Department of Probability and Mathematical Statistics

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

NMSA407 Linear Regression

Course Notes

2021–22

Arnošt Komárek

Last modified on September 2, 2021.

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 September 2, 2021.

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. Major revision of some parts conducted in Fall 2020.

Contents

Preface

Linear regression is a basic statistical modelling technique. Its principles, related mathematical theory and its applications are covered by a variety of textbooks or monographs. This text follows quite closely the course “NMSA407 Linear Regression” at Faculty of Mathematics and Physics (MatFyz) of Charles University which makes a part of the curriculum of the Master’s programs “Probability, Mathematical Statistics and Econometrics” and “Financial and Insurance Mathematics”. In current form, this course is being taught since fall 2013 when also development of this text started. During several decades before, a similar course, entitled just *Regression* was taught at MatFyz by Karel Zvára. Linear regression occupied indeed majority of the *Regression* course and complementary textbook *Regrese* (Zvára, 2008) also served as a primary source and inspiration for several chapters of this text, especially those devoted to a classical least squares theory in a (normal) linear model.

As complementary literature to this course, the following textbooks and monographs can be recommended, in English: Khuri (2010); Seber and Lee (2003); Draper and Smith (1998); Shao (2003); Weisberg (2005), in Czech: Zvára (2008); Anděl (2007); Cipra (2008); Zvára (1989). For practical analyzes, the R software (R Core Team, 2020) is perhaps the best choice. R is also used during the exercise classes to the *Linear Regression* course.

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

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

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

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

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

Notation related to the linear model

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

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

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

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

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

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

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

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

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

Chapter 1

Linear Model

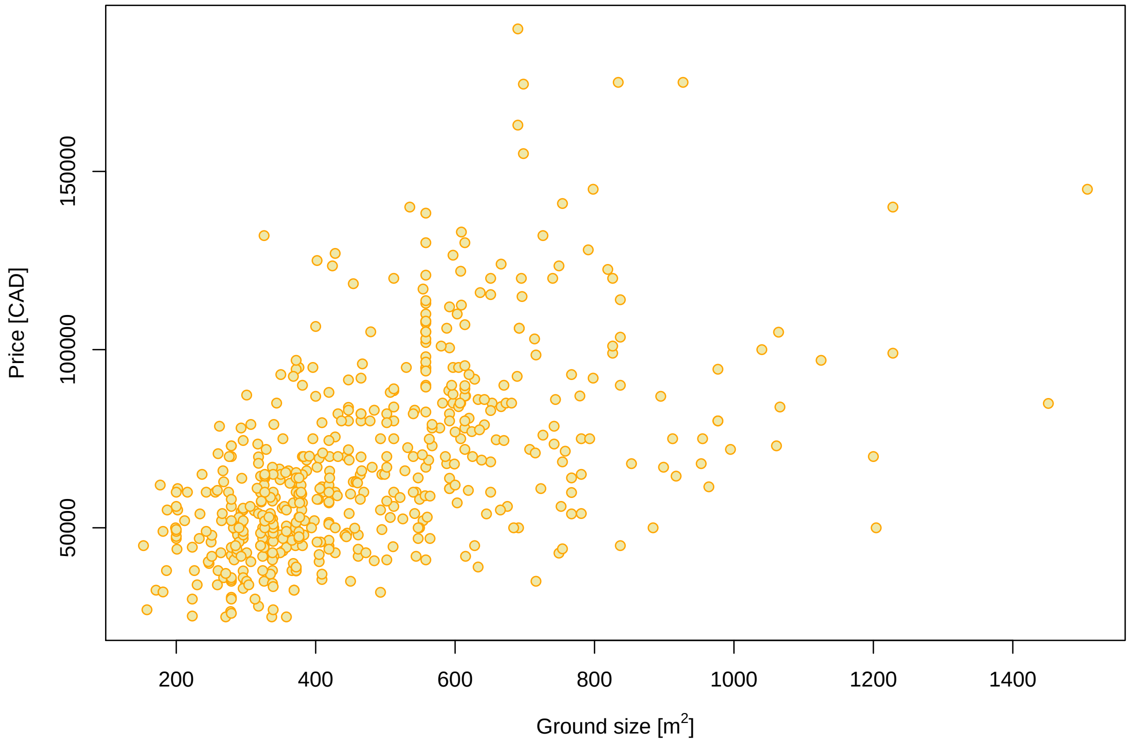
1.1 Regression analysis

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.

Illustrations

`Houses1987` ($n = 546$)

`price` $\sim$ `ground`

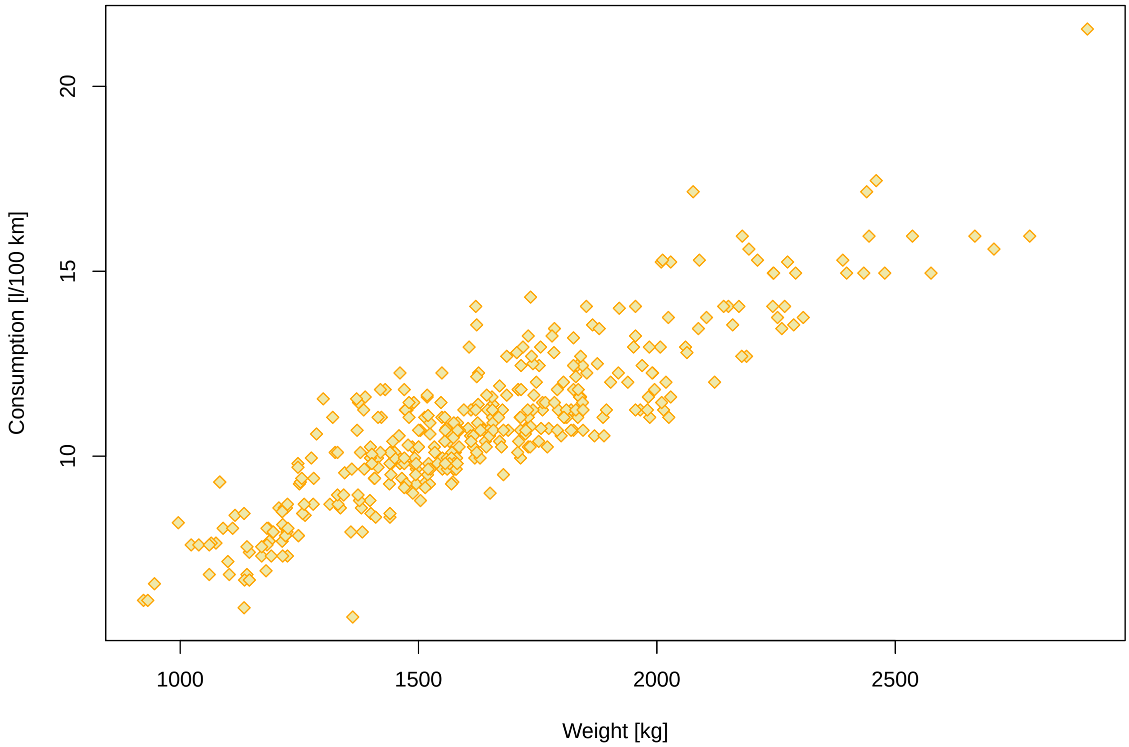


[1] *lineární regrese* [2] *regresní analýza* [3] *lineární model*

Illustrations

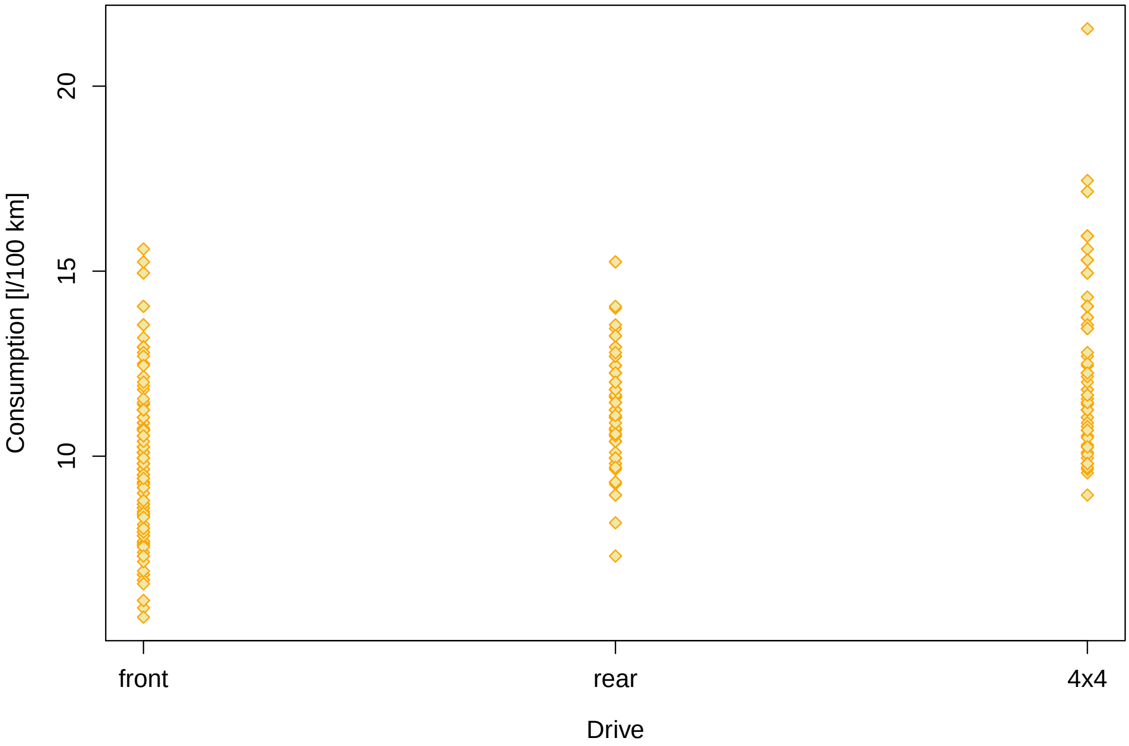
`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` ~ `weight`



`Cars2004nh` **(subset,** $n = 409$**)**

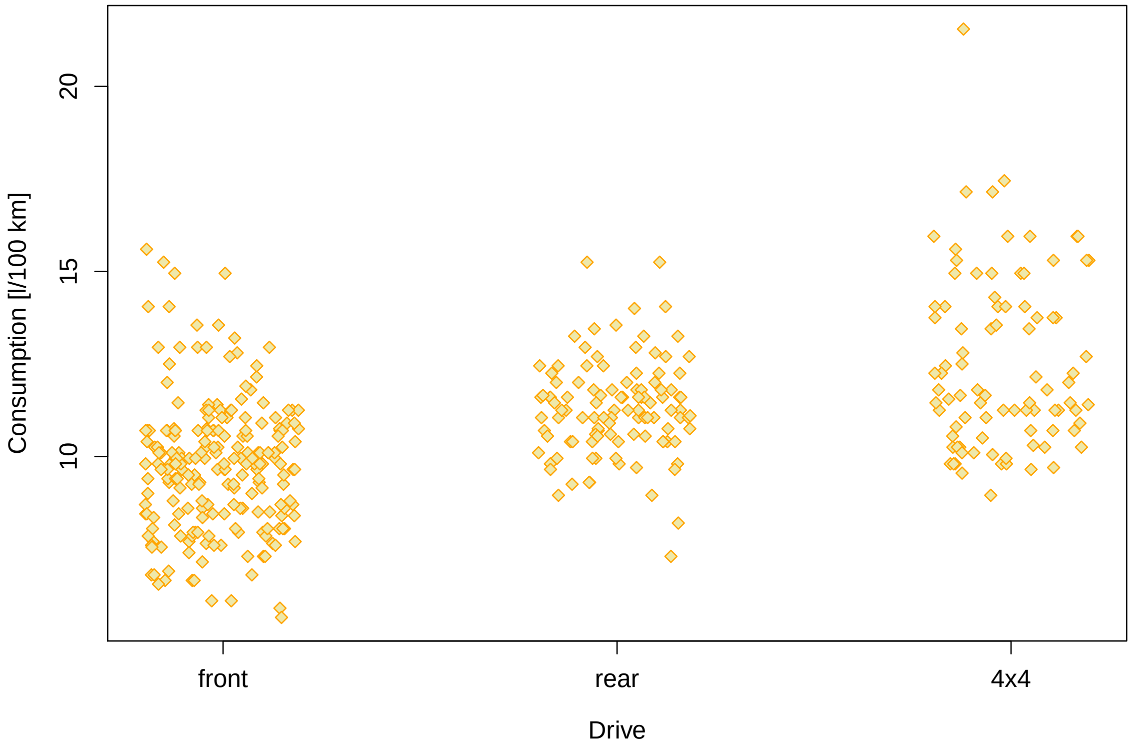
`consumption` ~ `drive`



Illustrations

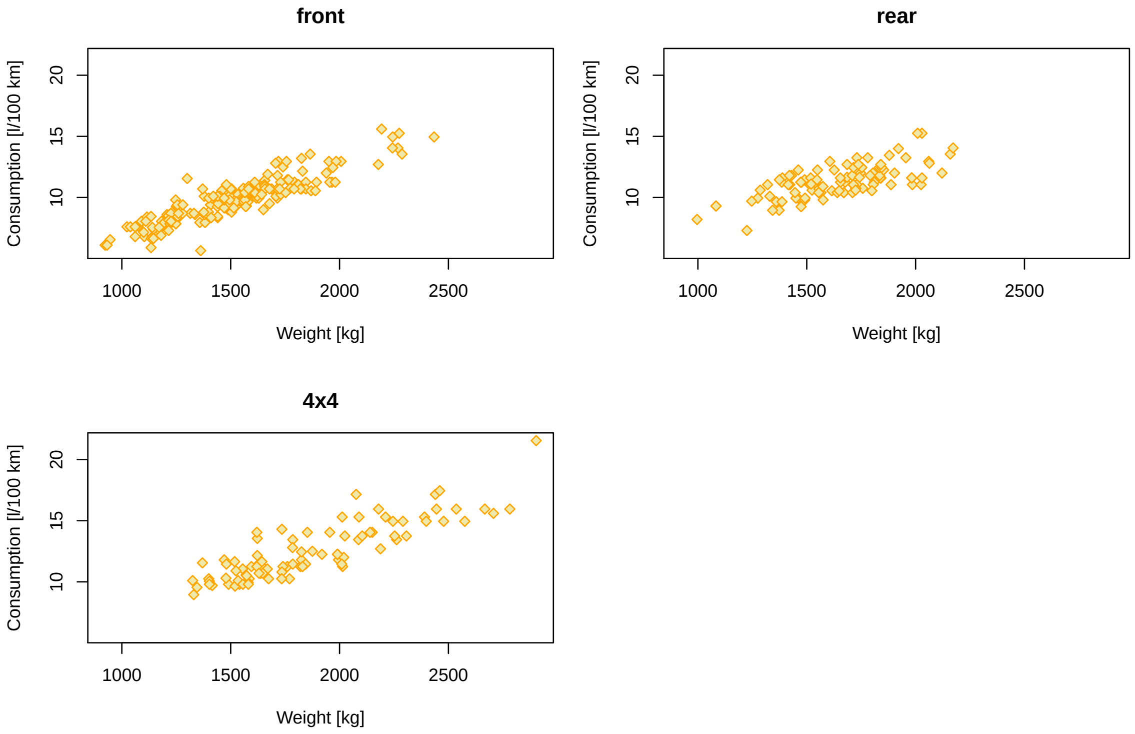
`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `drive`



`Cars2004nh` **(subset,** $n = 409$**)**

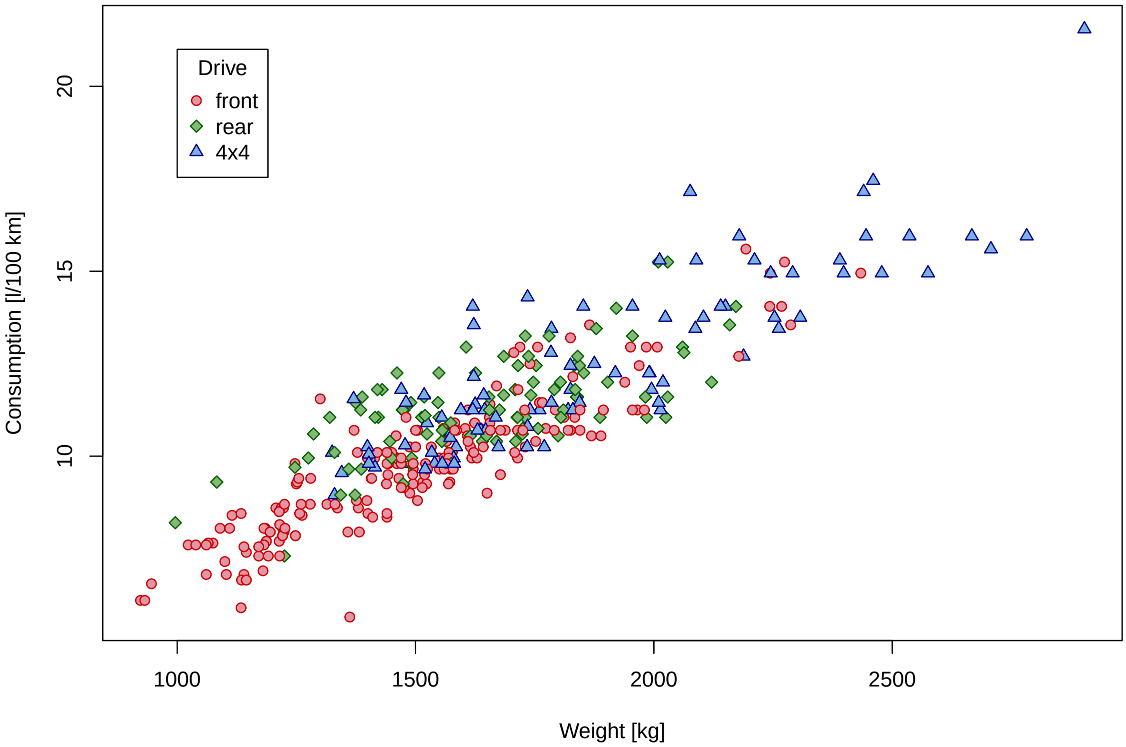
`consumption` $\sim$ `weight, drive`



Illustrations

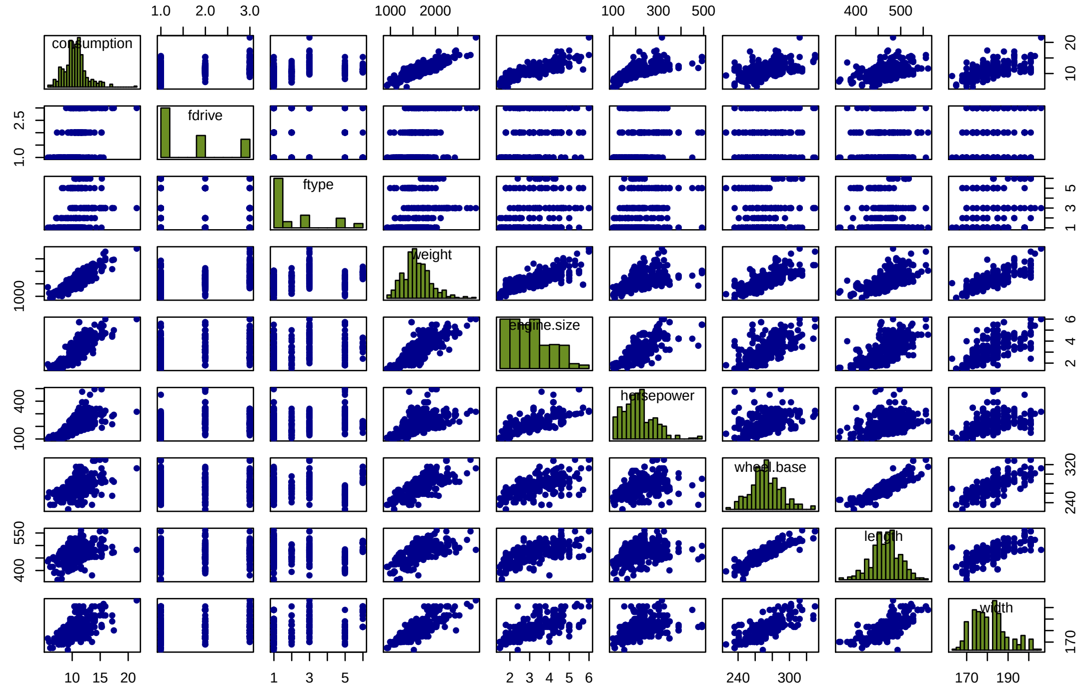
`Cars2004nh` **(subset,** $n = 409$)

`consumption ~ weight, drive`



`Cars2004nh` **(subset,** $n = 384$)

`consumption ~ drive, type, weight, engine.size, horsepower, wheel.base, length, width`



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 $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $i = 1, \dots, n$, being distributed as a generic random vector $\bigl(Y,\, \boldsymbol{Z}^\top\bigr)^\top$. That is,

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

where $\boldsymbol{Z} = \bigl(Z_1, \dots, Z_p\bigr)^\top$. 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, $\mathsf{P}(\boldsymbol{Z} \in \mathcal{Z}) = 1$.

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

Further, let

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

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

[4] *odezva* [5] *závisle proměnná* [6] *Nepřekládá se. Výraz „kovariáty" nepoužívat!* [7] *vysvětlující proměnné* [8] *prediktory* [9] *nezávisle proměnné* [10] *výběrový prostor* [11] *vektor odezvy* [12] *matice vysvětlujících proměnných*

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 $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $i = 1, \dots, n$, the data generating mechanism corresponds to a joint distribution of a generic random vector $\bigl(Y,\, \boldsymbol{Z}^\top\bigr)^\top$ 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 $\bigl(\mathbb{R},\, \mathcal{B}\bigr)$.

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

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

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 $\mathbb{E}\bigl(Y \,\big|\, \boldsymbol{Z}\bigr)$, even the density $f_{Y|\boldsymbol{Z}}$ might be left practically unspecified for many tasks.

1.1.3 Regressors

In the reminder of the lecture, we will mainly attempt to model the conditional mean $\mathbb{E}\bigl(Y \,\big|\, \boldsymbol{Z}\bigr)$. 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, $\boldsymbol{t} = \bigl(t_0,\, \dots,\, t_{k-1}\bigr)^\top$ (for reasons which become clear in a while, we start indexing of the elements of this transformation by zero). Further, let

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

Subsequently, we will assume that

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

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

Terminology *(Regressors, regression function).*

- The vectors $\boldsymbol{X}$, $\boldsymbol{X}_i$, $i = 1, \dots, n$, 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 $\boldsymbol{X}^j := \bigl(X_{1,j},\, \dots,\, X_{n,j}\bigr)^\top$ $(j = 0,\, \dots,\, k-1)$ is called the j*th regressor vector*.[15]

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

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 the 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.

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 $\bigl(Y_i,\,\boldsymbol{X}_i^\top\bigr)^\top \overset{\text{i.i.d.}}{\sim} \bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$*,* $i = 1,\dots,n$*, satisfy a* linear model *if*

$$\mathbb{E}\bigl(Y\,\big|\,\boldsymbol{X}\bigr) = \boldsymbol{X}^\top\boldsymbol{\beta}, \qquad \mathsf{var}\bigl(Y\,\big|\,\boldsymbol{X}\bigr) = \sigma^2,$$

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

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

- $\boldsymbol{\beta} = \bigl(\beta_0,\,\dots,\,\beta_{k-1}\bigr)^\top$ 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 $Y\,\big|\,\boldsymbol{X}$. 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

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

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 $\boldsymbol{X} = \bigl(1,\,X_1,\,\dots,\,X_{k-1}\bigr)^\top$ and the regression function becomes

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

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 $j \in \bigl\{0,\,1,\,\dots,k-1\bigr\}$

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

[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

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

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 $\delta \in \mathbb{R}$

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

we then have

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

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

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

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

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

1.2.3 Linear model with general data

Notation and terminology *(Model matrix).*

Let

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

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

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

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

The response random vector $\boldsymbol{Y} = \bigl(Y_1,\, \dots,\, Y_n\bigr)^\top$, 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 $\bigl(Y_1,\, \dots,\, Y_n,\, \boldsymbol{X}_1^\top,\, \dots,\, \boldsymbol{X}_n^\top\bigr)^\top \equiv \bigl(\boldsymbol{Y},\, \mathbb{X}\bigr)$ 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

$$f_{\boldsymbol{Y},\mathbb{X}}(\boldsymbol{y},\, \mathbf{x}) = f_{\boldsymbol{Y}|\mathbb{X}}\bigl(\boldsymbol{y} \,\big|\, \mathbf{x}\bigr)\, f_{\mathbb{X}}(\mathbf{x}). \tag{1.3}$$

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

$$f_{\boldsymbol{Y},\mathbb{X}}(\boldsymbol{y},\, \mathbf{x}) = \underbrace{\Bigl\{\prod_{i=1}^n f_{Y|\boldsymbol{X}}\bigl(y_i \,\big|\, \boldsymbol{x}_i\bigr)\Bigr\}}_{f_{\boldsymbol{Y}|\mathbb{X}}(\boldsymbol{y}\,|\,\mathbf{x})}\, \underbrace{\Bigl\{\prod_{i=1}^n f_{\boldsymbol{X}}(\boldsymbol{x}_i)\Bigr\}}_{f_{\mathbb{X}}(\mathbf{x})}. \tag{1.4}$$

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

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

[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 $\bigl(Y_i,\, \boldsymbol{X}_i^\top\bigr)^\top \overset{\text{i.i.d.}}{\sim} \bigl(Y,\, \boldsymbol{X}^\top\bigr)^\top$, $i = 1, \ldots, n$, where $\mathbb{E}\bigl(Y \,\big|\, \boldsymbol{X}\bigr) = \boldsymbol{X}^\top\boldsymbol{\beta}$, $\mathsf{var}\bigl(Y \,\big|\, \boldsymbol{X}\bigr) = \sigma^2$. To derive many results shown later in this lecture, it is sufficient to assume that the full data ($\equiv \bigl(\boldsymbol{Y},\, \mathbb{X}\bigr)$) satisfy just the weaker condition (1.5) without requesting that the random vectors $\bigl(Y_i,\, \boldsymbol{X}_i^\top\bigr)^\top$, $i = 1, \ldots, n$, 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 $\bigl(\boldsymbol{Y},\, \mathbb{X}\bigr)$, *satisfy a* linear model *if*

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

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

Notation.

