol{Z} \qquad \text{as } n \to \infty$$

if and only if for all $\mathbf{l} \in \mathbb{R}^p$

$$\mathbf{l}^\top \boldsymbol{Z}_n \xrightarrow{\mathcal{D}} \mathbf{l}^\top \boldsymbol{Z} \qquad \text{as } n \to \infty.$$

Proof. See *Probability Theory 1 (NMSA333)* lecture (3rd year of the Bc. study programme). ❑

Theorem C.7 Cramér-Slutsky.

Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ be a sequence of random vectors such that

$$\boldsymbol{Z}_n \xrightarrow{\mathcal{D}} \boldsymbol{Z} \qquad \text{as} \quad n \to \infty,$$

where $\boldsymbol{Z}$ be a random vector. Let $S_1, S_2, \ldots$ be a sequence of random variables such that

$$S_n \xrightarrow{\mathsf{P}} S \qquad \text{as} \quad n \to \infty,$$

where $S \in \mathbb{R}$ is a real constant.

Then

(i) $S_n \boldsymbol{Z}_n \xrightarrow{\mathcal{D}} S \boldsymbol{Z}$ *as* $n \to \infty$.

(ii) $\frac{1}{S_n} \boldsymbol{Z}_n \xrightarrow{\mathcal{D}} \frac{1}{S} \boldsymbol{Z}$ *as* $n \to \infty$, *if* $S \neq 0$.

Proof. See *Probability Theory 1 (NMSA333)* lecture (3rd year of the Bc. study programme).

See also Shao (2003, Theorem 1.11 in Section 1.5).

Bibliography

Anděl, J. (2007). *Základy matematické statistiky*. Matfyzpress, Praha. ISBN 80-7378-001-1.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, **160**(901), 268–282. doi: 10.1098/rspa.1937.0109.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47**(5), 1287–1294. doi: 10.2307/1911963.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**(346), 364–367. doi: 10.1080/01621459.1974.10482955.

Cipra, T. (2008). *Finanční ekonometrie*. Ekopress, Praha. ISBN 978-80-86929-43-9.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**(1), 1–10. doi: 10.1093/biomet/70.1.1.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, **45**(2), 215–233. doi: 10.1016/S0167-9473(02)00366-3.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York. ISBN 0-387-90356-9.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York, Revised edition. ISBN 0-387-95366-3.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon, Oxford. ISBN 0-19-853440-X.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons, New York, Third edition. ISBN 0-471-17082-8.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, **37**, 409–428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**(1/2), 159–177. doi: 10.2307/2332325.

Durbin, J. and Watson, G. S. (1971). Testing for serial correlation in least squares regression III. *Biometrika*, **58**(1), 1–19. doi: 10.2307/2334313.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26. doi: 10.1214/aos/1176344552.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statistical Science*, **11**(1), 89–121. doi: 10.1214/ss/1038425655.

Farebrother, R. W. (1980). Algorithm AS 153: Pan's procedure for the tail probabilities of the Durbin-Watson statistics. *Applied Statistics*, **29**(2), 224–227.

Farebrother, R. W. (1984). Remark AS R53: A remark on algorithm AS 106, AS 153, AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics*, **33**, 366–369.

Fligner, M. A. and Killeen, T. J. (1976). Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, **71**(353), 210–213. doi: 10.2307/2285771.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, **87**(417), 178–183. doi: 10.1080/01621459.1992.10475190.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Springer-Verlag, New York. ISBN 978-3-642-01688-2.

Goldfeld, S. M. and Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, **60**(310), 539–547. doi: 10.1080/01621459.1965.10480811.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *The Annals of Statistics*, **12**(1), 61–75. doi: 10.1214/aos/1176346392.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, **50**(3), 346–363. doi: 10.1002/bimj.200810425.

Hothorn, T., Bretz, F., and Westfall, P. (2011). *Multiple Comparisons Using R*. Chapman & Hall/CRC, Boca Raton. ISBN 978-1-5848-8574-0.

Khuri, A. I. (2010). *Linear Model Methodology*. Chapman & Hall/CRC, Boca Raton. ISBN 978-1-58488-481-1.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, **17**(1), 107–112. doi: 10.1016/0304-4076(81)90062-2.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, **12**(3), 307–310. doi: 10.2307/3001469.

Levene, H. (1960). Robust tests for equality of variances. In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., and Mann, H. B., editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press, Standord.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, **54**(3), 217–224. doi: 10.2307/2685594.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, **29**(3), 305–325. doi: 10.1016/0304-4076(85)90158-7.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons, New York, Second edition. ISBN 0-471-21875-8.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. John Wiley & Sons, New York. ISBN 0-471-84096-3.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*. John Wiley & Sons, New York, Second edition. ISBN 978-0-47141-540-4.

Shao, J. (2003). *Mathematical Statistics*. Springer Science+Business Media, New York, Second edition. ISBN 0-387-95382-5.

Sun, J. (2003). *Mathematical Statistics*. Springer Science+Business Media, New York, Second edition. ISBN 0-387-95382-5.

Tukey, J. W. (1949). Comparing individual means in the Analysis of variance. *Biometrics*, **5**(2), 99–114. doi: 10.2307/3001913.

Tukey, J. W. (1953). The problem of multiple comparisons (originally unpublished manuscript). In Braun, H. I., editor, *The Collected Works of John W. Tukey*, volume 8, 1994. Chapman & Hall, New York.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons, Hoboken, Third edition. ISBN 0-471-66379-4.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**(4), 817–838. doi: 10.2307/1912934.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, **11**(10), 1–17. URL http://www.jstatsoft.org/v11/i10/.

Zvára, K. (1989). *Regresní analýza*. Academia, Praha. ISBN 80-200-0125-5.

Zvára, K. (2008). *Regrese*. Matfyzpress, Praha. ISBN 978-80-7378-041-8.