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

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

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

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

Note. If $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$ 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 $r \leq k$,

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

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 $\mathsf{rank}\big(\mathbb{X}\big) = r$ 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

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

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

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

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*

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

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

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

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 $\bigl(Y_i,\,\boldsymbol{X}_i^\top\bigr)^\top \overset{\text{i.i.d.}}{\sim} \bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$, $i = 1, \ldots, n$, many results can be derived (see Chapter 6) if it is assumed

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

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 16), 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 $\mathbb{E}\bigl(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\bigr)$, $\boldsymbol{x} \in \mathcal{X}$, cannot be expressed as a linear function of $\boldsymbol{x}$.
  $\Rightarrow$ *Incorrect regression function.*

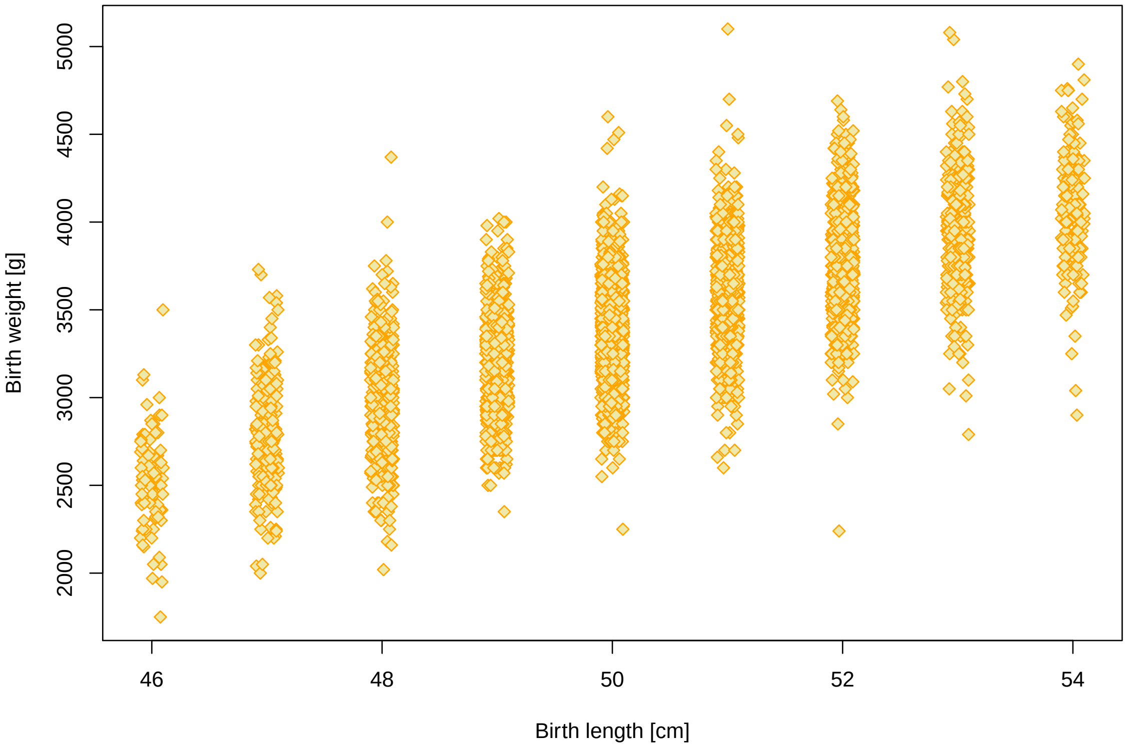
- The conditional variance $\mathsf{var}\bigl(Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\bigr)$, $\boldsymbol{x} \in \mathcal{X}$, is not constant. It may depend on $\boldsymbol{x}$ as well, it may depend on other factors.
  $\Rightarrow$ *Heteroscedasticity.*

- 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.

Illustrations

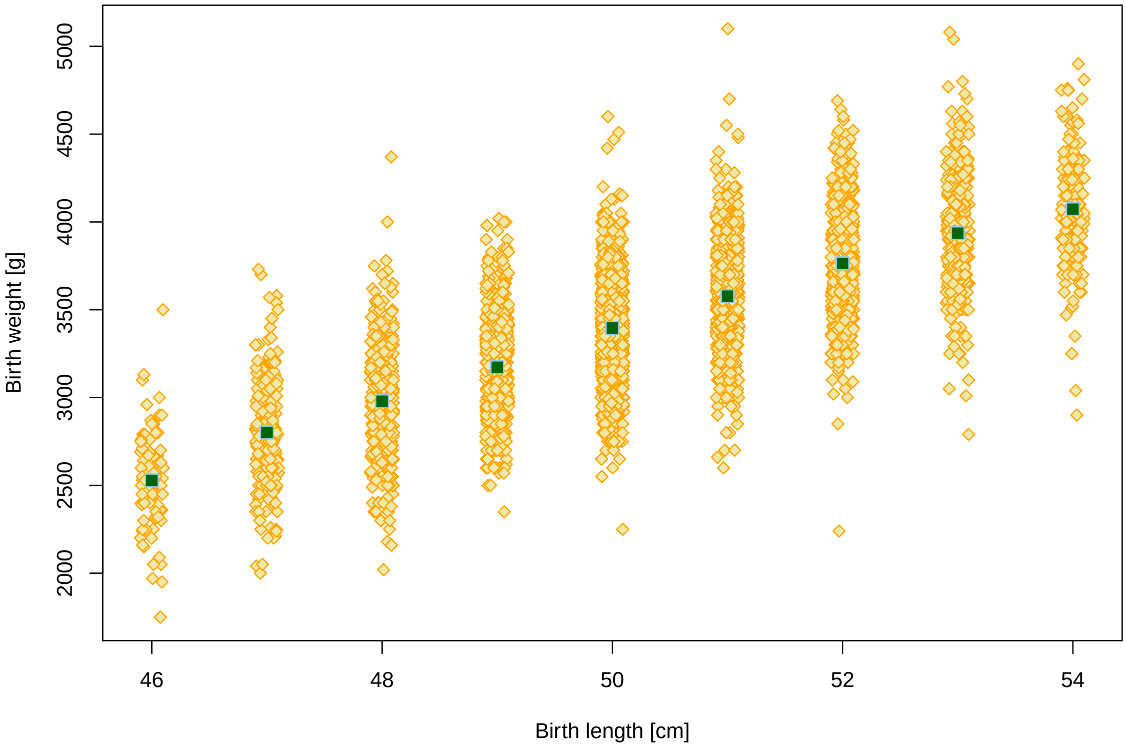
`Hosi0` ($n = 4838$)

`bweight` $\sim$ `blength`



`Hosi0` ($n = 4838$)

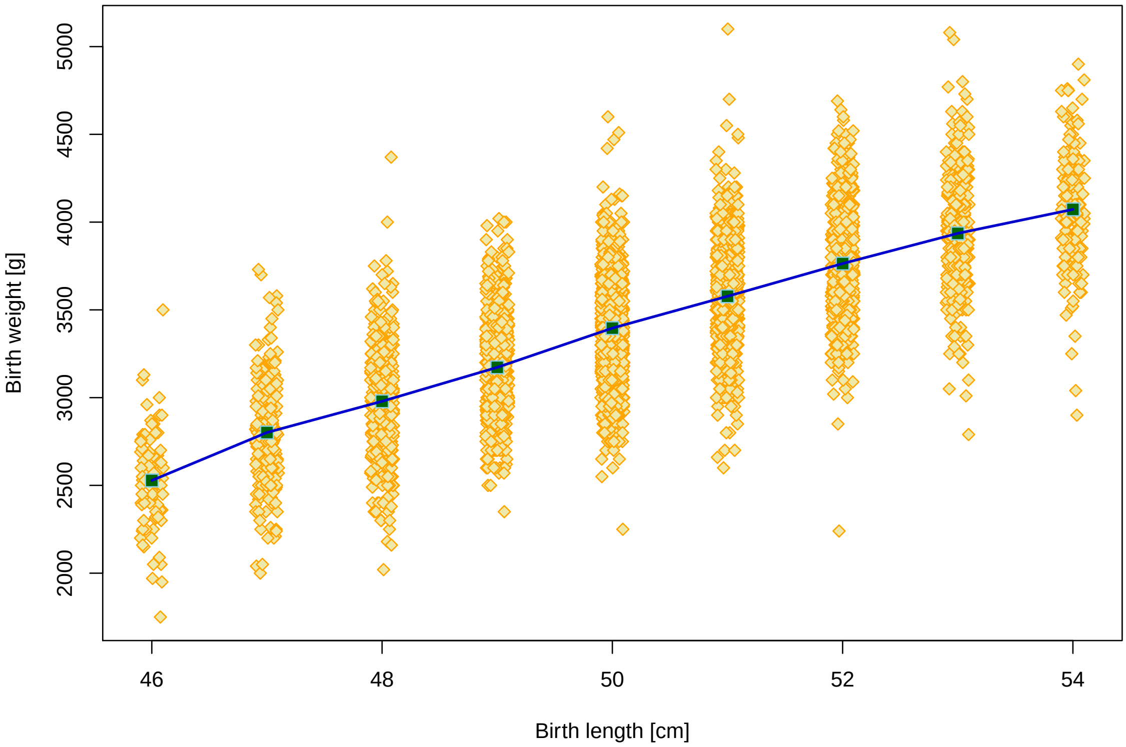
`bweight` $\sim$ `blength`



Illustrations

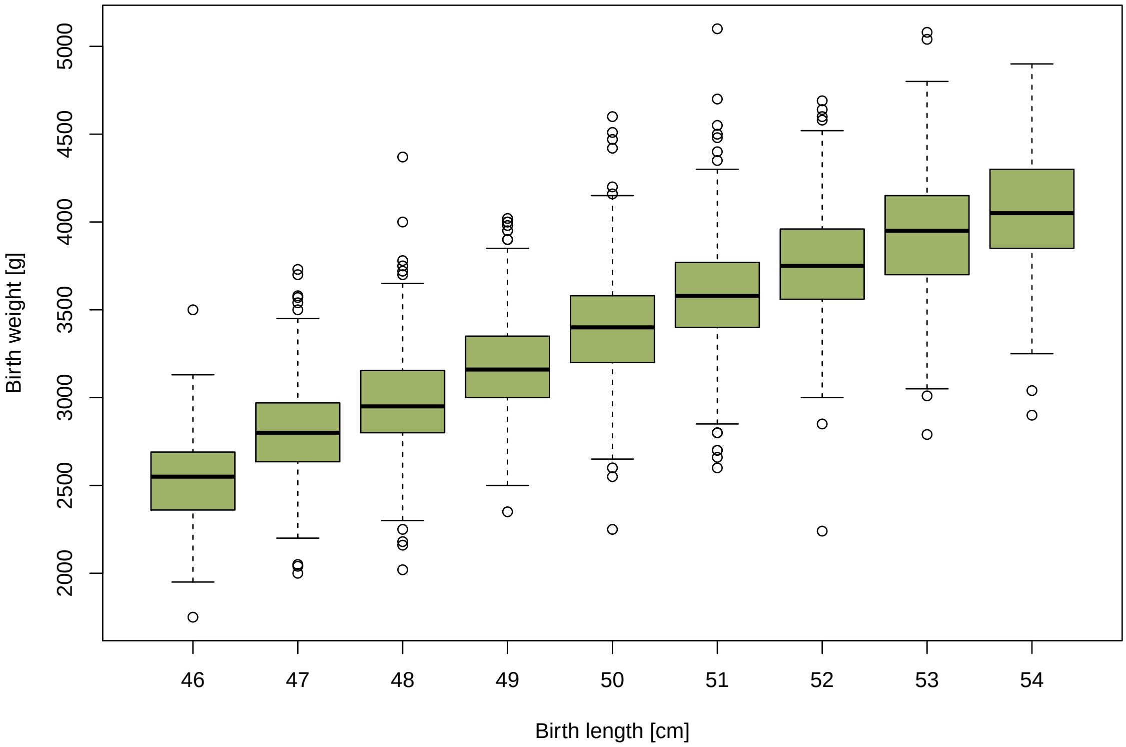
`Hosi0` ($n = 4838$)

`bweight` $\sim$ `blength`



`Hosi0` ($n = 4838$)

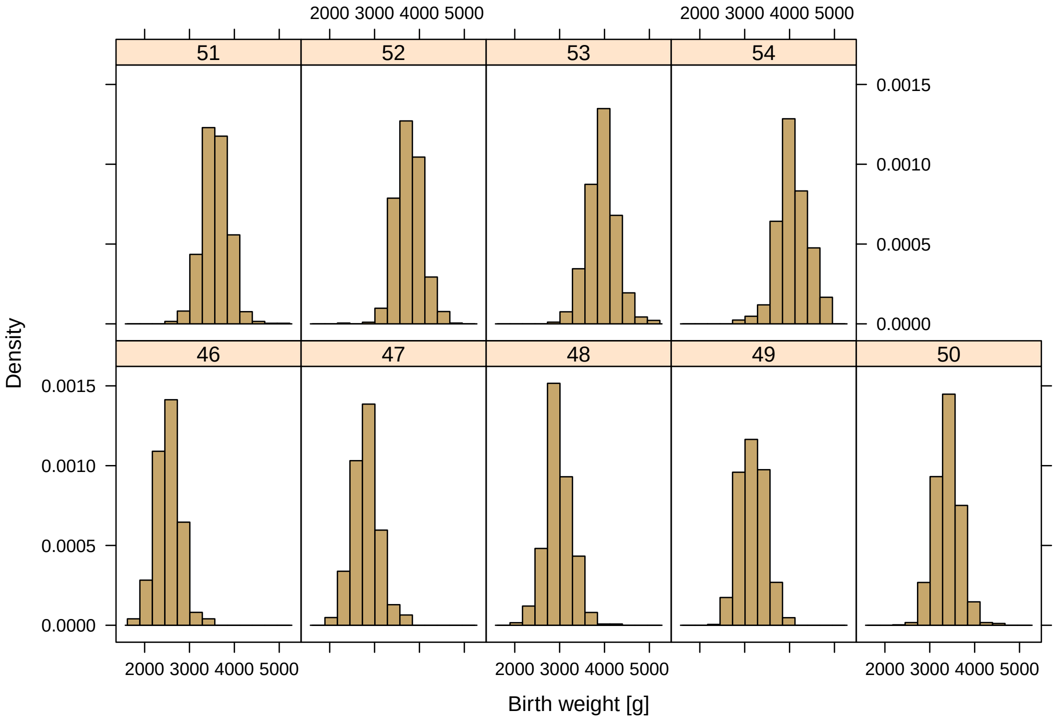
`bweight` $\sim$ `blength`



Illustrations

`Hosi0` ($n = 4838$)

`bweight` $\sim$ `blength`



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 $Y \mid \boldsymbol{X}$ are of (primary) interest.

Chapter 2

Least Squares Estimation

This chapter is not complete in the notes.

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

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

where $\boldsymbol{Y} = \bigl(Y_1, \ldots, Y_n\bigr)^\top$, $\mathbb{X}$ is a matrix with vectors $\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top$ in its rows and $\boldsymbol{\beta} = \bigl(\beta_0, \ldots, \beta_{k-1}\bigr) \in \mathbb{R}^k$ 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. Certain results will be derived while additionally assuming the full-rank model ($r = k$).

[1] *metoda nejmenších čtverců*

2.1 Sum of squares, least squares estimator and normal equations

Definition 2.1 Sum of squares.

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

$$\mathsf{SS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta})^2 = \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big\|^2 = \big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big)^\top\big(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big), \qquad \boldsymbol{\beta} \in \mathbb{R}^k$$

will be called the sum of squares[2] *of the model.*

Lemma 2.1 Least squares estimator.

Assume a full-rank linear model $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$*. There exist a unique minimizer to* $\mathsf{SS}(\boldsymbol{\beta})$ *given as*

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

Definition 2.2 Least squares estimator, normal equations.

Consider a linear model $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$*. The quantity* $\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ *will be called the* least squares estimator (LSE)[3] *of the vector of regression coefficients* $\boldsymbol{\beta}$*. The linear system* $\mathbb{X}^\top\mathbb{X}\,\boldsymbol{\beta} = \mathbb{X}^\top\boldsymbol{Y}$ *will be called the system of* normal equations.[4]

Lemma 2.2 Moments of the least squares estimator.

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

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

[2] *součet čtverců* [3] *odhad metodou nejmenších čtverců (MNČ)* [4] *normální rovnice*

2.2 Fitted values, residuals, projections

Definition 2.3 Regression and residual space of a linear model.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *The* regression space[5] *of the model is a vector space* $\mathcal{M}\big(\mathbb{X}\big)$. *The* residual space[6] *of the model is the orthogonal complement of the regression space, i.e., a vector space* $\mathcal{M}\big(\mathbb{X}\big)^{\perp}$.

Definition 2.4 Fitted values, residuals.

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

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

will be called the vector of fitted values[7] *of the model. The vector*

$$\boldsymbol{U} := \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

will be called the vector of residuals[8] *of the model.*

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

Lemma 2.3 Algebraic properties of fitted values, residuals and related projection matrices.

(i) $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$ *and* $\boldsymbol{U} = \mathbb{M}\boldsymbol{Y}$ *are projections of* $\boldsymbol{Y}$ *into* $\mathcal{M}\big(\mathbb{X}\big)$ *and* $\mathcal{M}\big(\mathbb{X}\big)^{\perp}$*, respectively;*

(ii) $\widehat{\boldsymbol{Y}} \perp \boldsymbol{U}$*;*

(iii) $\mathbb{H}$ *and* $\mathbb{M}$ *are projection matrices into* $\mathcal{M}\big(\mathbb{X}\big)$ *and* $\mathcal{M}\big(\mathbb{X}\big)^{\perp}$*, respectively;*

(iv) $\mathbb{H}^\top = \mathbb{H},\ \ \mathbb{M}^\top = \mathbb{M}$*;*

(v) $\mathbb{H}\,\mathbb{H} = \mathbb{H},\ \ \mathbb{M}\,\mathbb{M} = \mathbb{M}$*;*

(vi) $\mathbb{H}\,\mathbb{X} = \mathbb{X},\ \ \mathbb{M}\,\mathbb{X} = \mathbf{0}_{n\times k}$.

Terminology *(Hat matrix, residual projection matrix).*

For a linear model of (not necessarily full-rank) $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$.

[5] *regresní prostor* [6] *reziduální prostor* [7] *vyrovnané hodnoty* [8] *rezidua*

- $\mathbb{H} = \mathbb{Q}\,\mathbb{Q}^\top = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top$: *hat matrix*[9],
  where $\mathbb{Q}_{n\times r} = \big(\boldsymbol{q}_1,\,\dots,\,\boldsymbol{q}_r\big)$ is an orthonormal vector basis of the regression space $\mathcal{M}\big(\mathbb{X}\big)$;
- $\mathbb{M} = \mathbb{N}\,\mathbb{N}^\top = \mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^-\mathbb{X}^\top$: *residual projection matrix*[10],
  where $\mathbb{N}_{n\times r} = \big(\boldsymbol{n}_1,\,\dots,\,\boldsymbol{n}_{n-r}\big)$ is an orthonormal vector basis of the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$.

[9] *regresní projekční matice, lze však užívat též výrazu „hat matice“*

[10] *reziduální projekční matice*

2.3 Gauss-Markov theorem

Theorem 2.4 Gauss–Markov.

Assume a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *Then the vector of fitted values* $\widehat{\boldsymbol{Y}}$ *is, conditionally given* $\mathbb{X}$, *the* best linear unbiased estimator (BLUE)[11] *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 $\boldsymbol{\beta} = \big(0,\ldots,1,\ldots,0\big)^\top$ 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,

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

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

$$\begin{aligned} \mathbb{B}\mathbb{X} &= \mathbb{X}, && \Big|\;\cdot\;\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top\\ \mathbb{B}\mathbb{X}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top &= \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top, &&\\ \mathbb{B}\mathbb{H} &= \mathbb{H}, && (2.3)\\ \mathbb{H}^\top\mathbb{B}^\top &= \mathbb{H}^\top, &&\\ \mathbb{H}\mathbb{B}^\top &= \mathbb{H}. && (2.4) \end{aligned}$$

[11] *nejlepší lineární nestranný odhad*

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

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

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

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

  where $\mathbb{H}(\mathbb{B}-\mathbb{H})^\top = (\mathbb{B}-\mathbb{H})\mathbb{H}^\top = \mathbf{0}_n$ follow from (2.3) and (2.4) and from the fact that $\mathbb{H}$ is symmetric and idempotent.

- Hence finally,

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

  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

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

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.

  ⇒ The name Gauss–Markov theorem.

Theorem 2.5 Gauss–Markov for linear combinations.

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

(i) *For a vector* $\mathbf{l} = \big(l_0,\, \ldots,\, l_{k-1}\big)^\top \in \mathbb{R}^k$, $\mathbf{l} \neq \mathbf{0}$, *the statistic* $\widehat{\theta} = \mathbf{l}^\top\widehat{\boldsymbol{\beta}}$ *is the* best linear unbiased estimator (BLUE) *of the parameter* $\theta = \mathbf{l}^\top\boldsymbol{\beta}$ *with*

$$\mathsf{var}\big(\widehat{\theta}\,\big|\,\mathbb{X}\big) \;=\; \sigma^2\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l} > 0.$$

(ii) *For a given matrix*

$$\mathbb{L} = \begin{pmatrix} \mathbf{l}_1^\top \\ \vdots \\ \mathbf{l}_m^\top \end{pmatrix}, \quad \mathbf{l}_j \in \mathbb{R}^k,\ \mathbf{l}_j \neq \mathbf{0}, \quad j = 1,\ldots,m, \quad m \leq k$$

with linearly independent rows ($\mathsf{rank}\big(\mathbb{L}_{m\times k}\big) = m$*), the statistic* $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}}$ *is the* best linear unbiased estimator (BLUE) *of the vector parameter* $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ *with*

$$\mathsf{var}\big(\widehat{\boldsymbol{\theta}}\,\big|\,\mathbb{X}\big) \;=\; \sigma^2\,\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top,$$

which is a positive definite matrix.

2.4 Residuals, properties

Definition 2.5 Residual sum of squares.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *The quantity* $\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2 = \sum_{i=1}^n U_i^2 = \sum_{i=1}^n \big(Y_i - \widehat{Y}_i\big)^2 = \big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2$ *will be called the* residual sum of squares[12] *of the model.*

Lemma 2.6 Altenative expressions of residuals and residual sum of squares.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *The following then holds.*

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

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

Lemma 2.7 Moments of residuals and residual sum of squares.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *Then*

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

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

Definition 2.6 Residual mean square and residual degrees of freedom.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$.

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

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

(ii) The residual degrees of freedom[14] *of the model is the vector dimension of the residual space* $\mathcal{M}\big(\mathbb{X}\big)^\perp$ *and will be denotes as* ν_e. *That is,*

$$\nu_e = n - r.$$

[12] *reziduální součet čtverců* [13] *reziduální střední čtverec* [14] *reziduální stupně volnosti*

2.5 Parameterizations of a linear model

For given response $\boldsymbol{Y} = \bigl(Y_1, \dots, Y_n\bigr)^\top$ and given set of covariates $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$, many different sets of regressors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ 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.

Definition 2.7 Equivalent linear models.

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

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

Notes.

- The two equivalent models:
  - have the same hat matrix $\mathbb{H} = \mathbb{X}_1\bigl(\mathbb{X}_1^\top\mathbb{X}_1\bigr)^{-}\mathbb{X}_1^\top = \mathbb{X}_2\bigl(\mathbb{X}_2^\top\mathbb{X}_2\bigr)^{-}\mathbb{X}_2^\top$ 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.

2.6 Matrix algebra and a method of least squares

In principle, any linear model $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$, even if being of less than full rank can be reparameterized such that the model matrix $\mathbb{X}$ has linearly independent columns. For instance, the orthonormal vector basis $\mathbb{Q}$ of the regression space provides such a model matrix. That is, for practical calculations, we can assume, without loss of generality that $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$. 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:

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

$$\begin{aligned}
\widehat{\boldsymbol{Y}} &= \mathbb{H}\boldsymbol{Y} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}, & \mathsf{var}\big(\widehat{\boldsymbol{Y}}\,\big|\,\mathbb{X}\big) &= \sigma^2\mathbb{H} = \sigma^2\,\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top,\\
\boldsymbol{U} &= \mathbb{M}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}, & \mathsf{var}\big(\boldsymbol{U}\,\big|\,\mathbb{X}\big) &= \sigma^2\mathbb{M} = \sigma^2\Big\{\mathbf{I}_n - \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\Big\},\\
\widehat{\boldsymbol{\beta}} &= \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}, & \mathsf{var}\big(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{X}\big) &= \sigma^2\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}.
\end{aligned}$$

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, 2020) 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

$$\mathbb{X}^\top\mathbb{X} = \boldsymbol{R}^\top\underbrace{\mathbb{Q}^\top\mathbb{Q}}_{\mathbf{I}_k}\boldsymbol{R} = \boldsymbol{R}^\top\boldsymbol{R}. \tag{2.5}$$

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 (2.5) can now be used to get easily (i) matrix $\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}$, (ii) a value of its determinant or a value of determinant of $\mathbb{X}^\top\mathbb{X}$, (iii) solution to normal equations.

(i) Matrix $\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}$.

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

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 $\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}$.
Let $r_1, \dots, r_k$ denote diagonal elements of the matrix $\boldsymbol{R}$. We then have

$$\det\bigl(\mathbb{X}^\top\mathbb{X}\bigr) = \det\bigl(\boldsymbol{R}^\top\boldsymbol{R}\bigr) = \bigl\{\det(\boldsymbol{R})\bigr\}^2 = \Bigl(\prod_{j=1}^{k} r_j\Bigr)^2,$$

$$\det\Bigl\{\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\Bigr\} = \Bigl\{\det\bigl(\mathbb{X}^\top\mathbb{X}\bigr)\Bigr\}^{-1}.$$

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

$$\begin{aligned} \mathbb{X}^\top\mathbb{X}\boldsymbol{b} &= \mathbb{X}^\top\boldsymbol{Y} \\ \boldsymbol{R}^\top\boldsymbol{R}\boldsymbol{b} &= \boldsymbol{R}^\top\mathbb{Q}^\top\boldsymbol{Y} \\ \boldsymbol{R}\boldsymbol{b} &= \mathbb{Q}^\top\boldsymbol{Y}. \end{aligned} \tag{2.6}$$

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 $\boldsymbol{c} = \bigl(c_1, \dots, c_k\bigr)^\top := \mathbb{Q}^\top\boldsymbol{Y}$ of the linear system (2.6) 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 $\mathcal{M}\bigl(\mathbb{X}\bigr)$ 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**

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 $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $\boldsymbol{Z}_i = \bigl(Z_{i,1},\, \ldots,\, Z_{i,p}\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$. We keep considering that the principal aim of the statistical analysis is to find a suitable model to express the (conditional) response expectation $\mathbb{E}\bigl(Y_i \bigm| \boldsymbol{Z}_i\bigr)$, $i = 1, \ldots, n$, in summary the response vector conditional expectation $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr)$, 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

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

3.1 (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 $\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z}\bigr) = \boldsymbol{t}^\top(\mathbb{Z})\boldsymbol{\beta} = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$ and $\mathsf{var}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z}\bigr) = \sigma^2\,\mathbf{I}_n$. For the data $\bigl(Y_i,\,\boldsymbol{X}_i^\top\bigr)^\top$, $i = 1,\dots,n$, where we directly work with the response-regressors pairs, this means the following assumptions ($i = 1,\dots,n$):

(A1) $\mathbb{E}\bigl(Y_i\,\big|\,\boldsymbol{X}_i = \boldsymbol{x}\bigr) = \boldsymbol{x}^\top\boldsymbol{\beta}$ 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) $\mathsf{var}\bigl(Y_i\,\big|\,\boldsymbol{X}_i = \boldsymbol{x}\bigr) = \sigma^2$ 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) $\mathsf{cov}\bigl(Y_i,\,Y_l\,\big|\,\mathbb{X} = \mathbf{x}\bigr) = 0$, $i \neq l$, for (almost all) $\mathbf{x} \in \mathcal{X}^n$.

$\equiv$ The responses are conditionally uncorrelated.

Some of our results (especially those shown in Chapter 6) will be derived while additionally assuming normality of the response, i.e., while assuming

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

$\equiv$ Normality of the response.

If we take the error terms of the linear model, i.e., the vector $\bigl(\varepsilon_1,\,\dots,\,\varepsilon_n\bigr)^\top = \boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} = \bigl(Y_1 - \boldsymbol{X}_1^\top\boldsymbol{\beta},\,\dots,\,Y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\bigr)^\top$, 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) $\mathbb{E}\bigl(\varepsilon_i\,\big|\,\boldsymbol{X}_i = \boldsymbol{x}\bigr) = 0$ for (almost all) $\boldsymbol{x} \in \mathcal{X}$, and consequently also $\mathbb{E}\bigl(\varepsilon_i\bigr) = 0$, $i = 1,\dots,n$.

$\equiv$ This again means that a structural part of the model stating that $\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr) = \mathbb{X}\boldsymbol{\beta}$ 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) $\mathsf{var}\bigl(\varepsilon_i\,\big|\,\boldsymbol{X}_i = \boldsymbol{x}\bigr) = \sigma^2$ for some σ^2 which is constant irrespective of (almost all) values of $\boldsymbol{x} \in \mathcal{X}$. Consequently also $\mathsf{var}\bigl(\varepsilon_i\bigr) = \sigma^2$, $i = 1,\dots,n$.

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

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

$\equiv$ The errors are uncorrelated.

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

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

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

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 16 that

[1] *homoskedasticita*

also majority of inferential tools based on results of Chapters 6 and 8 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 9 and 11.

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 $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big) \in \mathcal{M}\big(\mathbb{X}\big)$. 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

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

When doing so, it is exploited that each of assumptions (A1)–(A4) implies a certain property of the residuals stated earlier in Lemma 2.7 (Moments of residuals and residual sum of squares) and Theorem 6.2 (Properties of the LSE under the normality). It follows from those theorems (or their proofs) the following:

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

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.

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

[2] *regresní diagnostika*

3.2 Standardized residuals

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, 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 $\big(m_{1,1}, \ldots, m_{n.n}\big)^\top$. The following definition is motivated by the facts following the properties of residuals shown in Lemma 2.7:

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

Definition 3.1 Standardized residuals.

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

$$U_i^{std} = \begin{cases} \dfrac{U_i}{\sqrt{\mathsf{MS}_e\,m_{i,i}}}, & m_{i,i} > 0, \\[2ex] \textit{undefined}, & m_{i,i} = 0, \end{cases} \qquad\qquad i = 1, \ldots, n.$$

Lemma 3.1 Moments of standardized residuals under normality.

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

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

Proof. **The proof is not requested for exam.**

For each $i = 1,\, \ldots,\, n$ for which $m_{i,i} > 0$:

$$U_i^{std} = \frac{U_i}{\sqrt{\mathsf{MS}_e\,m_{i,i}}} = \frac{U_i}{\sqrt{\mathsf{SS}_e}}\sqrt{\frac{n-r}{m_{i,i}}} = \frac{U_i}{\|\boldsymbol{U}\|}\sqrt{\frac{n-r}{m_{i,i}}}.$$

Further, $\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = \mathbb{N}\mathbb{N}^\top\boldsymbol{Y}$, where $\mathbb{N}_{n\times(n-r)}$ is a matrix with orthonormal basis of the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$ in its columns.

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

$$\mathbb{N}^\top\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_{n-r}\big(\underbrace{\mathbb{N}^\top\mathbb{X}}_{\mathbf{0}_{n-r}}\boldsymbol{\beta},\, \sigma^2\,\underbrace{\mathbb{N}^\top\mathbb{N}}_{\mathbf{I}_{n-r}}\big).$$

That is, $\mathbb{N}^\top\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_{n-r}\big(\mathbf{0},\, \sigma^2\mathbf{I}_{n-r}\big)$.

We now use Lemma B.2 with $\boldsymbol{Z} = \mathbb{N}^\top\boldsymbol{Y}$ and function T defined as

$$T(\boldsymbol{Z}) := \frac{\boldsymbol{j}_i^\top\mathbb{N}\boldsymbol{Z}}{\|\mathbb{N}\boldsymbol{Z}\|} = \frac{\boldsymbol{j}_i^\top\mathbb{N}\mathbb{N}^\top\boldsymbol{Y}}{\|\mathbb{N}\mathbb{N}^\top\boldsymbol{Y}\|} = \frac{U_i}{\|\boldsymbol{U}\|},$$

[3] *standardizovaná rezidua*

where $\boldsymbol{j}_i = \bigl(0, \ldots, 1, \ldots, 0\bigr)^\top$ is a unity vector.

It is easily seen that for any $c > 0$

$$T(c\boldsymbol{Z}) = \frac{\boldsymbol{j}_i^\top \mathbb{N}\boldsymbol{Z}\,c}{\|c\,\mathbb{N}\boldsymbol{Z}\|} = \frac{\boldsymbol{j}_i^\top \mathbb{N}\boldsymbol{Z}}{\|\mathbb{N}\boldsymbol{Z}\|} = T(\boldsymbol{Z}).$$

Hence, by Lemma B.2, random variables

$$T(\boldsymbol{Z}) = \frac{U_i}{\|\boldsymbol{U}\|}$$

$$\text{and} \quad \|\boldsymbol{Z}\| = \|\mathbb{N}^\top\boldsymbol{Y}\| = \sqrt{\boldsymbol{Y}^\top\mathbb{N}\mathbb{N}^\top\boldsymbol{Y}} = \sqrt{\boldsymbol{Y}^\top\mathbb{M}\boldsymbol{Y}} = \sqrt{\boldsymbol{Y}^\top\mathbb{M}^\top\mathbb{M}\boldsymbol{Y}} = \sqrt{\boldsymbol{U}^\top\boldsymbol{U}} = \|\boldsymbol{U}\|$$

are independent, in our case conditionally given $\mathbb{X}$.

We use this independence in calculation of the following (conditional) expectations. At the same time, we shall use known moments of a (raw) residual U_i. First,

$$0 \overset{\text{Lemma 2.7}}{=} \mathbb{E}\bigl(U_i \bigm| \mathbb{X}\bigr) = \mathbb{E}\biggl(\frac{U_i}{\|\boldsymbol{U}\|}\,\|\boldsymbol{U}\| \biggm| \mathbb{X}\biggr) \overset{\text{indep.}}{=} \mathbb{E}\biggl(\frac{U_i}{\|\boldsymbol{U}\|} \biggm| \mathbb{X}\biggr)\,\mathbb{E}\bigl(\|\boldsymbol{U}\| \bigm| \mathbb{X}\bigr). \tag{3.1}$$

Under normality assumption, $\frac{1}{\sigma^2}\,\|\boldsymbol{U}\|^2 \Bigm| \mathbb{X} \sim \chi^2_{n-r}$ which is a *continuous* distribution whose support is $(0,\,\infty)$. Hence, $\mathsf{P}\Bigl(\|\boldsymbol{U}\|^2 = 0 \Bigm| \mathbb{X}\Bigr) = 0$ and consequently, also $\mathsf{P}\bigl(\|\boldsymbol{U}\| = 0 \bigm| \mathbb{X}\bigr) = 0$. This implies that $\mathbb{E}\bigl(\|\boldsymbol{U}\| \bigm| \mathbb{X}\bigr) > 0$. From the relationship (3.1) we now conclude that

$$\mathbb{E}\biggl(\frac{U_i}{\|\boldsymbol{U}\|} \biggm| \mathbb{X}\biggr) = 0.$$

Finally, we can calculate

$$\mathbb{E}\bigl(U_i^{std} \bigm| \mathbb{X}\bigr) = \mathbb{E}\biggl(\frac{U_i}{\|\boldsymbol{U}\|}\,\sqrt{\frac{n-r}{m_{i,i}}} \biggm| \mathbb{X}\biggr) = \sqrt{\frac{n-r}{m_{i,i}}}\,\mathbb{E}\biggl(\frac{U_i}{\|\boldsymbol{U}\|} \biggm| \mathbb{X}\biggr) = 0.$$

Second,

$$\begin{aligned}\sigma^2\,m_{i,i} \overset{\text{Lemma 2.7}}{=} \mathsf{var}\bigl(U_i \bigm| \mathbb{X}\bigr) = \mathbb{E}\bigl(U_i^2 \bigm| \mathbb{X}\bigr) &= \mathbb{E}\biggl(\frac{U_i^2}{\|\boldsymbol{U}\|^2}\,\|\boldsymbol{U}\|^2 \biggm| \mathbb{X}\biggr)\\ &\overset{\text{indep.}}{=} \mathbb{E}\biggl(\frac{U_i^2}{\|\boldsymbol{U}\|^2} \biggm| \mathbb{X}\biggr)\,\mathbb{E}\Bigl(\|\boldsymbol{U}\|^2 \Bigm| \mathbb{X}\Bigr).\end{aligned} \tag{3.2}$$

As mentioned above, under normality assumptions, $\frac{1}{\sigma^2}\,\|\boldsymbol{U}\|^2 \Bigm| \mathbb{X} \sim \chi^2_{n-r}$. Hence, $\mathbb{E}\Bigl(\|\boldsymbol{U}\|^2 \Bigm| \mathbb{X}\Bigr) = \sigma^2\,(n-r)$. Relationship (3.2) now implies

$$\sigma^2\,m_{i,i} = \sigma^2\,(n-r)\,\mathbb{E}\biggl(\frac{U_i^2}{\|\boldsymbol{U}\|^2} \biggm| \mathbb{X}\biggr).$$

That is,

$$\mathbb{E}\biggl(\frac{U_i^2}{\|\boldsymbol{U}\|^2} \biggm| \mathbb{X}\biggr) = \frac{m_{i,i}}{n-r}.$$

Finally,

$$\mathsf{var}\bigl(U_i^{std}\,\big|\,\mathbb{X}\bigr) = \mathbb{E}\Bigl((U_i^{std})^2\,\Big|\,\mathbb{X}\Bigr) = \mathbb{E}\left(\frac{U_i^2}{\|\boldsymbol{U}\|^2}\,\frac{n-r}{m_{i,i}}\,\middle|\,\mathbb{X}\right) = \frac{n-r}{m_{i,i}}\,\mathbb{E}\left(\frac{U_i^2}{\|\boldsymbol{U}\|^2}\,\middle|\,\mathbb{X}\right) = 1.$$

❏

Notes.

- Unfortunately, even in a normal linear model, the standardized residuals $U_1^{std}, \dots, U_n^{std}$ 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 11.

[4] *studentizovaná rezidua*

3.3 Graphical tools of regression diagnostics

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

$$\mathbb{X} = \bigl(\boldsymbol{X}^0,\, \dots,\, \boldsymbol{X}^{k-1}\bigr).$$

Remember that usually $\boldsymbol{X}^0 = \bigl(1, \dots, 1\bigr)^\top$ is an intercept column. Further, in many situations, see Section 8.2 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 8.2) and perhaps additional regressors are available to model the response expectation $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr)$. Let us denote them as $\boldsymbol{V}^1, \dots, \boldsymbol{V}^m$. That is, the matrix

$$\mathbb{V} := \bigl(\boldsymbol{V}^1,\, \dots,\, \boldsymbol{V}^m\bigr).$$

may play a role of omitted covariates (matrix $\mathbb{X}^1$ in the notation of Section 8.2).

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.

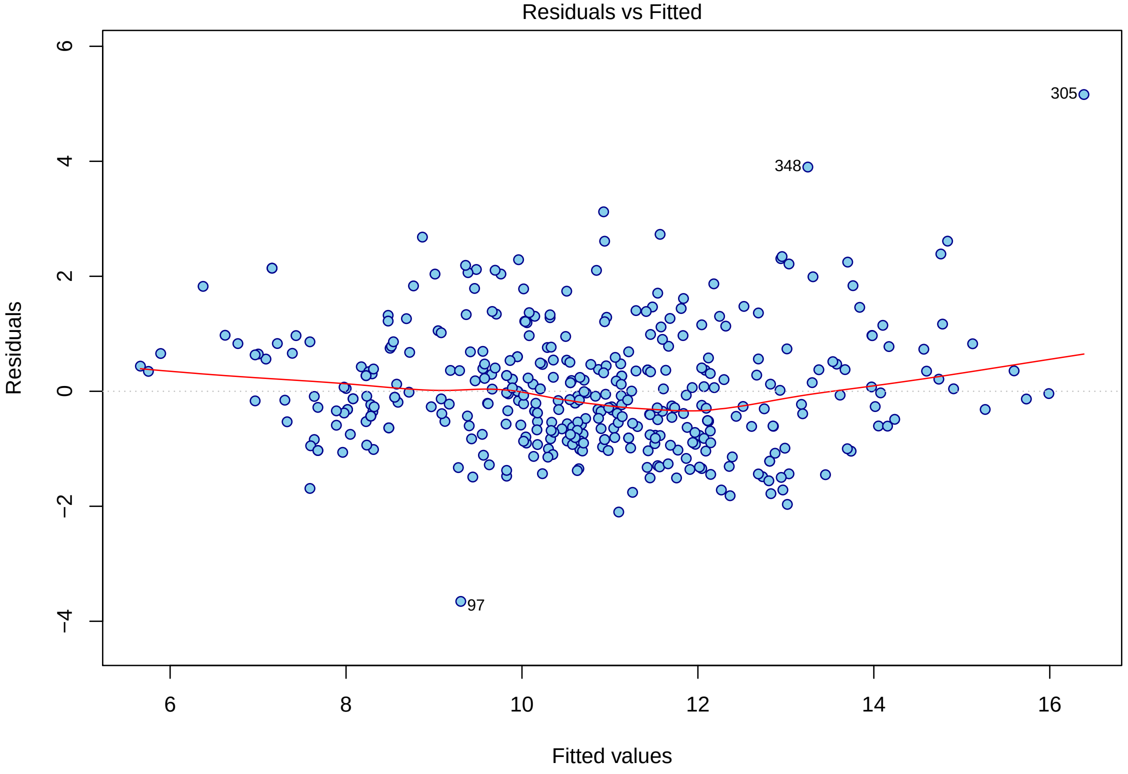
3.3.1 (A1) Correctness of the regression function

To detect:

Overall inappropriateness of the regression function

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

Illustrations



Nonlinearity of the regression function with respect to a particular regressor X^j

⇒ scatterplot $\bigl(X^j, U\bigr)$ of residuals versus that regressor.

Possibly omitted regressor V

⇒ scatterplot $\bigl(V, U\bigr)$ of residuals versus that regressor.

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

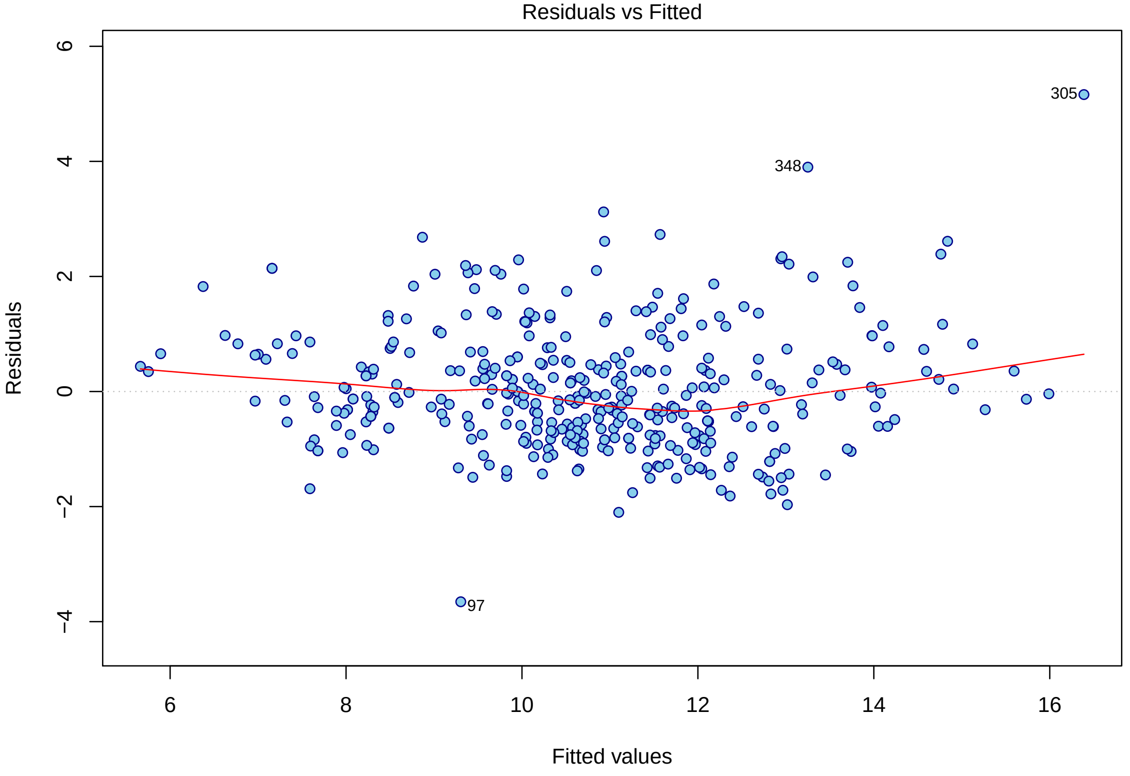
3.3.2 (A2) Homoscedasticity of the errors

To detect

Residual variance that depends on the response expectation

⇒ scatterplot $\bigl(\widehat{Y}, U\bigr)$ of residuals versus fitted values.

Illustrations



Residual variance that depends on a particular regressor X^j

⇒ scatterplot $\bigl(X^j, U\bigr)$ of residuals versus that regressor.

Residual variance that depends on a regressor V not included in the model

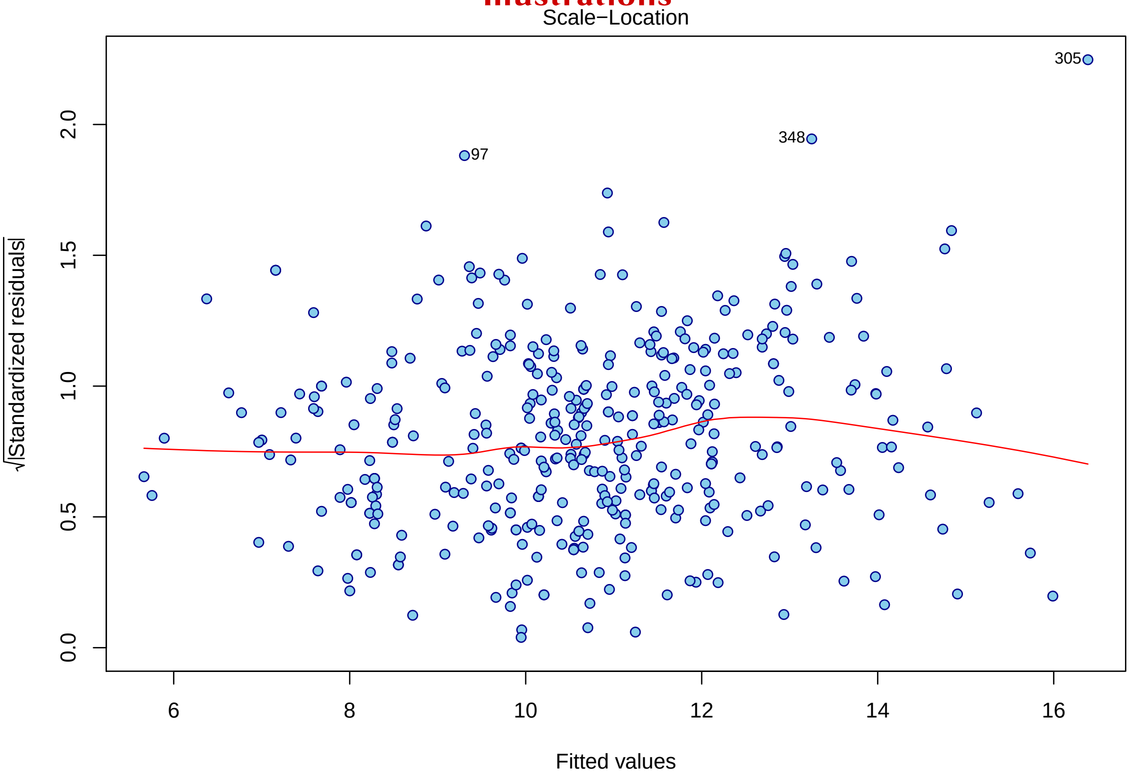
⇒ scatterplot $\bigl(V, U\bigr)$ of residuals versus that regressor.

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 ($\mathsf{var}\bigl(\boldsymbol{U}\,\big|\,\mathbb{Z}\bigr) = \sigma^2\,\mathbb{M}$), 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

$$\Bigl(\sqrt{|U_1^{std}|},\,\dots,\,\sqrt{|U_n^{std}|}\Bigr).$$

Illustrations



3.3.3 (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.

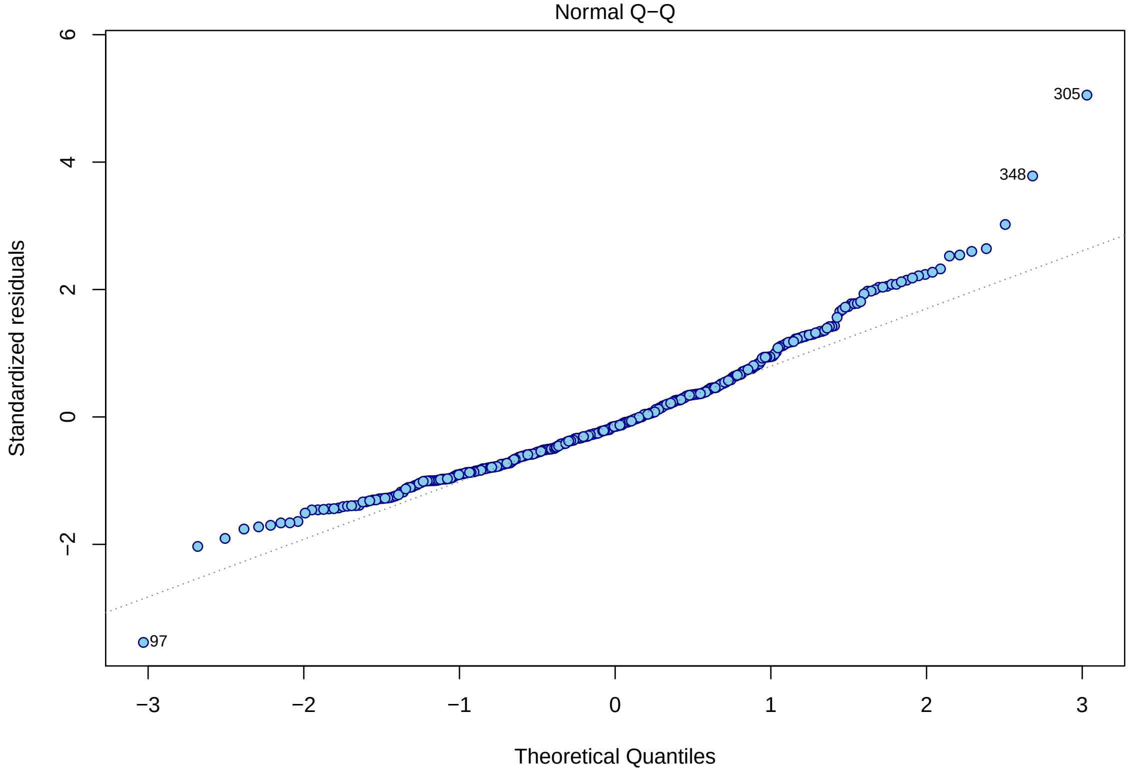
As stated above, even if the errors are uncorrelated and assumption (A3) is fulfilled, the residuals $\boldsymbol{U}$ are in general correlated. Nevertheless, the correlation is usually rather low and the residuals are typically used to check assumption (A3) and possibly to detect a form of the serial 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 $(U_1,\, U_2)$, $(U_2,\, U_3)$, …, $(U_{n-1},\, U_n)$.

3.3.4 (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).

Illustrations



[5] *časová řada* [6] *bílý šum*

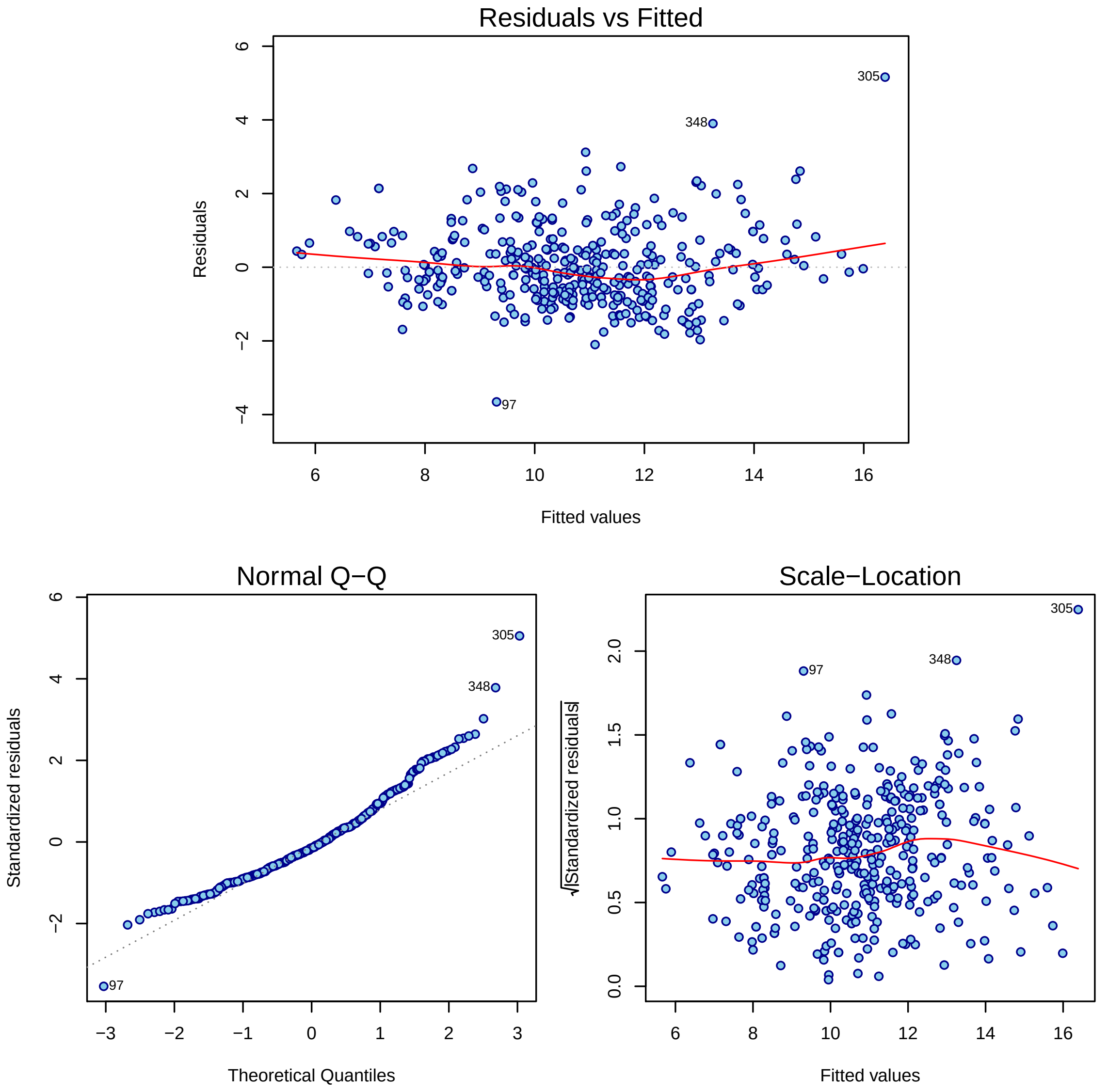
Usage of both the raw residuals $\boldsymbol{U}$ and the standardized residuals $\boldsymbol{U}^{std}$ to check the normality assumption (A4) bears certain inconveniences. If all assumptions of the normal linear model are fulfilled, then

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

The standardized residuals $\boldsymbol{U}^{std}$ satisfy $\mathbb{E}\bigl(U_i^{std} \,\big|\, \mathbb{Z}\bigr) = 0$, $\mathsf{var}\bigl(U_i^{std} \,\big|\, \mathbb{Z}\bigr) = 1$ for all $i = 1, \dots, n$. 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.

3.3.5 The three basic diagnostic plots

Illustrations

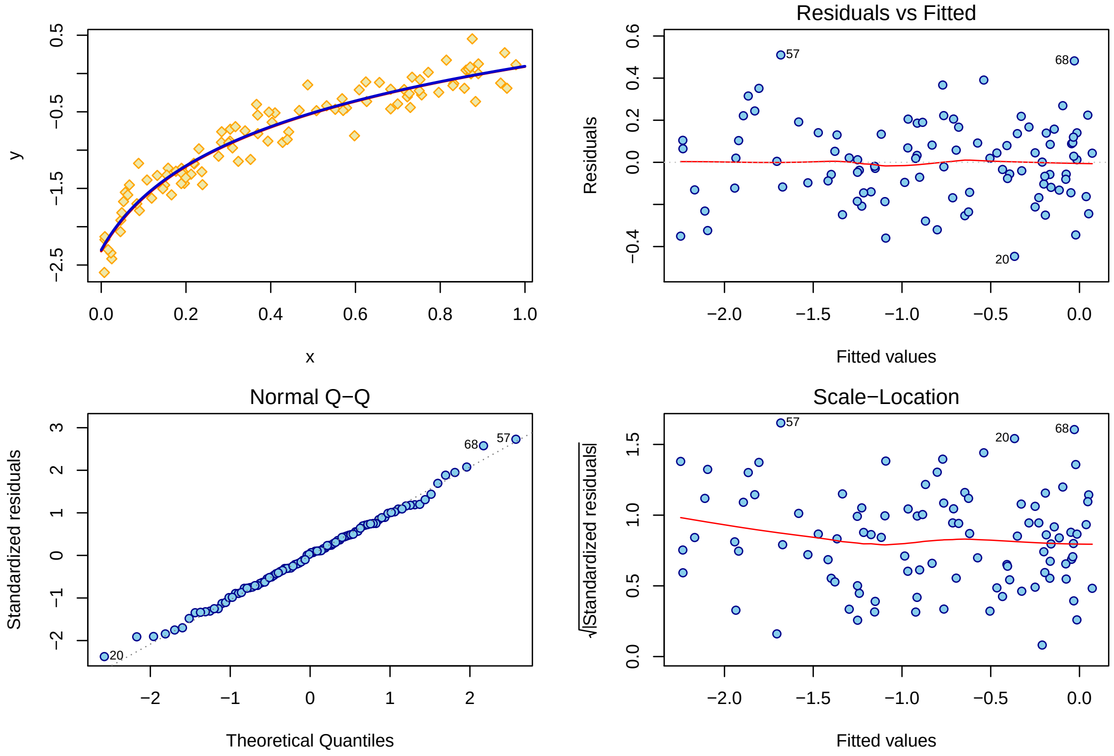


Illustrations

Correct model

True: $Y = \log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, 0.2^2).$

Model: $Y = \beta_0 + \beta_1\,\log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, \sigma^2).$

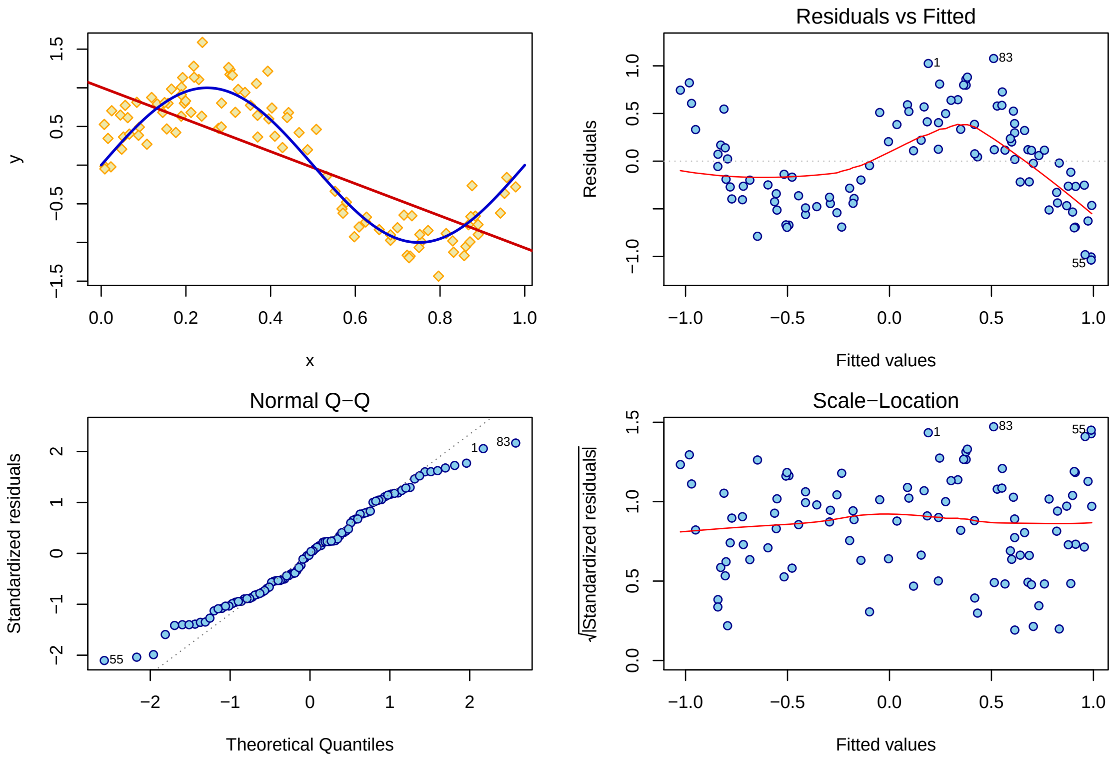


Illustrations

Incorrect regression function

True: $Y = \sin(2\pi\, x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, 0.3^2).$

Model: $Y = \beta_0 + \beta_1\, x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, \sigma^2).$

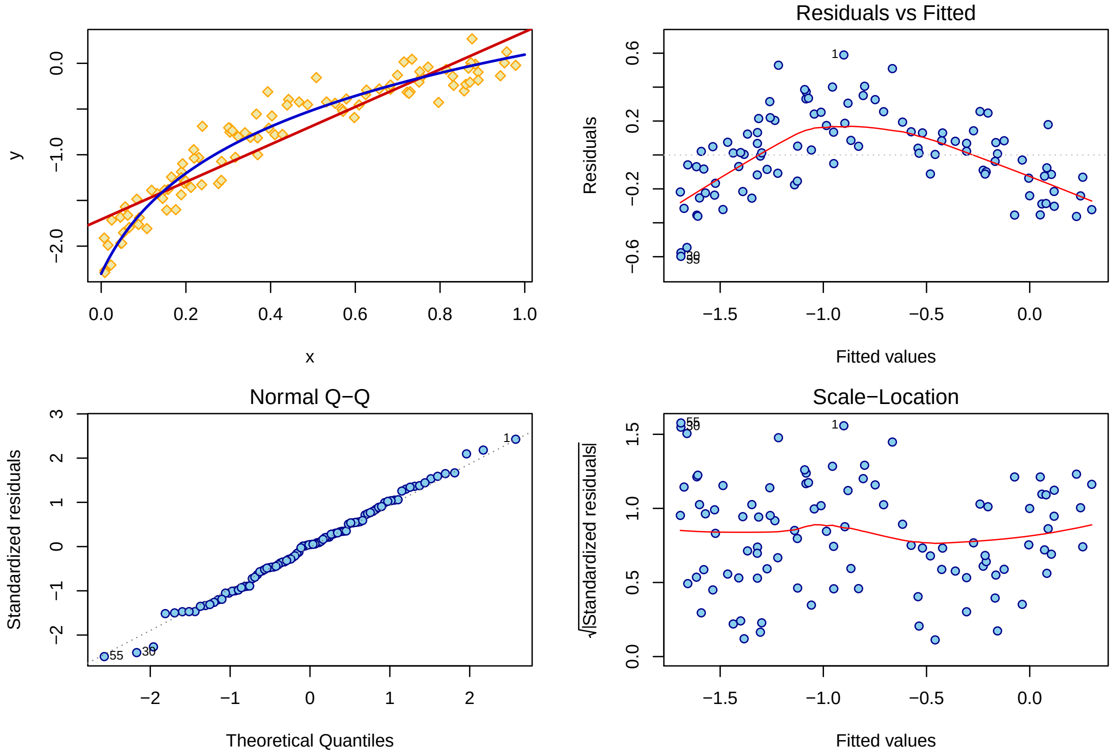


Illustrations

Incorrect regression function

True: $Y = \log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, 0.2^2).$

Model: $Y = \beta_0 + \beta_1\, x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, \sigma^2).$

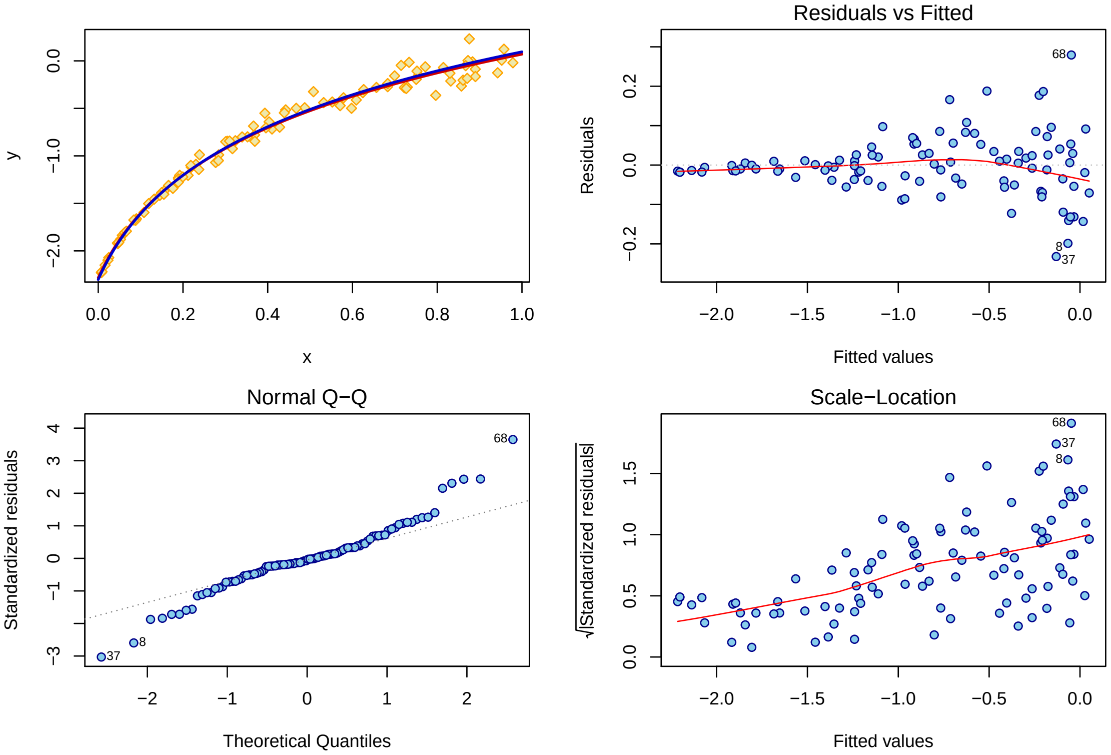


Illustrations

Heteroscedasticity

True: $Y = \log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\ (0.2\,x)^2).$

Model: $Y = \beta_0 + \beta_1\ \log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\ \sigma^2).$

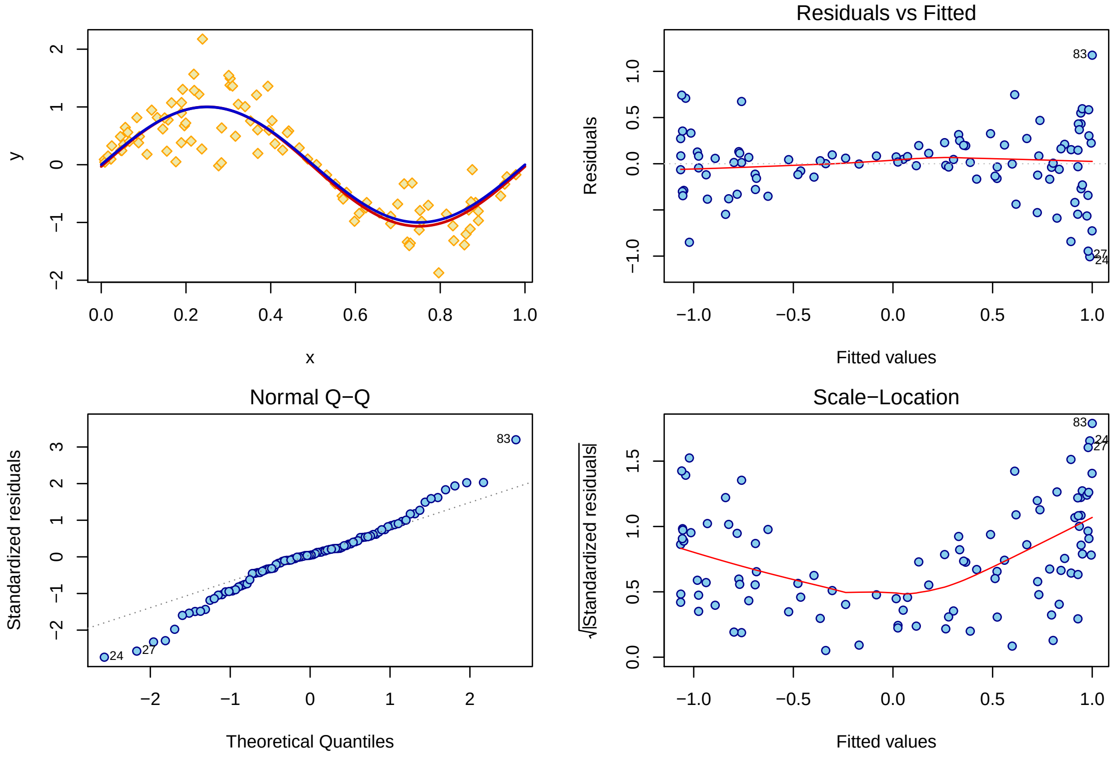


Illustrations

Heteroscedasticity

True: $Y = \sin(2\pi\, x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, \{0.6\, \sin(2\pi\, x)\}^2).$

Model: $Y = \beta_0 + \beta_1\, \sin(2\pi\, x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, \sigma^2).$

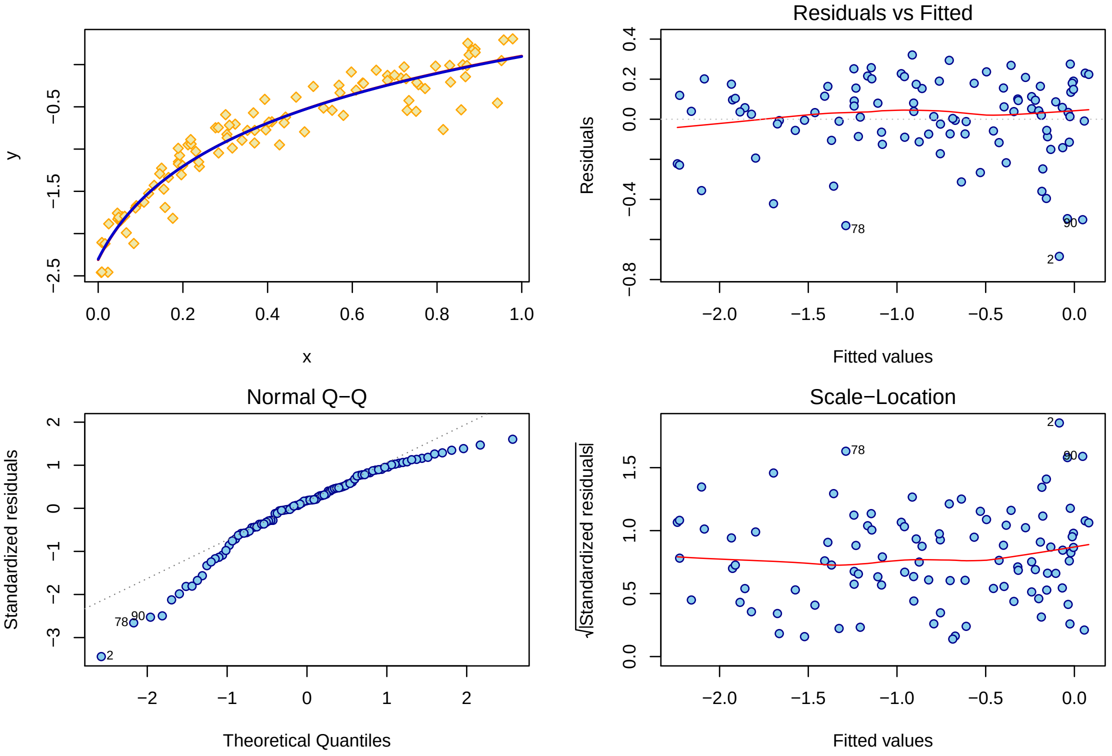


Illustrations

Nonnormal errors

True: $Y = \log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \text{Gumbel}.$

Model: $Y = \beta_0 + \beta_1 \log(0.1 + x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$



Chapter **4**

Parameterizations of Covariates

4.1 Linearization of the dependence of the response on the covariates

As it is usual in this lecture, we represent data by n random vectors $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $\boldsymbol{Z}_i = \bigl(Z_{i,1},\, \ldots,\, Z_{i,p}\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$. The principal problem we consider is to find a suitable model to express the (conditional) response expectation $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr)$, where $\boldsymbol{Y} = \bigl(Y_1,\, \ldots,\, Y_n\bigr)^\top$ and $\mathbb{Z}$ is a matrix with vectors $\boldsymbol{Z}_1,\, \ldots,\, \boldsymbol{Z}_n$ in its rows. To this end, we consider a linear model, where $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr)$ can be expressed as $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$, where

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}, \qquad \begin{array}{rcl} \boldsymbol{X}_1 = \bigl(X_{1,0},\, \ldots,\, X_{1,k-1}\bigr)^\top & = & \boldsymbol{t}(\boldsymbol{Z}_1), \\ & \vdots & \\ \boldsymbol{X}_n = \bigl(X_{n,0},\, \ldots,\, X_{n,k-1}\bigr)^\top & = & \boldsymbol{t}(\boldsymbol{Z}_n), \end{array}$$

and $\boldsymbol{t} : \mathcal{Z} \longrightarrow \mathcal{X} \subseteq \mathbb{R}^k$, $\boldsymbol{t}(\boldsymbol{z}) = \bigl(t_0(\boldsymbol{z}),\, \ldots,\, t_{k-1}(\boldsymbol{z})\bigr)^\top = \bigl(x_0,\, \ldots,\, x_{k-1}\bigr)^\top = \boldsymbol{x}$, 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

$$m(\boldsymbol{z}) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} = \beta_0\, t_0(\boldsymbol{z}) + \cdots + \beta_{k-1}\, t_{k-1}(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathcal{Z}. \tag{4.1}$$

One of main problems of any practical 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 $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr)$ and in general to express $\mathbb{E}\bigl(Y \bigm| \boldsymbol{Z} = \boldsymbol{z}\bigr)$, $\boldsymbol{z} \in \mathcal{Z}$, for a generic response Y being generated, given the covariate value $\boldsymbol{Z} = \boldsymbol{z}$, by the same probabilistic mechanism as the original data.

4.2 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.

4.2.1 Parameterization

Our aim is to propose transformations $\boldsymbol{t} : \mathcal{Z} \longrightarrow \mathbb{R}^k$, $\boldsymbol{t}(z) = \big(t_0(z),\,\dots,\,t_{k-1}(z)\big)^\top$ such that a regression function (4.1) 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

$$m(z) = \beta_0 + \beta_1\, s_1(z) + \cdots + \beta_{k-1}\, s_{k-1}(z), \quad z \in \mathcal{Z}, \tag{4.2}$$

where the non-intercept part of the transformation $\boldsymbol{t}$ will be denoted as $\boldsymbol{s}$. That is, for $z \in \mathcal{Z}$,

$$s_j(z) = t_j(z), \qquad j = 1, \dots, k-1,$$

$$\boldsymbol{s} : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}, \quad \boldsymbol{s}(z) = \big(s_1(z),\,\dots,\,s_{k-1}(z)\big)^\top = \big(t_1(z),\,\dots,\,t_{k-1}(z)\big)^\top.$$

Definition 4.1 Parameterization of a covariate.
Let $Z_1,\,\dots,\,Z_n$ be values of a given univariate covariate $Z \in \mathcal{Z} \subseteq \mathbb{R}$. By a parameterization *of this covariate we mean*

(i) the function $\boldsymbol{s} : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}$, $\boldsymbol{s}(z) = \big(s_1(z),\,\dots,\,s_{k-1}(z)\big)^\top$, $z \in \mathcal{Z}$, where all $s_1,\,\dots,\,s_{k-1}$ are non-constant functions on $\mathcal{Z}$, and

(ii) an $n \times (k-1)$ matrix $\mathbb{S}$, where

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{s}^\top(Z_1) \\ \vdots \\ \boldsymbol{s}^\top(Z_n) \end{pmatrix} = \begin{pmatrix} s_1(Z_1) & \dots & s_{k-1}(Z_1) \\ \vdots & \vdots & \vdots \\ s_1(Z_n) & \dots & s_{k-1}(Z_n) \end{pmatrix}.$$

Terminology *(Reparameterizing matrix, regressors).*

Matrix $\mathbb{S}$ from Definition 4.1 is called *reparameterizing matrix*[1] of a covariate. Its columns, i.e., vectors

$$\boldsymbol{X}^1 = \begin{pmatrix} s_1(Z_1) \\ \vdots \\ s_1(Z_n) \end{pmatrix}, \;\dots,\; \boldsymbol{X}^{k-1} = \begin{pmatrix} s_{k-1}(Z_1) \\ \vdots \\ s_{k-1}(Z_n) \end{pmatrix}$$

determine the *regressors* of the linear model based on the covariate values $Z_1,\,\dots,\,Z_n$.

[1] *reparametrizační matice*

Notes.

- A model matrix $\mathbb{X}$ of the model with the regression function (4.2) is

$$\mathbb{X} = \Bigl(\mathbf{1}_n,\ \ \mathbb{S}\Bigr) = \Bigl(\mathbf{1}_n,\ \boldsymbol{X}^1,\ \dots,\ \boldsymbol{X}^{k-1}\Bigr) = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,k-1} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{X}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{X}_n^\top \end{pmatrix},$$

$$\boldsymbol{X}_i = \boldsymbol{s}(Z_i), \qquad X_{i,j} = s_j(Z_i),\ i = 1,\dots,n,\ j = 1,\dots,k-1.$$

- Definition 4.1 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 4.3.4), we will have $\mathbf{1}_n \in \mathcal{M}\bigl(\mathbb{S}\bigr)$.

4.2.2 Covariate types

The covariate space $\mathcal{Z}$ and the corresponding univariate covariates $Z_1,\dots,Z_n$ 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.,

$$\mathcal{Z} = \bigl\{\omega_1,\ \dots,\ \omega_G\bigr\},$$

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).

Illustrations

Cars2004nh ($n = 425$)

```
data(Cars2004nh, package = "mffSM")
head(Cars2004nh)
```

```
                        vname type drive price.retail price.dealer   price
1          Chevrolet.Aveo.4dr    1     1        11690        10965 11327.5
2 Chevrolet.Aveo.LS.4dr.hatch    1     1        12585        11802 12193.5
3      Chevrolet.Cavalier.2dr    1     1        14610        13697 14153.5
4      Chevrolet.Cavalier.4dr    1     1        14810        13884 14347.0
5   Chevrolet.Cavalier.LS.2dr    1     1        16385        15357 15871.0
6           Dodge.Neon.SE.4dr    1     1        13670        12849 13259.5

  cons.city cons.highway consumption engine.size ncylinder horsepower
1       8.4          6.9        7.65         1.6         4        103
2       8.4          6.9        7.65         1.6         4        103
3       9.0          6.4        7.70         2.2         4        140
4       9.0          6.4        7.70         2.2         4        140
5       9.0          6.4        7.70         2.2         4        140
6       8.1          6.5        7.30         2.0         4        132

  weight      iweight  lweight wheel.base length width    ftype fdrive
1   1075 0.0009302326 6.980076        249    424   168 personal  front
2   1065 0.0009389671 6.970730        249    389   168 personal  front
3   1187 0.0008424600 7.079184        264    465   175 personal  front
4   1214 0.0008237232 7.101676        264    465   173 personal  front
5   1187 0.0008424600 7.079184        264    465   175 personal  front
6   1171 0.0008539710 7.065613        267    442   170 personal  front
```

[5] *nominální* [6] *ordinální*

Illustrations

Cars2004nh ($n = 425$)

```
summary(subset(Cars2004nh,
   select = c("price.retail", "price.dealer", "price", "cons.city", "cons.highway",
              "consumption", "engine.size", "horsepower", "weight",
              "wheel.base", "length", "width")))
```

```
 price.retail     price.dealer        price          cons.city
Min.   : 10280   Min.   :  9875   Min.   : 10078   Min.   : 6.20
1st Qu.: 20370   1st Qu.: 18973   1st Qu.: 19600   1st Qu.:11.20
Median : 27905   Median : 25672   Median : 26656   Median :12.40
Mean   : 32866   Mean   : 30096   Mean   : 31481   Mean   :12.36
3rd Qu.: 39235   3rd Qu.: 35777   3rd Qu.: 37514   3rd Qu.:13.80
Max.   :192465   Max.   :173560   Max.   :183012   Max.   :23.50
                                                   NA's   :14

 cons.highway     consumption     engine.size      horsepower
Min.   : 5.100   Min.   : 5.65   Min.   :1.300   Min.   :100.0
1st Qu.: 8.100   1st Qu.: 9.65   1st Qu.:2.400   1st Qu.:165.0
Median : 9.000   Median :10.70   Median :3.000   Median :210.0
Mean   : 9.142   Mean   :10.75   Mean   :3.208   Mean   :216.8
3rd Qu.: 9.800   3rd Qu.:11.65   3rd Qu.:3.900   3rd Qu.:255.0
Max.   :19.600   Max.   :21.55   Max.   :8.300   Max.   :500.0
NA's   :14       NA's   :14

    weight       wheel.base        length          width
Min.   : 923   Min.   :226.0   Min.   :363.0   Min.   :163.0
1st Qu.:1412   1st Qu.:262.0   1st Qu.:450.0   1st Qu.:175.0
Median :1577   Median :272.0   Median :472.0   Median :180.0
Mean   :1626   Mean   :274.9   Mean   :470.6   Mean   :181.1
3rd Qu.:1804   3rd Qu.:284.0   3rd Qu.:490.0   3rd Qu.:185.0
Max.   :3261   Max.   :366.0   Max.   :577.0   Max.   :206.0
NA's   :2      NA's   :2       NA's   :26      NA's   :28
```

Illustrations

Cars2004nh ($n = 425$)

```
summary(subset(Cars2004nh, select = c("type", "drive")))
```

```
      type           drive      
 Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:1.000  
 Median :1.000   Median :1.000  
 Mean   :2.219   Mean   :1.692  
 3rd Qu.:3.000   3rd Qu.:2.000  
 Max.   :6.000   Max.   :3.000  
```

```
table(Cars2004nh[, "type"], useNA = "ifany")
```

```
  1   2   3   4   5   6 
242  30  60  24  49  20 
```

```
table(Cars2004nh[, "drive"], useNA = "ifany")
```

```
  1   2   3 
223 110  92 
```

Cars2004nh ($n = 425$)

```
summary(subset(Cars2004nh, select = c("ftype", "fdrive")))
```

```
      ftype       fdrive   
 personal:242   front:223  
 wagon   : 30   rear :110  
 SUV     : 60   4x4  : 92  
 pickup  : 24              
 sport   : 49              
 minivan : 20              
```

Illustrations

Cars2004nh ($n = 425$)

```
summary(subset(Cars2004nh, select = "ncylinder"))
```

```
  ncylinder
Min.   :-1.000
1st Qu.: 4.000
Median : 6.000
Mean   : 5.791
3rd Qu.: 6.000
Max.   :12.000
```

```
table(Cars2004nh[, "ncylinder"], useNA = "ifany")
```

```
-1   4   5   6   8  10  12
 2 134   7 190  87   2   3
```

4.3 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.

4.3.1 Simple transformation of the covariate

The regression function is

$$m(z) = \beta_0 + \beta_1 \, s(z), \quad z \in \mathcal{Z}, \tag{4.3}$$

where $s : \mathcal{Z} \longrightarrow \mathbb{R}$ is a suitable *non-constant* function. The corresponding reparameterizing matrix is

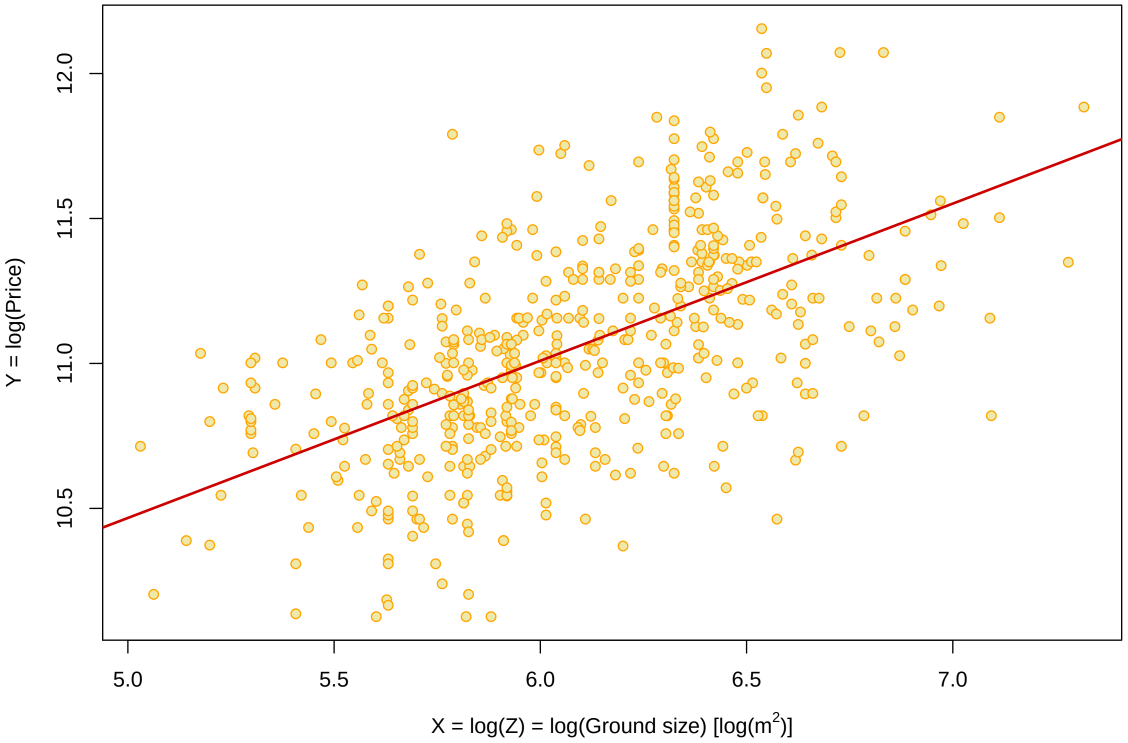
$$\mathbb{S} = \begin{pmatrix} s(Z_1) \\ \vdots \\ s(Z_n) \end{pmatrix}.$$

Due to interpretability issues, "simple" functions like: identity, logarithm, exponential, square root, reciprocal, . . . , are considered in place of the transformation s.

Illustrations

`Houses1987` ($n = 546$)

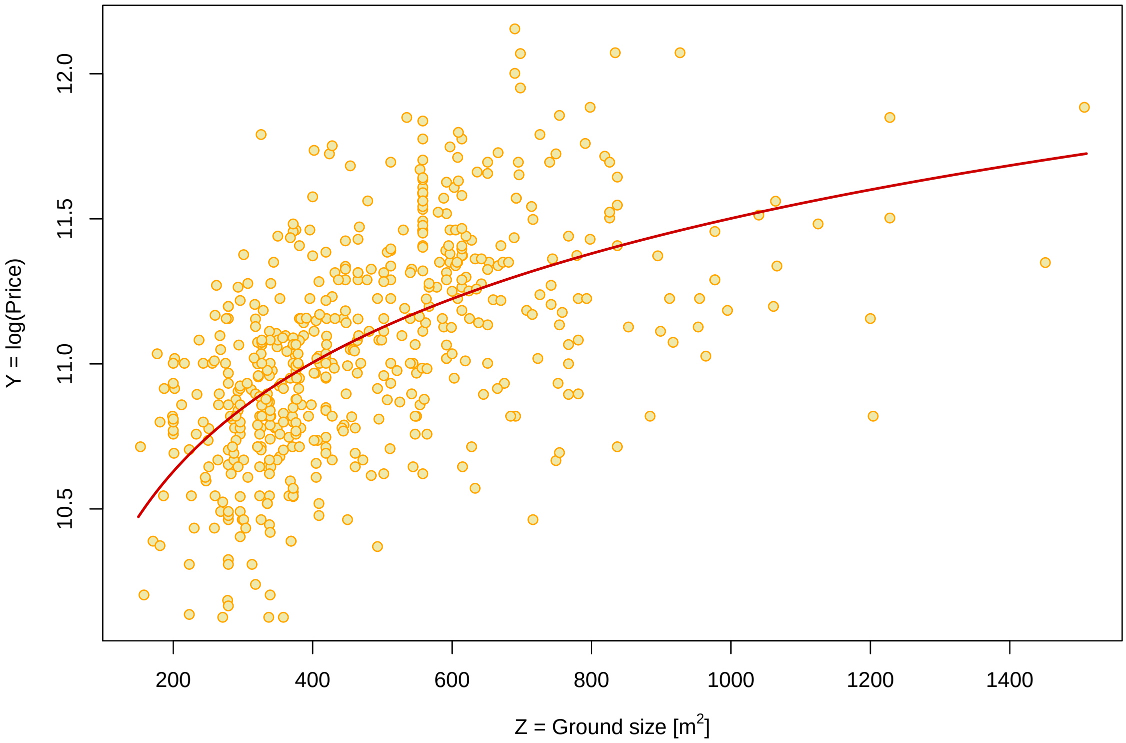
`log(price) ~ log(ground)`, $\widehat{m}(z) = 7.76 + 0.54 \log(z)$



Illustrations

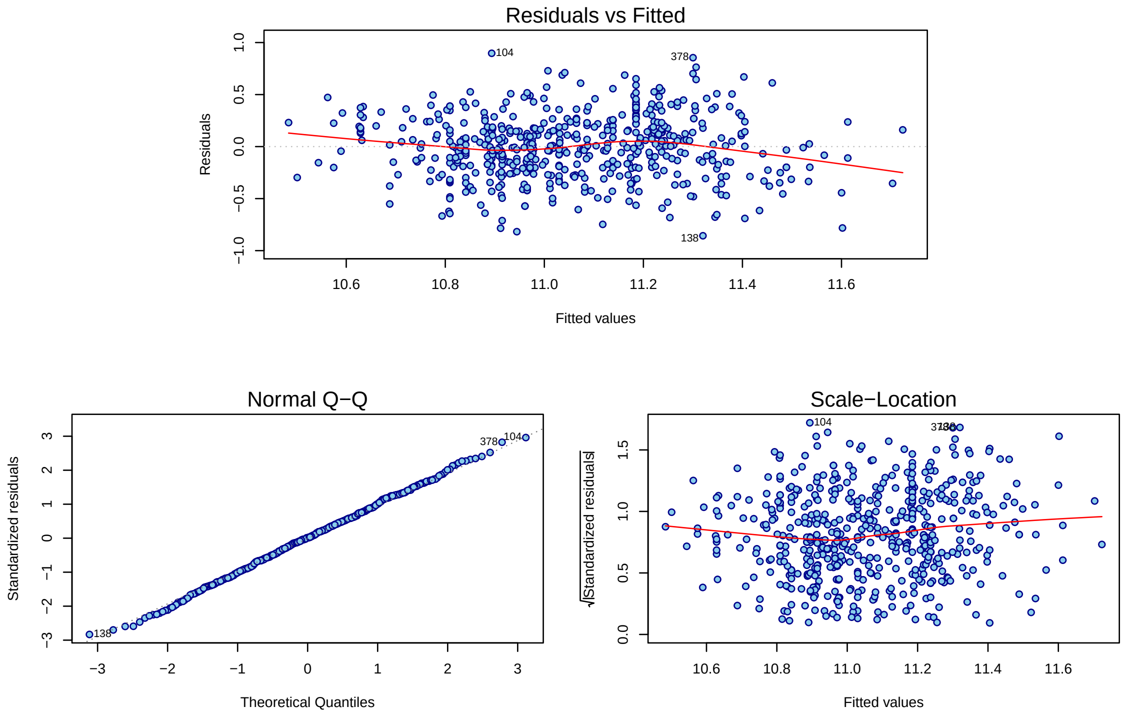
`Houses1987` ($n = 546$)

`log(price)` $\sim$ `log(ground)`, $\quad \widehat{m}(z) = 7.76 + 0.54\,\log(z)$



`Houses1987` ($n = 546$)

`log(price)` $\sim$ `log(ground)`, **residual plots**



Evaluation of the effect of the original covariate

Advantage of a model with the regression function (4.3) 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 will be introduced in Section 6.2.

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 (4.3) 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 $r = k = 2$.

Illustrations

`Houses1987` ($n = 546$)

Effect of the covariate, interpretation of the regression coefficients

```
summary(lm(log(price) ~ log(ground), data = Houses1987))
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.8571 -0.1988  0.0046  0.1929  0.8969 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.75625    0.19933   38.91   <2e-16 ***
log(ground)  0.54216    0.03265   16.61   <2e-16 ***
---

Residual standard error: 0.3033 on 544 degrees of freedom
Multiple R-squared:  0.3364,	Adjusted R-squared:  0.3351 
F-statistic: 275.7 on 1 and 544 DF,  p-value: < 2.2e-16
```

4.3.2 Raw polynomials

The regression function is polynomial of a chosen degree $k-1$, i.e.,

$$m(z) = \beta_0 + \beta_1 z + \cdots + \beta_{k-1} z^{k-1}, \quad z \in \mathcal{Z}. \tag{4.4}$$

The parameterization is

$$\boldsymbol{s} : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}, \qquad \boldsymbol{s}(z) = \bigl(z, \ldots, z^{k-1}\bigr)^\top, \; z \in \mathcal{Z}$$

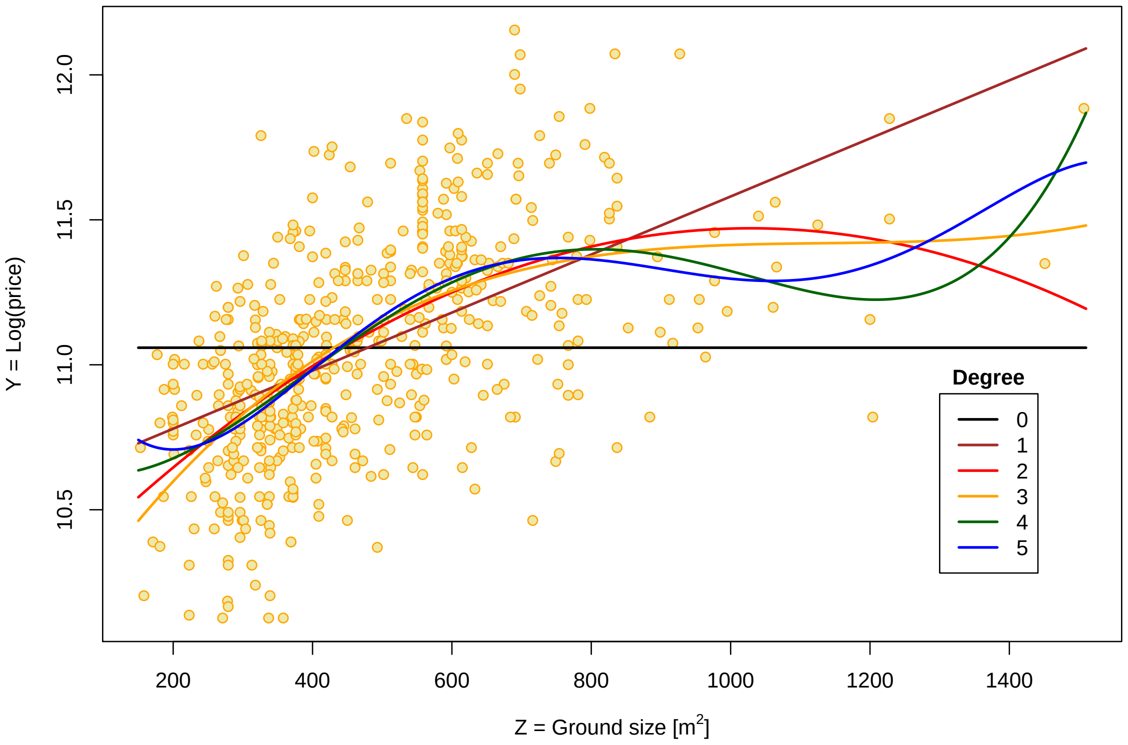
and the corresponding reparameterizing matrix is

$$\mathbb{S} = \begin{pmatrix} Z_1 & \ldots & Z_1^{k-1} \\ \vdots & \vdots & \vdots \\ Z_n & \ldots & Z_n^{k-1} \end{pmatrix}.$$

Illustrations

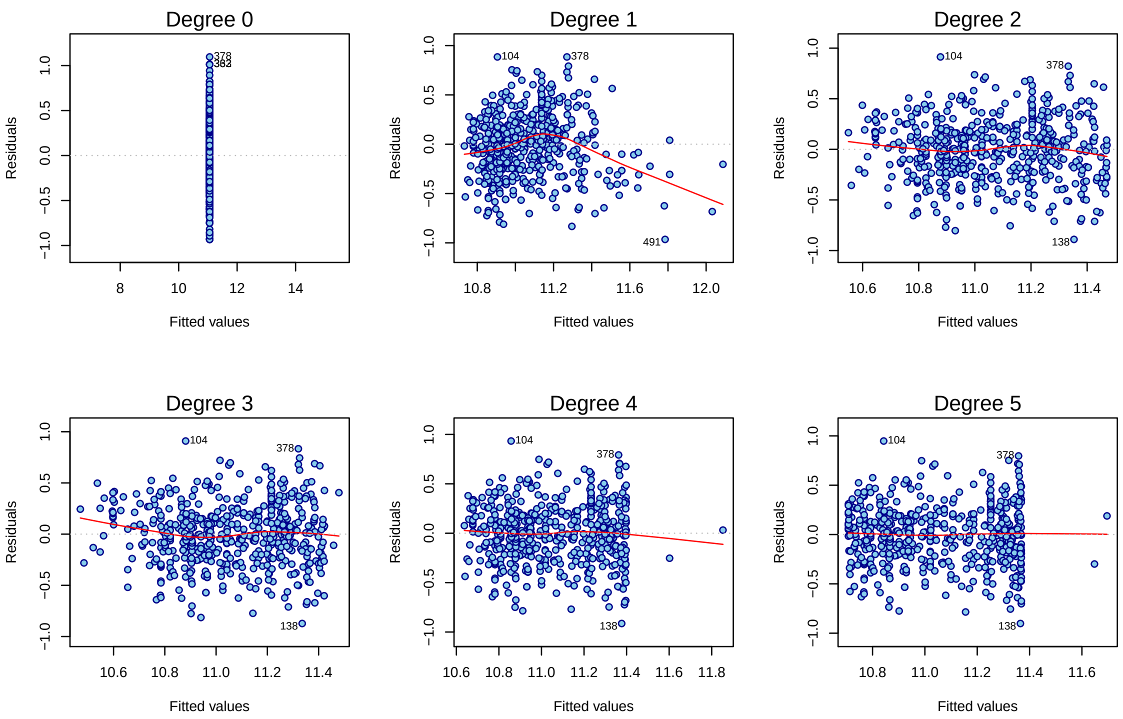
`Houses1987` ($n = 546$)

`log(price)` $\sim$ `rawpoly(ground, d)`



`Houses1987` ($n = 546$)

`log(price)` $\sim$ `rawpoly(ground, d)`, **residuals vs. fitted plots**

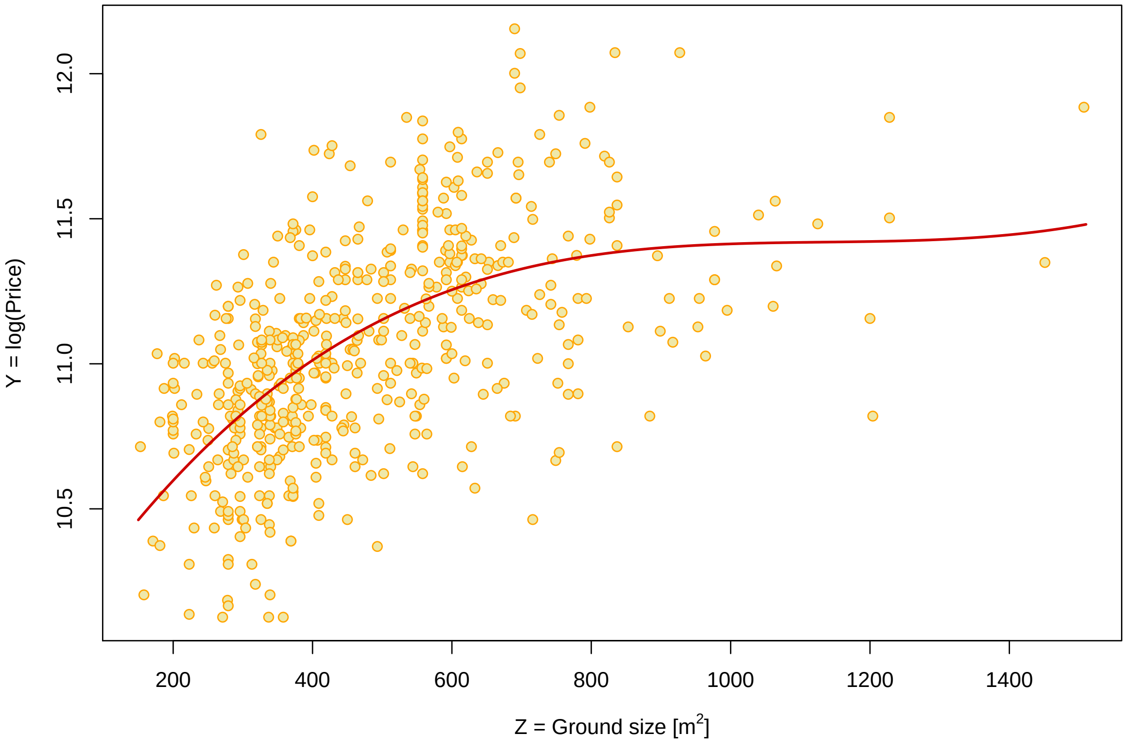


Illustrations

`Houses1987` ($n = 546$)

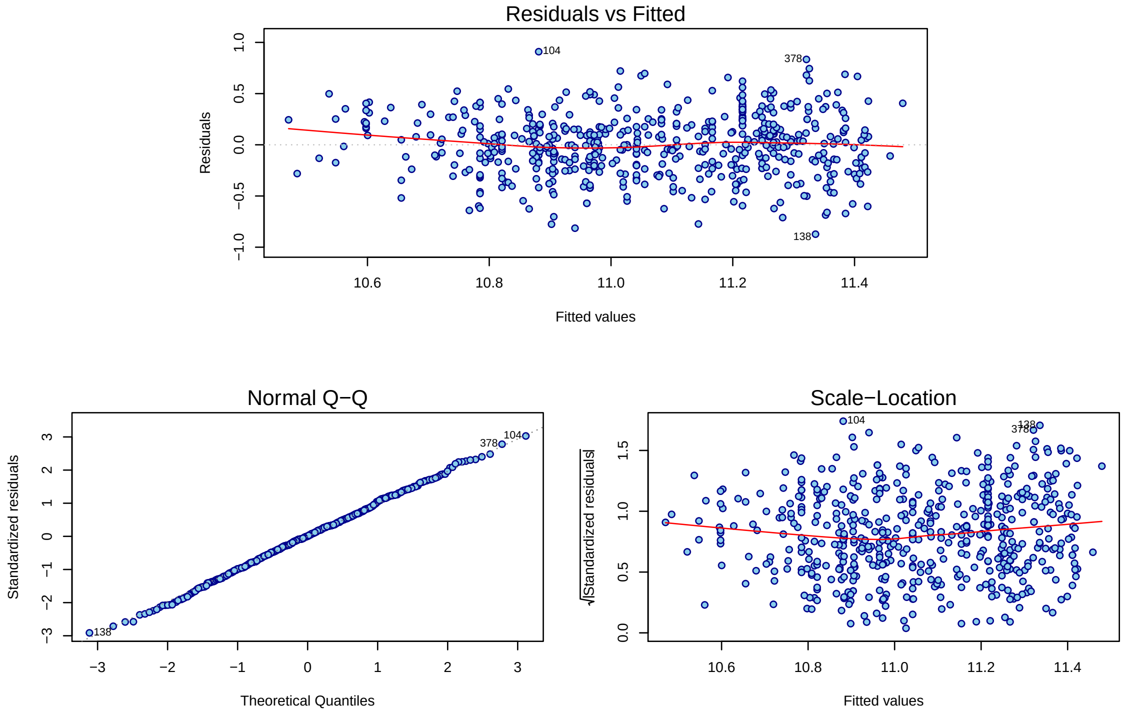
`log(price)` $\sim$ `rawpoly(ground, 3)`,

$\widehat{m}(z) = 9.97 + 3.78 \cdot 10^{-3}\,z - 3.31 \cdot 10^{-6}\,z^2 + 9.70 \cdot 10^{-10}\,z^3$



`Houses1987` ($n = 546$)

`log(price)` $\sim$ `rawpoly(ground, 3)`, **residual plots**



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 $\boldsymbol{\beta}^Z := \bigl(\beta_1, \ldots, \beta_{k-1}\bigr)^\top$. 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 will be introduced in Section 6.2.

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:

$$\begin{aligned}\mathbb{E}\bigl(Y \bigm| Z = z+1\bigr) \;-\; \mathbb{E}\bigl(Y \bigm| Z = z\bigr)& \\ = \beta_1 \;+\; \beta_2\,\bigl\{(z+1)^2 - z^2\bigr\} \;+\; \cdots \;+\; \beta_{k-1}\,\bigl\{(z+1)^{k-1} - z^{k-1}\bigr\},& \qquad z \in \mathcal{Z}.\end{aligned}$$

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 $\bigl(\mathbf{1}_n, \mathbb{S}\bigr)$ is of a full-rank $r = k$.

Illustrations

`Houses1987` ($n = 546$)

Effect of the covariate, interpretation of the regression coefficients

```
summary(lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.87279 -0.19903  0.00212  0.19780  0.90934

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.965e+00  1.371e-01  72.682  < 2e-16 ***
ground       3.784e-03  7.109e-04   5.323 1.49e-07 ***
I(ground^2) -3.306e-06  1.092e-06  -3.028  0.00258 **
I(ground^3)  9.700e-10  4.958e-10   1.957  0.05091 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,	Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

Degree of a polynomial

Test on a subset of regression coefficients (Section 6.2) or a submodel test (Section 8.2) can be used to infer on the degree of a polynomial in the regression function (4.4). The null hypothesis expressing, for $d < k$, belief that the regression function is a polynomial of degree $d-1$ corresponds to the null hypothesis

$$\mathrm{H}_0: \;\beta_d = 0 \;\; \& \;\; \ldots \;\; \& \;\; \beta_{k-1} = 0.$$

Illustrations

Houses1987 ($n = 546$)

Degree? Cubic versus quadratic, cubic versus linear polynomial

```
summary(lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987))
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.965e+00  1.371e-01  72.682  < 2e-16 ***
ground       3.784e-03  7.109e-04   5.323 1.49e-07 ***
I(ground^2) -3.306e-06  1.092e-06  -3.028  0.00258 **
I(ground^3)  9.700e-10  4.958e-10   1.957  0.05091 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

```
rp3 <- lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987)
rp1 <- lm(log(price) ~ ground, data = Houses1987)
anova(rp1, rp3)
```

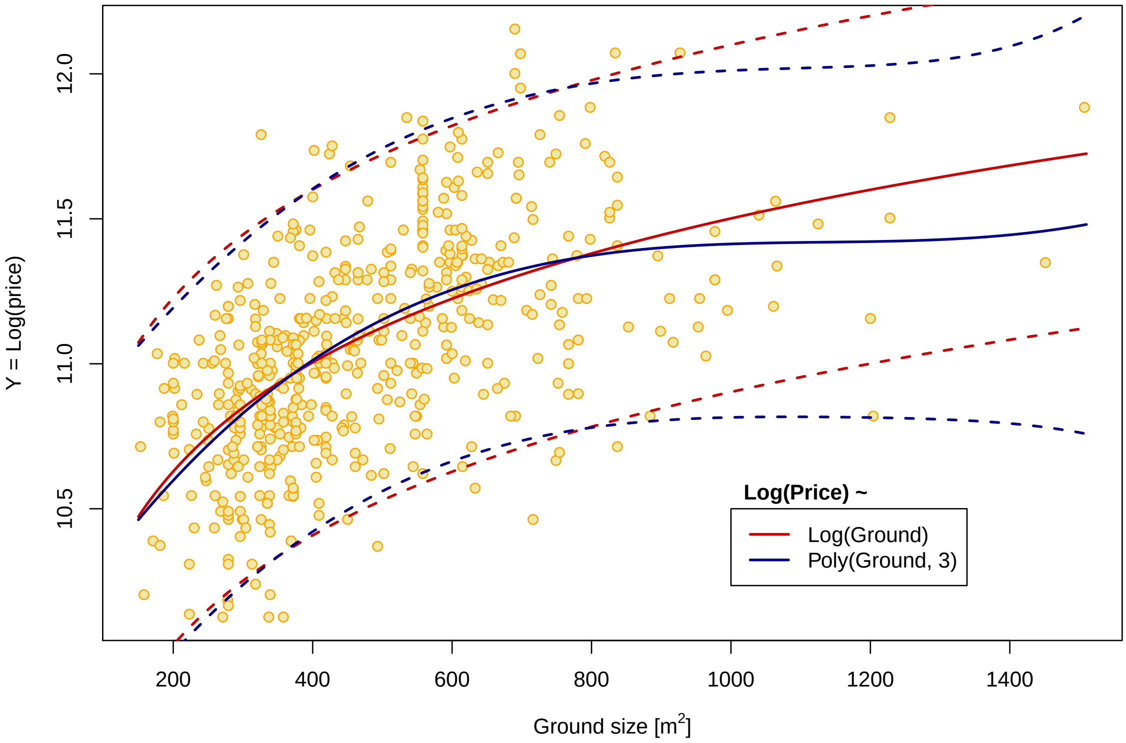
```
Analysis of Variance Table

Model 1: log(price) ~ ground
Model 2: log(price) ~ ground + I(ground^2) + I(ground^3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    544 53.186
2    542 48.968  2    4.2181 23.344 1.883e-10 ***
---
```

Illustrations

Houses1987 ($n = 546$)

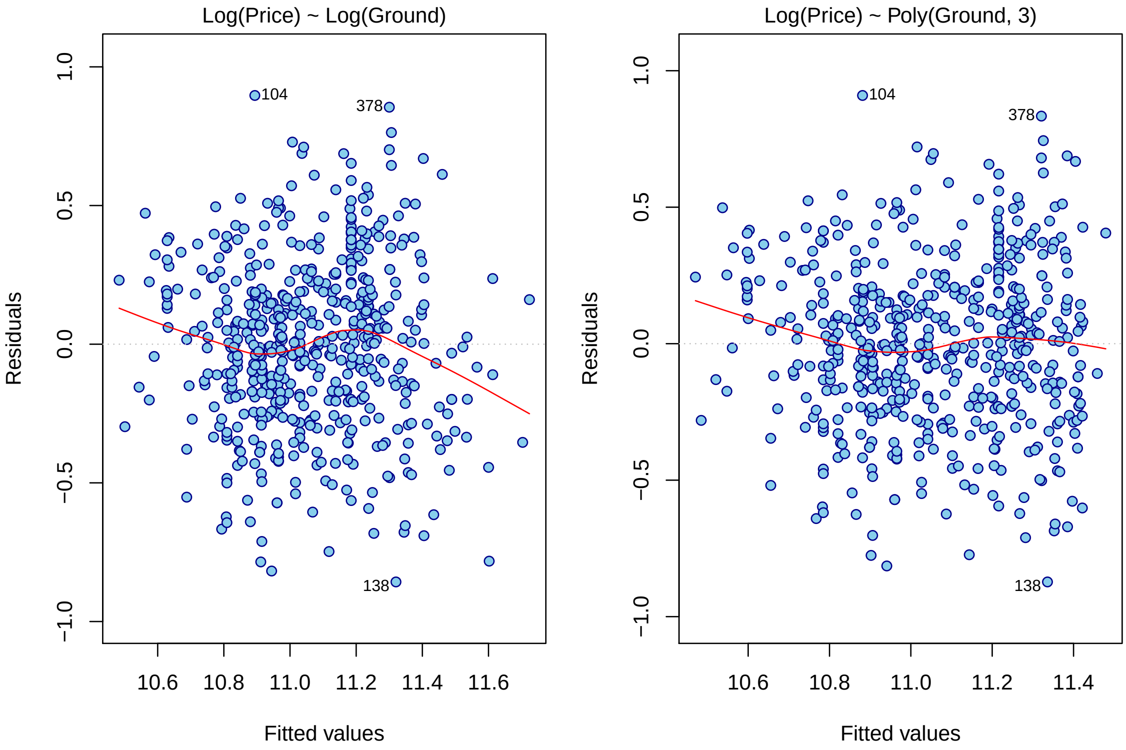
`log(price)` $\sim$ `log(ground)` **and** `log(price)` $\sim$ `rawpoly(ground, d)`**,**

$\widehat{m}$ **with 95% prediction band**



Houses1987 ($n = 546$)

`log(price)` $\sim$ `log(ground)` **and** `log(price)` $\sim$ `rawpoly(ground, d)`**, residuals vs. fitted plots**



Illustrations

Houses1987 ($n = 546$)

Practical importance of higher order polynomials?

```
summary(lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.87279 -0.19903  0.00212  0.19780  0.90934

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.965e+00  1.371e-01  72.682  < 2e-16 ***
ground       3.784e-03  7.109e-04   5.323 1.49e-07 ***
I(ground^2) -3.306e-06  1.092e-06  -3.028  0.00258 **
I(ground^3)  9.700e-10  4.958e-10   1.957  0.05091 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

4.3.3 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

$$m(z) = \beta_0 \,+\, \beta_1\, P^1(z) \,+\, \cdots \,+\, \beta_{k-1}\, P^{k-1}(z), \quad z \in \mathcal{Z}, \tag{4.5}$$

where P^j is an *orthonormal polynomial* of degree j, $j = 1, \dots, k-1$ built above a set of the covariate datapoints $Z_1, \dots, Z_n$. That is,

$$P^j(z) = a_{j,0} + a_{j,1}\, z + \cdots + a_{j,j}\, z^j, \qquad j = 1, \dots, k-1, \tag{4.6}$$

and the polynomial coefficients $a_{j,l}$, $j = 1, \dots, k-1$, $l = 0, \dots, j$ are such that vectors

$$\boldsymbol{P}^j = \begin{pmatrix} P^j(Z_1) \\ \vdots \\ P^j(Z_n) \end{pmatrix}, \quad j = 1, \dots, k-1,$$

are all *orthonormal* and also *orthogonal* to an intercept vector $\boldsymbol{P}^0 = \bigl(1, \dots, 1\bigr)^\top$. The corresponding reparameterizing matrix is

$$\mathbb{S} = \Bigl(\boldsymbol{P}^1, \quad \dots, \quad \boldsymbol{P}^{k-1}\Bigr) = \begin{pmatrix} P^1(Z_1) & \dots & P^{k-1}(Z_1) \\ \vdots & \vdots & \vdots \\ P^1(Z_n) & \dots & P^{k-1}(Z_n) \end{pmatrix}, \tag{4.7}$$

which leads to the model matrix $\mathbb{X} = \bigl(\mathbf{1}_n, \quad \mathbb{S}\bigr)$ 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 (4.6), see lectures on linear algebra. It can only be mentioned here that as soon as the data contain at least k different values among $Z_1, \dots, Z_n$, those polynomial coefficients exist and are unique.

Note. For given dataset and given polynomial degree $k-1$, the model matrix $\mathbb{X} = \bigl(\mathbf{1}_n, \ \mathbb{S}\bigr)$ 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.

Illustrations

`Houses1987` ($n = 546$)
`log(price)` $\sim$ `orthpoly(ground, 3)`

```
summary(lm(log(price) ~ poly(ground, degree = 3), data = Houses1987))
```

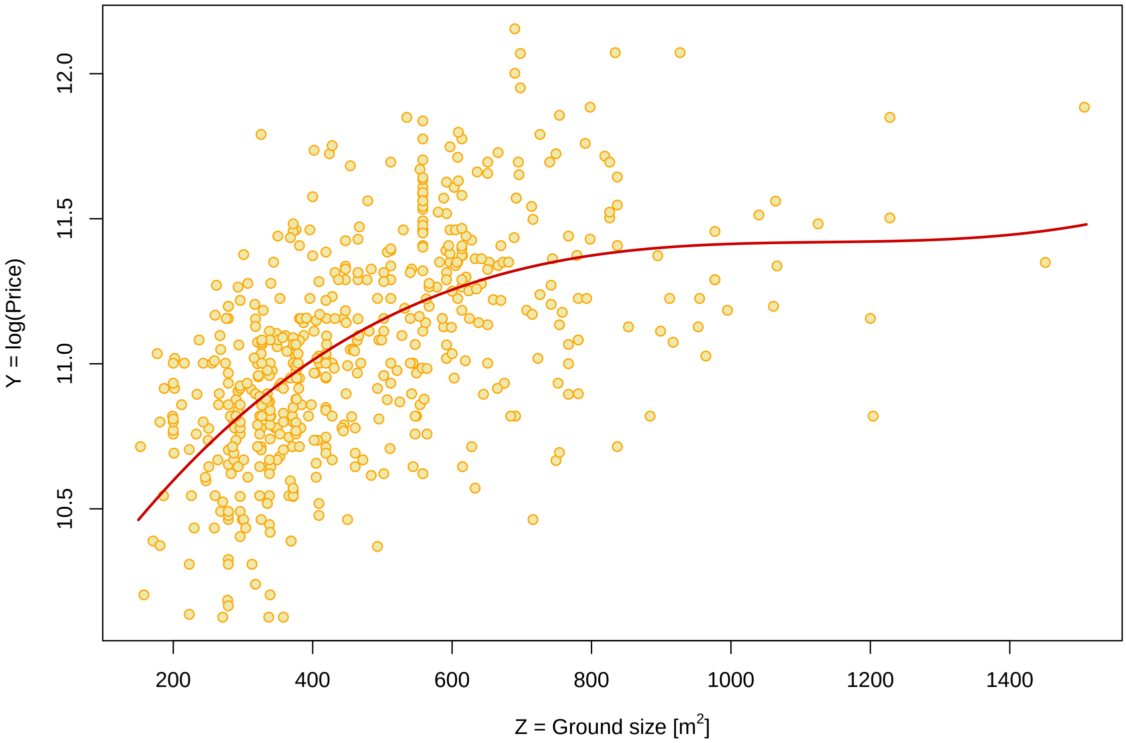
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.87279 -0.19903  0.00212  0.19780  0.90934 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               11.05896    0.01286 859.717  < 2e-16 ***
poly(ground, degree = 3)1  4.71459    0.30058  15.685  < 2e-16 ***
poly(ground, degree = 3)2 -1.96780    0.30058  -6.547 1.37e-10 ***
poly(ground, degree = 3)3  0.58811    0.30058   1.957   0.0509 .  
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,	Adjusted R-squared:  0.3471 
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

Illustrations

`Houses1987` ($n = 546$)

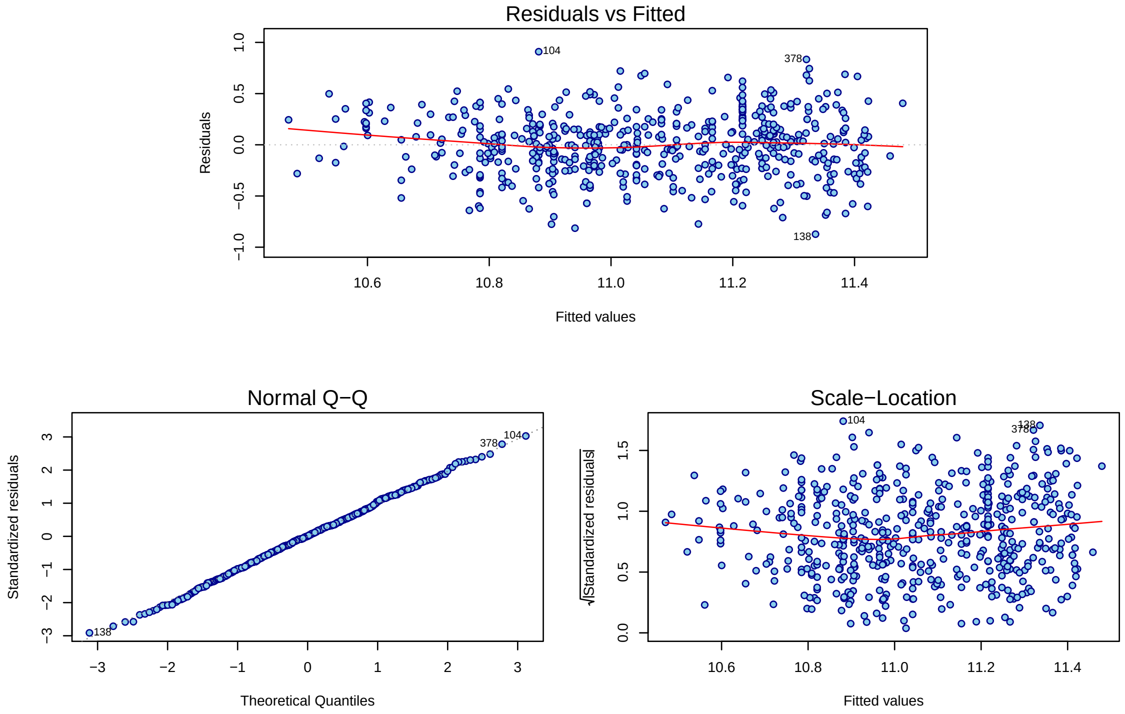
`log(price)` $\sim$ `orthpoly(ground, 3)`,

$\widehat{m}(z) = 11.06 + 4.71\,P^1(z) - 1.97\,P^2(z) + 0.59\,P^3(z)$



`Houses1987` ($n = 546$)

`log(price)` $\sim$ `orthpoly(ground, 3)`, **residual plots**



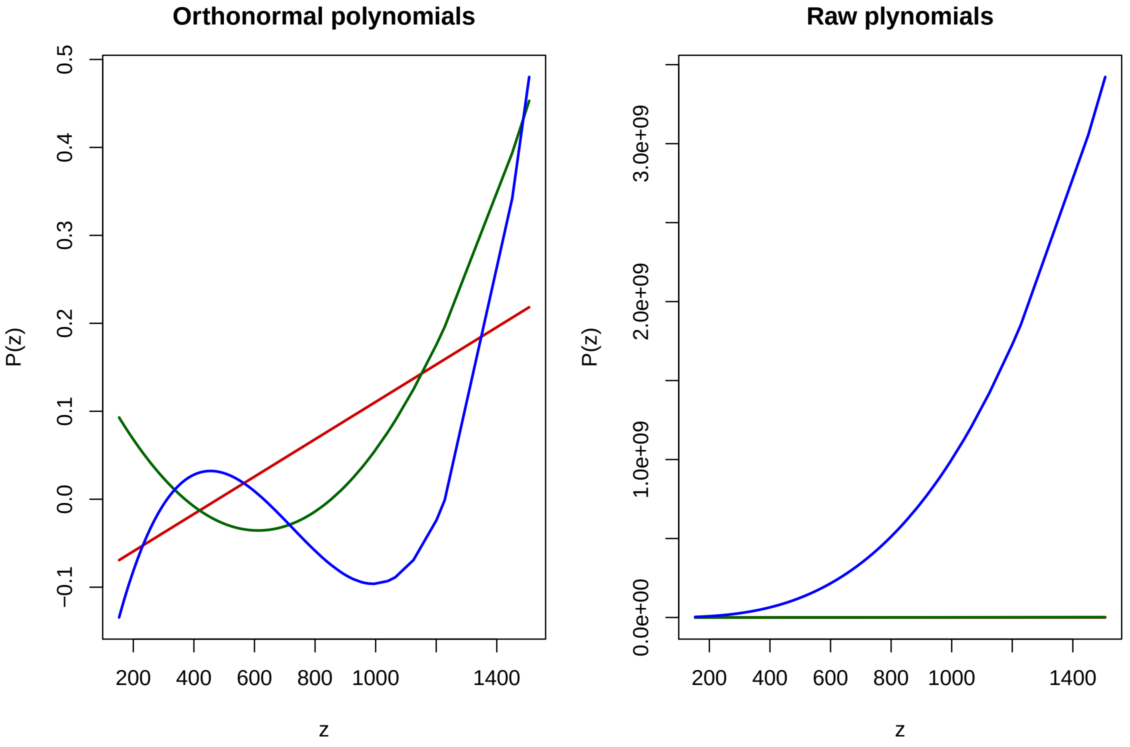
Advantages of orthonormal polynomials compared to raw polynomials

- 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.

Illustrations

Houses1987 ($n = 546$)

Basis orthonormal and raw polynomials



Illustrations

Houses1987 ($n = 546$)

Advantages of orthonormal polynomials compared to raw polynomials

```
summary(lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987))
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.965e+00  1.371e-01  72.682  < 2e-16 ***
ground       3.784e-03  7.109e-04   5.323 1.49e-07 ***
I(ground^2) -3.306e-06  1.092e-06  -3.028  0.00258 **
I(ground^3)  9.700e-10  4.958e-10   1.957  0.05091 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

```
summary(lm(log(price) ~ poly(ground, degree = 3), data = Houses1987))
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.05896    0.01286 859.717  < 2e-16 ***
poly(ground, degree = 3)1  4.71459    0.30058  15.685  < 2e-16 ***
poly(ground, degree = 3)2 -1.96780    0.30058  -6.547 1.37e-10 ***
poly(ground, degree = 3)3  0.58811    0.30058   1.957   0.0509 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

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 $\boldsymbol{\beta}^Z := \bigl(\beta_1, \dots, \beta_{k-1}\bigr)^\top$. 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 6.2 for a possible test procedure.

Illustrations

Houses1987 ($n = 546$)

Effect of the covariate (cubic versus constant regression function)

```
summary(lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987))
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.965e+00  1.371e-01  72.682  < 2e-16 ***
ground       3.784e-03  7.109e-04   5.323 1.49e-07 ***
I(ground^2) -3.306e-06  1.092e-06  -3.028  0.00258 **
I(ground^3)  9.700e-10  4.958e-10   1.957  0.05091 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

```
summary(lm(log(price) ~ poly(ground, degree = 3), data = Houses1987))
```

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.05896    0.01286 859.717  < 2e-16 ***
poly(ground, degree = 3)1  4.71459    0.30058  15.685  < 2e-16 ***
poly(ground, degree = 3)2 -1.96780    0.30058  -6.547 1.37e-10 ***
poly(ground, degree = 3)3  0.58811    0.30058   1.957   0.0509 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

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 (will be introduced in Sections 6.2 and 8.2) can again be used to infer on the degree of a polynomial in the regression function (4.5) 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

$$\mathrm{H}_0: \ \beta_d = 0 \ \ \& \ \ \ldots \ \ \& \ \ \beta_{k-1} = 0.$$

Illustrations

`Houses1987` ($n = 546$)

Degree? Cubic versus quadratic regression function

```
summary(lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987))
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.965e+00  1.371e-01  72.682  < 2e-16 ***
ground        3.784e-03  7.109e-04   5.323 1.49e-07 ***
I(ground^2)  -3.306e-06  1.092e-06  -3.028  0.00258 **
I(ground^3)   9.700e-10  4.958e-10   1.957  0.05091 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

```
summary(lm(log(price) ~ poly(ground, degree = 3), data = Houses1987))
```

```
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                11.05896    0.01286 859.717  < 2e-16 ***
poly(ground, degree = 3)1   4.71459    0.30058  15.685  < 2e-16 ***
poly(ground, degree = 3)2  -1.96780    0.30058  -6.547 1.37e-10 ***
poly(ground, degree = 3)3   0.58811    0.30058   1.957   0.0509 .
---

Residual standard error: 0.3006 on 542 degrees of freedom
Multiple R-squared:  0.3507,        Adjusted R-squared:  0.3471
F-statistic: 97.57 on 3 and 542 DF,  p-value: < 2.2e-16
```

Illustrations

Houses1987 ($n = 546$)

Degree? Cubic versus linear regression function

```
rp3 <- lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987)
rp1 <- lm(log(price) ~ ground, data = Houses1987)
anova(rp1, rp3)
```

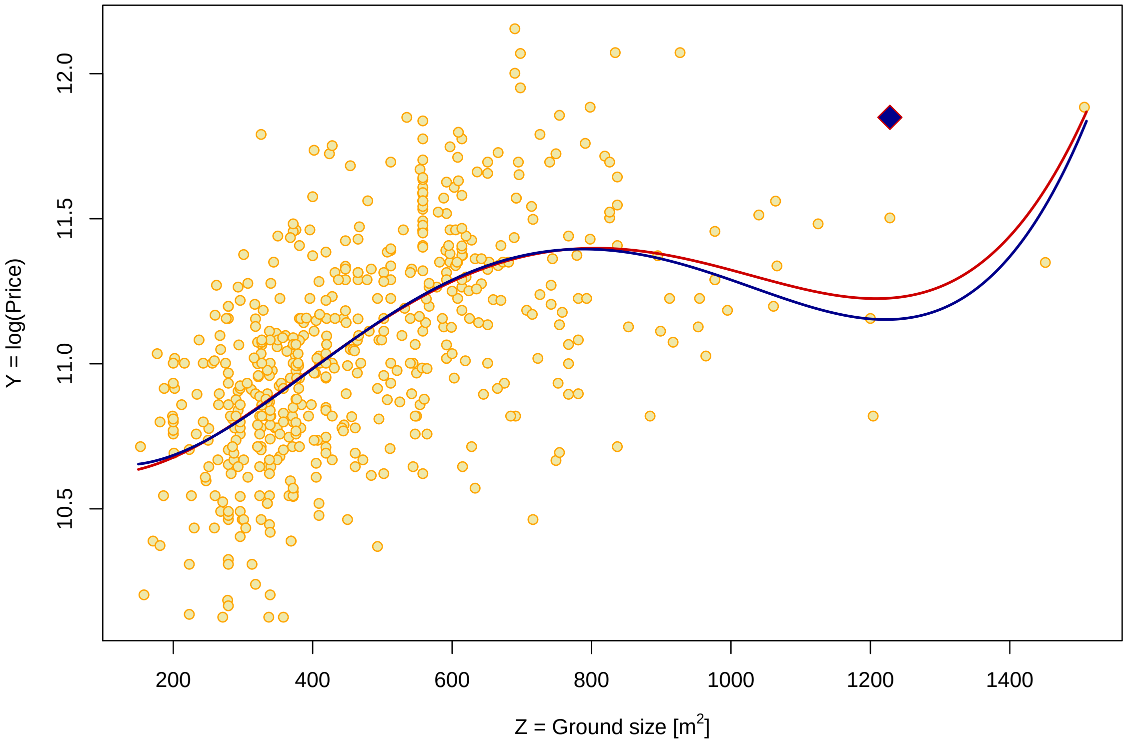
```
Analysis of Variance Table

Model 1: log(price) ~ ground
Model 2: log(price) ~ ground + I(ground^2) + I(ground^3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    544 53.186
2    542 48.968  2    4.2181 23.344 1.883e-10 ***
---
```

```
op3 <- lm(log(price) ~ poly(ground, degree = 3), data = Houses1987)
op1 <- lm(log(price) ~ poly(ground, degree = 1), data = Houses1987)
anova(op1, op3)
```

```
Analysis of Variance Table

Model 1: log(price) ~ poly(ground, degree = 1)
Model 2: log(price) ~ poly(ground, degree = 3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    544 53.186
2    542 48.968  2    4.2181 23.344 1.883e-10 ***
---
```

Houses1987 ($n = 546$)

`log(price)` $\sim$ `poly(ground, 4)`, **global effect**



4.3.4 Regression splines

Basis splines

The advantage of a polynomial regression function introduced in Sections 4.3.2 and 4.3.3 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 $\mathbb{E}(Y \,|\, Z = z)$ 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 4.2 Basis spline with distinct knots.

Let $d \in \mathbb{N}_0$ and $\boldsymbol{\lambda} = \bigl(\lambda_1, \dots, \lambda_{d+2}\bigr)^\top \in \mathbb{R}^{d+2}$, 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 $(\lambda_j,\, \lambda_{j+1})$, $j = 1, \dots, d+1$, $B^d(\cdot;\, \boldsymbol{\lambda})$ *is a polynomial of degree* d;

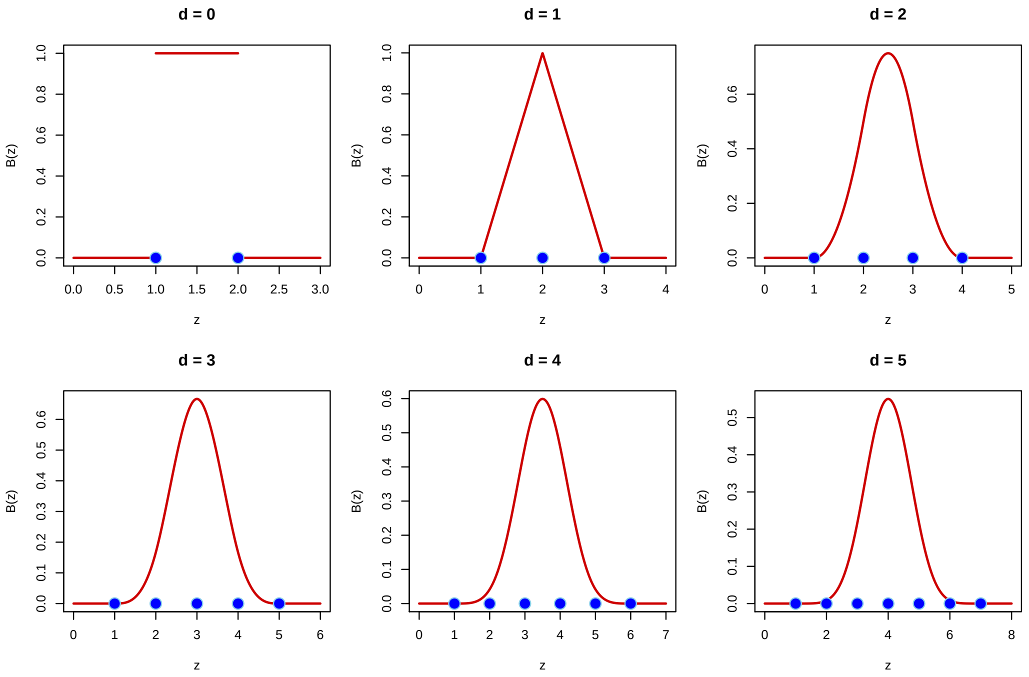
(iii) $B^d(\cdot;\, \boldsymbol{\lambda})$ *has continuous derivatives up to an order* $d-1$ *on* $\mathbb{R}$.

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 $\lambda_2, \dots, \lambda_{d+1}$.
- On the boundary (λ_1 and λ_{d+2}), the polynomial pieces are connected smoothly (of order $d-1$) with a constant zero.

Illustrations

Some basis splines of degree $d = 0, \dots, 5$



[7] *bazický spline [čti splajn] stupně* d *se vzájemně různými uzly*

[8] *po částech*

Definition 4.3 Basis spline with coincident left boundary knots.

Let $d \in \mathbb{N}_0$, $1 < r < d+2$ *and* $\boldsymbol{\lambda} = \bigl(\lambda_1, \ldots, \lambda_{d+2}\bigr)^\top \in \mathbb{R}^{d+2}$, *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*

(i) $B^d(z;\,\boldsymbol{\lambda}) = 0$, *for* $z \le \lambda_r$ *and* $z \ge \lambda_{d+2}$*;*

(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* $(\lambda_r,\,\infty)$*;*

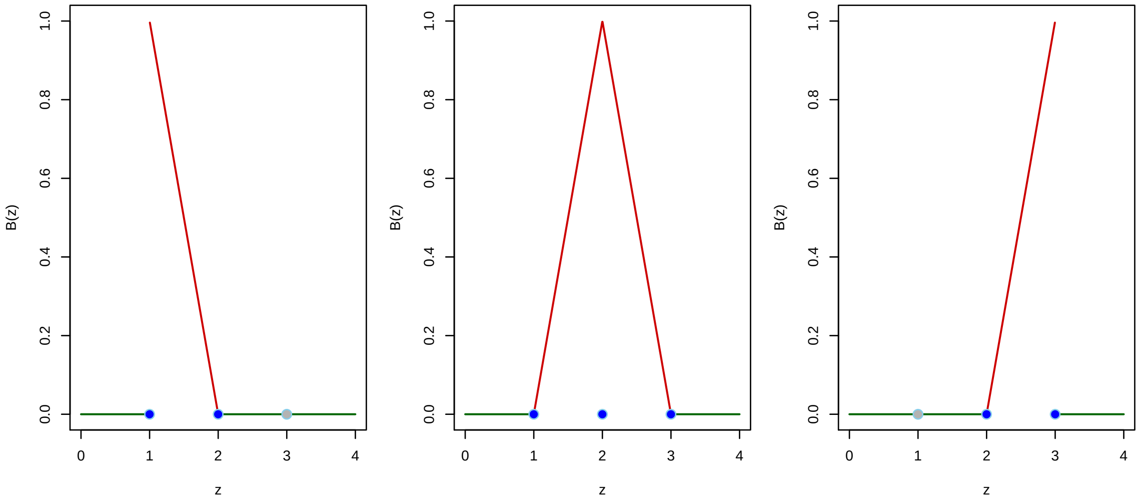
(iv) $B^d(\cdot;\,\boldsymbol{\lambda})$ *has continuous derivatives up to an order* $d-r$ *in* λ_r.

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 4.3 to the right boundary, basis spline with coincident right boundary knots is defined.

Illustrations

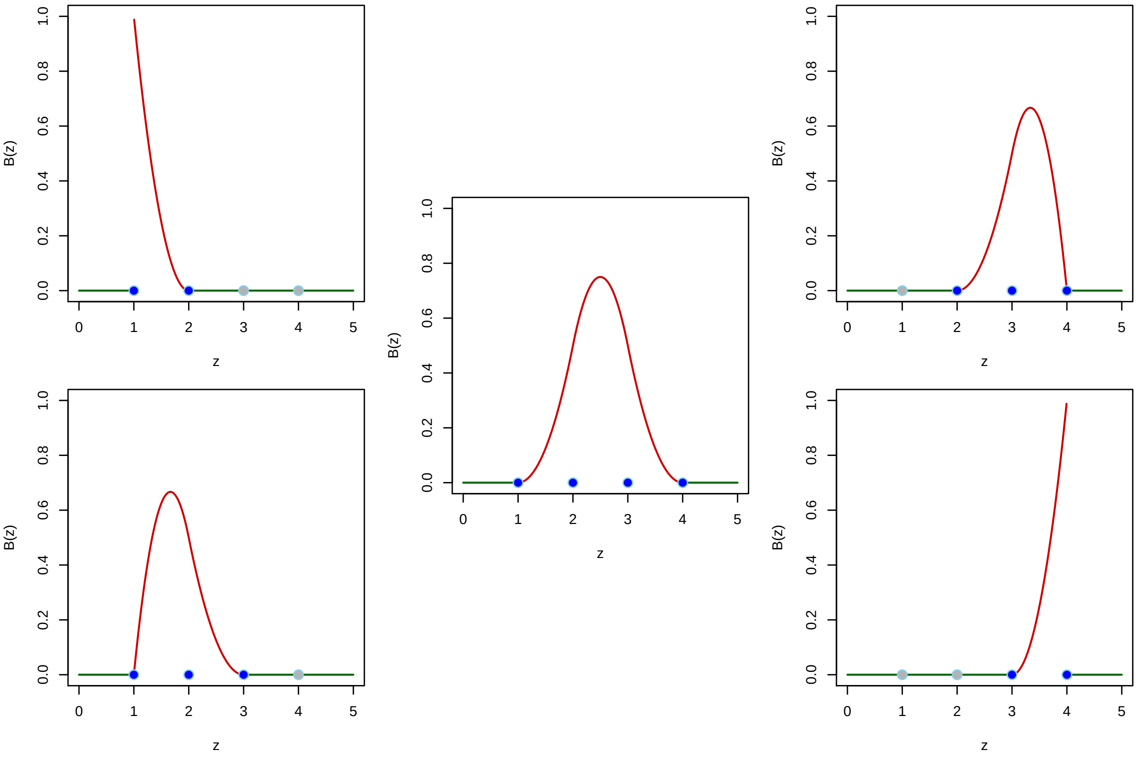
Some basis splines of degree $d = 1$ with possibly coincident boundary knots



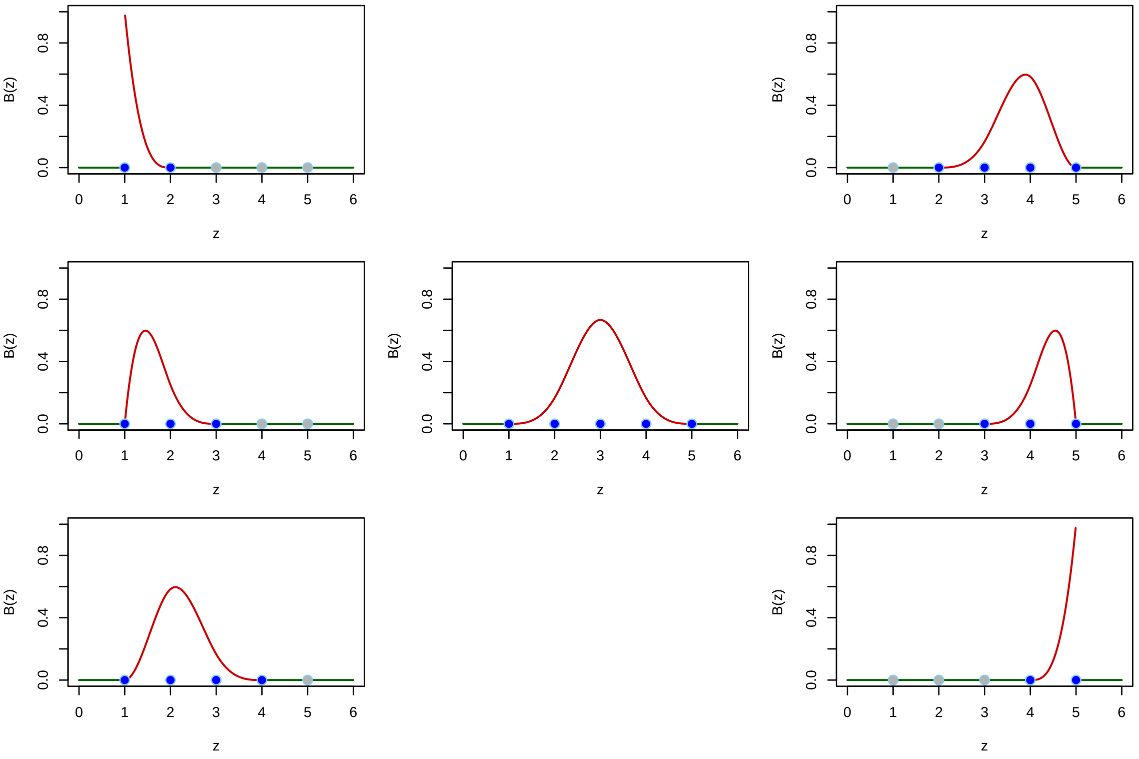
[9] *bazický spline stupně d s r překrývajícími se levými uzly*

Illustrations

Some basis splines of degree $d = 2$ with possibly coincident boundary knots



Some basis splines of degree $d = 3$ with possibly coincident boundary knots



Basis B-splines

There are many ways on how to construct the basis splines that satisfy conditions of Definitions 4.2 and 4.3, 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 $\boldsymbol{\lambda} = \bigl(\lambda_1, \ldots, \lambda_{d+2}\bigr)^\top$ 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

$$\begin{aligned} B^d(z, \boldsymbol{\lambda}) &> 0, \qquad \lambda_1 < z < \lambda_{d+2}, \\ B^d(z, \boldsymbol{\lambda}) &= 0, \qquad z \le \lambda_1,\; z \ge \lambda_{d+2}. \end{aligned}$$

Spline basis

Definition 4.4 Spline basis.

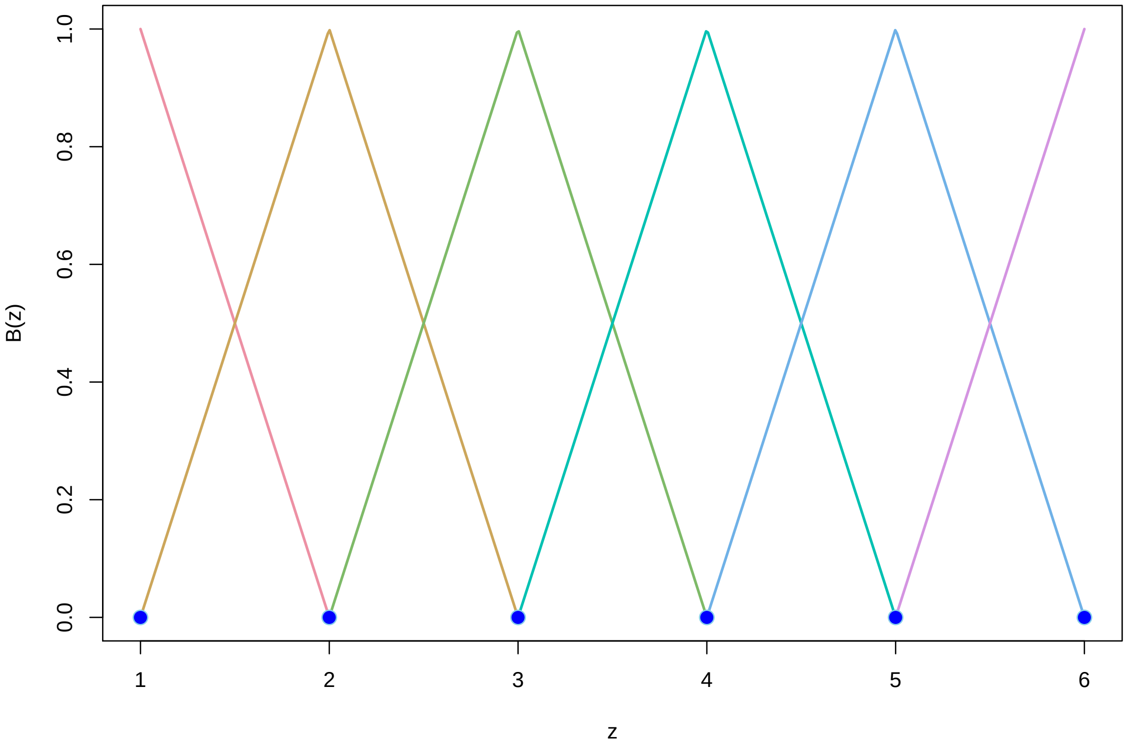
Let $d \in \mathbb{N}_0$, $k \ge d+1$ and $\boldsymbol{\lambda} = \bigl(\lambda_1, \ldots, \lambda_{k-d+1}\bigr)^\top \in \mathbb{R}^{k-d+1}$, where $-\infty < \lambda_1 < \ldots < \lambda_{k-d+1} < \infty$. The spline basis[10] *of degree d with knots $\boldsymbol{\lambda}$ is a set of basis splines $B_1, \ldots, B_k$, where for $z \in \mathbb{R}$,*

$$\begin{aligned}
B_1(z) &= B^d(z;\, \underbrace{\lambda_1, \ldots, \lambda_1}_{(d+1)\times},\, \lambda_2), \\
B_2(z) &= B^d(z;\, \underbrace{\lambda_1, \ldots, \lambda_1}_{d\times},\, \lambda_2,\, \lambda_3), \\
&\;\vdots \\
B_d(z) &= B^d(z;\, \underbrace{\lambda_1, \lambda_1}_{2\times},\, \lambda_2, \ldots, \lambda_{d+1}), \\
B_{d+1}(z) &= B^d(z;\, \lambda_1, \lambda_2, \ldots, \lambda_{d+2}), \\
B_{d+2}(z) &= B^d(z;\, \lambda_2, \ldots, \lambda_{d+3}), \\
&\;\vdots \\
B_{k-d}(z) &= B^d(z;\, \lambda_{k-2d}, \ldots, \lambda_{k-d+1}), \\
B_{k-d+1}(z) &= B^d(z;\, \lambda_{k-2d+1}, \ldots, \underbrace{\lambda_{k-d+1}, \lambda_{k-d+1}}_{2\times}), \\
&\;\vdots \\
B_{k-1}(z) &= B^d(z;\, \lambda_{k-d-1}, \lambda_{k-d} \ldots, \underbrace{\lambda_{k-d+1}, \ldots, \lambda_{k-d+1}}_{d\times}), \\
B_k(z) &= B^d(z;\, \lambda_{k-d} \ldots, \underbrace{\lambda_{k-d+1}, \ldots, \lambda_{k-d+1}}_{(d+1)\times}).
\end{aligned}$$

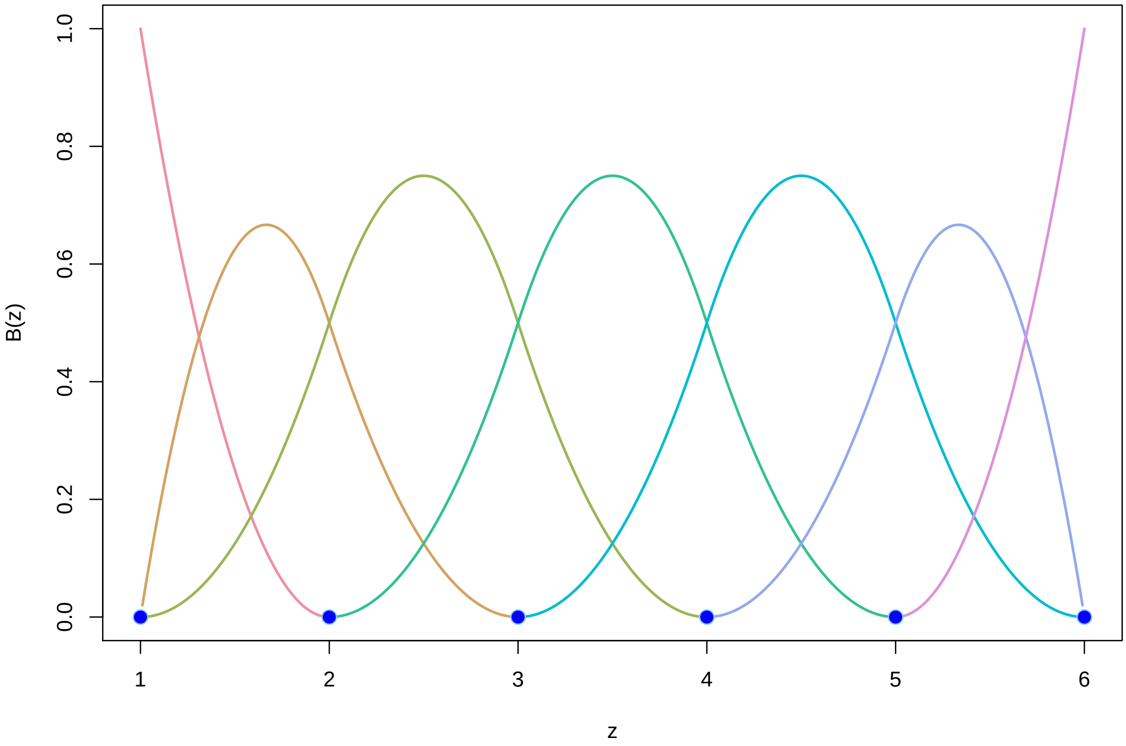
[10] *splinová báze*

Illustrations

Linear B-spline basis (of degree $d = 1$)

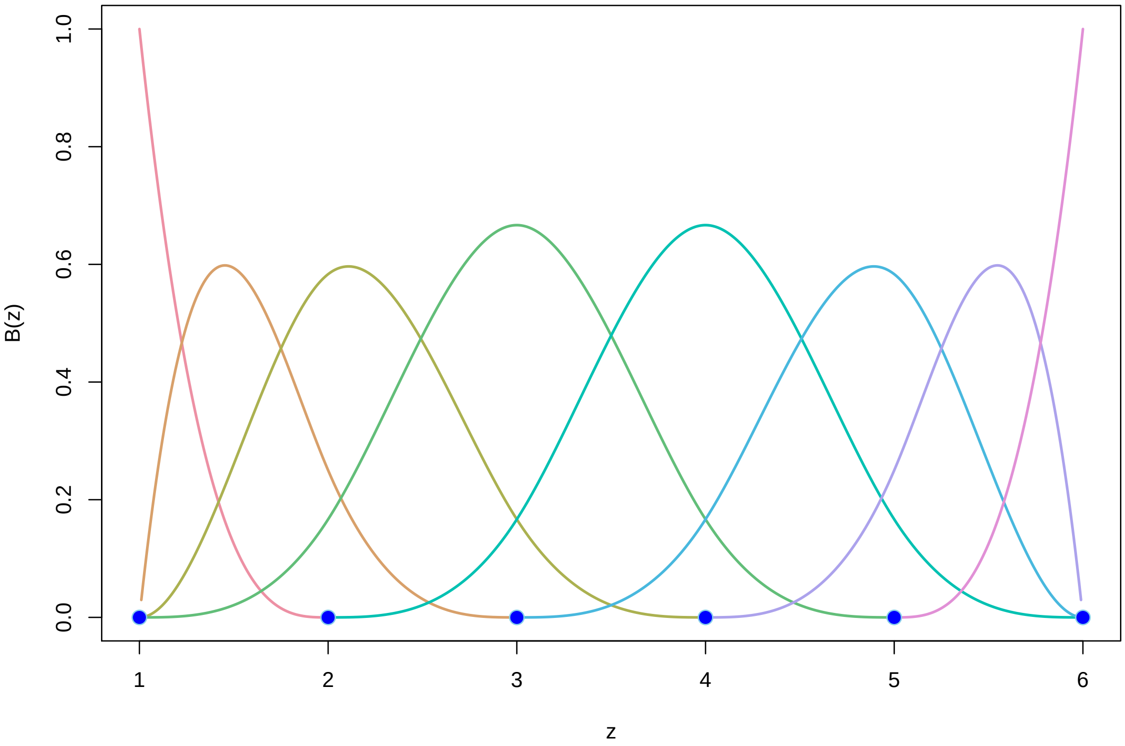


Quadratic B-spline basis (of degree $d = 2$)



Illustrations

Cubic B-spline basis (of degree $d = 3$)



Properties of the B-spline basis

If $k \geq d+1$, a set of knots $\boldsymbol{\lambda} = \bigl(\lambda_1, \dots, \lambda_{k-d+1}\bigr)^\top$, $-\infty < \lambda_1 < \ldots < \lambda_{k-d+1} < \infty$ is given and $B_1, \dots, B_k$ is the spline basis of degree d with knots $\boldsymbol{\lambda}$ composed of basis B-splines $k \geq d+1$, a set of knots $\boldsymbol{\lambda} = \bigl(\lambda_1, \dots, \lambda_{k-d+1}\bigr)^\top$, $-\infty < \lambda_1 < \ldots < \lambda_{k-d+1} < \infty$ is given and $B_1, \dots, B_k$ is the spline basis of degree d with knots $\boldsymbol{\lambda}$ composed of basis B-splines then

(a)
$$\sum_{j=1}^{k} B_j(z) = 1 \qquad \text{for all } z \in \bigl(\lambda_1, \lambda_{k-d+1}\bigr); \tag{4.8}$$

(b) for each $m \leq d$ there exist a set of coefficients $\gamma_1^m, \dots, \gamma_k^m$ such that

$$\sum_{j=1}^{k} \gamma_j^m B_j(z) \text{ is on } (\lambda_1, \lambda_{k-d+1}) \text{ a polynomial in } z \text{ of degree } m. \tag{4.9}$$

Regression spline

It will now be assumed that the covariate space is a bounded interval, i.e., $\mathcal{Z} = \bigl(z_{min}, z_{max}\bigr)$, $-\infty < z_{min} < z_{max} < \infty$. The regression function that exploits the regression splines is

$$m(z) = \beta_1 B_1(z) + \cdots + \beta_k B_k(z), \quad z \in \mathcal{Z}, \tag{4.10}$$

where $B_1, \dots, B_k$ is the spline basis of chosen degree $d \in \mathbb{N}_0$ composed of basis B-splines built above a set of chosen knots $\boldsymbol{\lambda} = \bigl(\lambda_1, \dots, \lambda_{k-d+1}\bigr)^\top$, $z_{min} = \lambda_1 < \ldots < \lambda_{k-d+1} = z_{max}$. The corresponding reparameterizing matrix coincided with the model matrix and is

$$\mathbb{X} = \mathbb{S} = \begin{pmatrix} B_1(Z_1) & \dots & B_k(Z_1) \\ \vdots & \vdots & \vdots \\ B_1(Z_n) & \dots & B_k(Z_n) \end{pmatrix} =: \mathbb{B}. \tag{4.11}$$

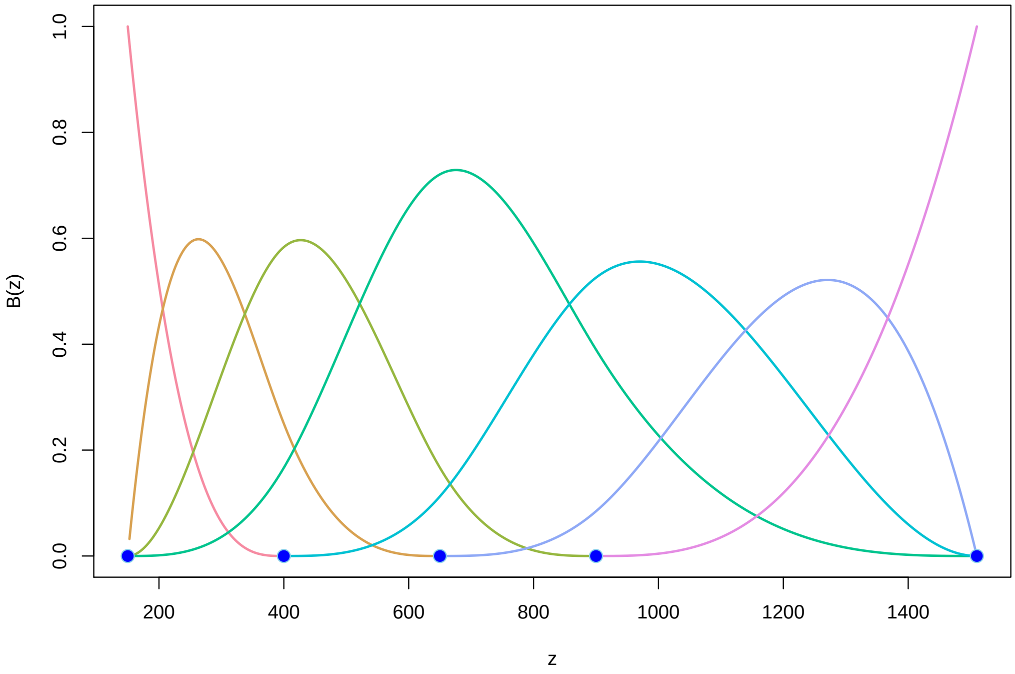
Notes.

- It follows from (4.8) that
$$\mathbf{1}_n \in \mathcal{M}\bigl(\mathbb{B}\bigr).$$
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 $\boldsymbol{\beta} = \bigl(\beta_1, \dots, \beta_k\bigr)^\top$.
- It also follows from (4.9) that for any $m \le d$, 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 (4.10) 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 $\lambda_1, \dots, \lambda_{k-d+1}$) is quite important to obtain the regression function that sufficiently well approximates the response expectations $\mathbb{E}\bigl(Y \bigm| Z = z\bigr)$, $z \in \mathcal{Z}$. 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.

Illustrations

`Houses1987` ($n = 546$)

B-spline basis (cubic, $d = 3$, $\boldsymbol{\lambda} = \bigl(150,\ 400,\ 650,\ 900,\ 1510\bigr)^\top$)



Illustrations

Houses1987 ($n = 546$)

log(price) $\sim$ spline(ground, degree = 3), **model matrix** $\mathbb{X} = \mathbb{B}$

```
lambda.inner <- c(400, 650, 900)
lambda.bound <- c(150, 1510)
Bx <- bs(Houses1987[, "ground"],
         knots = lambda.inner, Boundary.knots = lambda.bound,
         degree = 3, intercept = TRUE)
showBx <- data.frame(ground = Houses1987[, "ground"],
              B1 = Bx[,1], B2 = Bx[,2], B3 = Bx[,3],
              B4 = Bx[,4], B5 = Bx[,5], B6 = Bx[,6], B7 = Bx[,7])
print(showBx)
```

```
   ground    B1    B2    B3    B4    B5 B6 B7
1     544 0.000 0.019 0.424 0.535 0.022  0  0
2     372 0.001 0.341 0.541 0.117 0.000  0  0
3     285 0.097 0.583 0.293 0.026 0.000  0  0
4     619 0.000 0.000 0.235 0.689 0.076  0  0
5     592 0.000 0.003 0.302 0.644 0.051  0  0
6     387 0.000 0.291 0.567 0.142 0.000  0  0
7     361 0.004 0.379 0.517 0.100 0.000  0  0
8     387 0.000 0.291 0.567 0.142 0.000  0  0
9     447 0.000 0.134 0.590 0.275 0.001  0  0
10    512 0.000 0.042 0.497 0.451 0.010  0  0
11    670 0.000 0.000 0.130 0.729 0.142  0  0
12    279 0.113 0.590 0.273 0.023 0.000  0  0
13    158 0.907 0.091 0.002 0.000 0.000  0  0
14    268 0.147 0.597 0.238 0.018 0.000  0  0
15    335 0.018 0.465 0.450 0.068 0.000  0  0
...
```

Houses1987 ($n = 546$)

log(price) $\sim$ spline(ground, degree = 3)

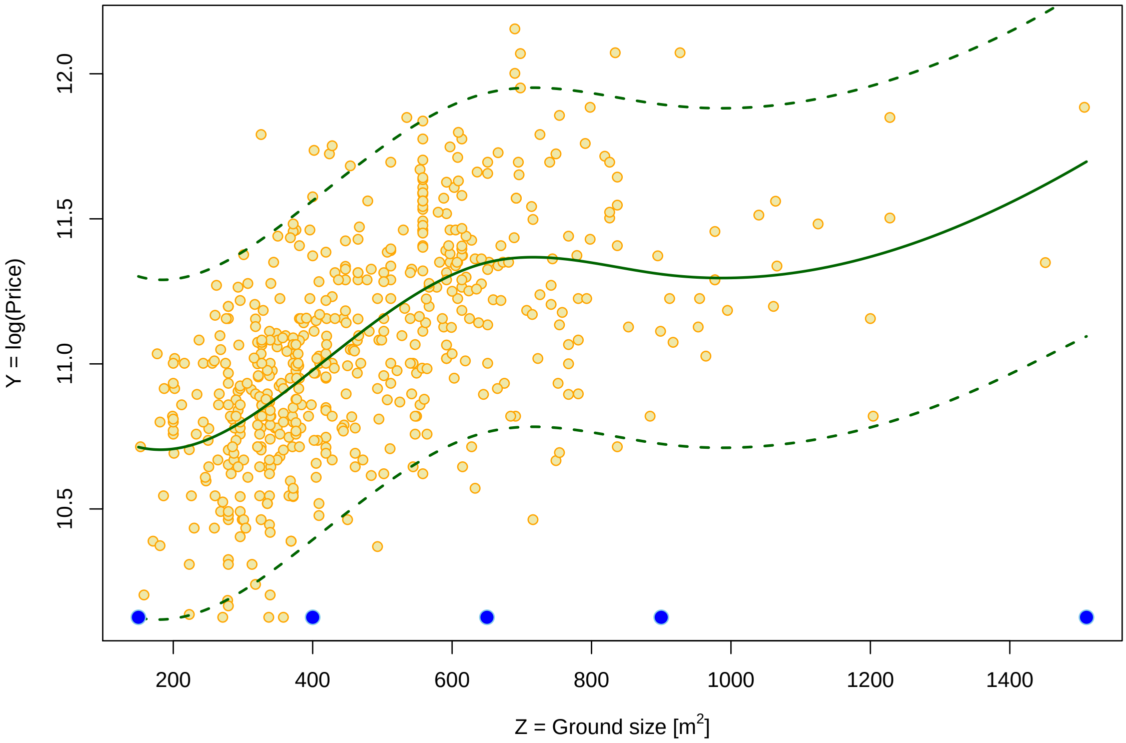
```
summary(lm(log(price) ~ Bx - 1, data = Houses1987))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.90457 -0.19497  0.00698  0.19693  0.94698

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
Bx1 10.71312    0.12078   88.70   <2e-16 ***
Bx2 10.66519    0.07956  134.06   <2e-16 ***
Bx3 10.97388    0.07464  147.03   <2e-16 ***
Bx4 11.46283    0.06699  171.11   <2e-16 ***
Bx5 11.17900    0.16773   66.65   <2e-16 ***
Bx6 11.41145    0.31448   36.29   <2e-16 ***
Bx7 11.69708    0.25076   46.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2974 on 539 degrees of freedom
Multiple R-squared:  0.9993,       Adjusted R-squared:  0.9993
F-statistic: 1.079e+05 on 7 and 539 DF,  p-value: < 2.2e-16
```

Illustrations

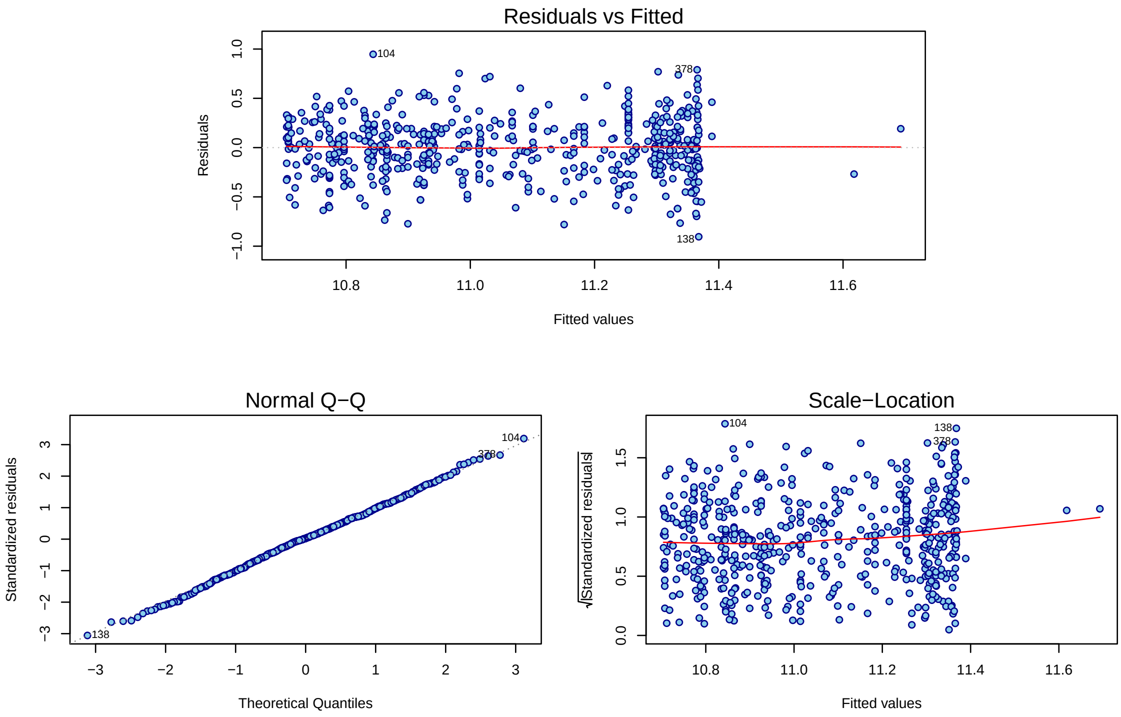
`Houses1987` ($n = 546$)

`log(price)` $\sim$ `spline(ground)`, $\widehat{m}(z) = 10.71\,B_1(z) + 10.67\,B_2(z) + 10.97\,B_3(z) +$ $11.46\,B_4(z) + 11.18\,B_5(z) + 11\;41\,B_6(z) + 11.70\,B_7(z)$ **and the 95% prediction band**



`Houses1987` ($n = 546$)

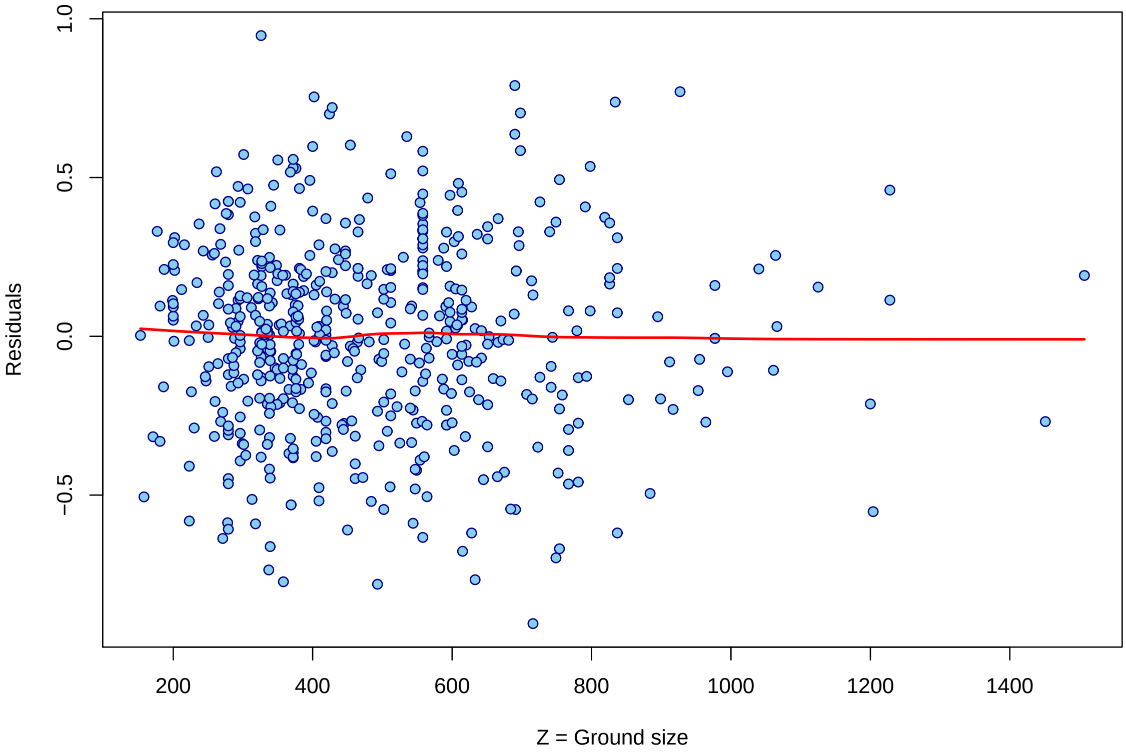
`log(price)` $\sim$ `spline(ground)`, **residual plots**



Illustrations

`Houses1987` ($n = 546$)

`log(price)` $\sim$ `spline(ground)`, **residuals versus covariate plot**



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 (4.8), this null hypothesis corresponds to assuming that $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbf{1}_n\big) \subset \mathcal{M}\big(\mathbb{B}\big)$. Consequently, it is possible to use a test on submodel that will be introduced in Section 8.1 to test the above null hypothesis.

Illustrations

Houses1987 ($n = 546$)

Effect of the covariate

```
mB <- lm(log(price) ~ Bx - 1, data = Houses1987)
m0 <- lm(log(price) ~ 1, data = Houses1987)
anova(m0, mB)
```

```
Analysis of Variance Table

Model 1: log(price) ~ 1
Model 2: log(price) ~ Bx - 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    545 75.413
2    539 47.663  6     27.75 52.302 < 2.2e-16 ***
---
```

Houses1987 ($n = 546$)

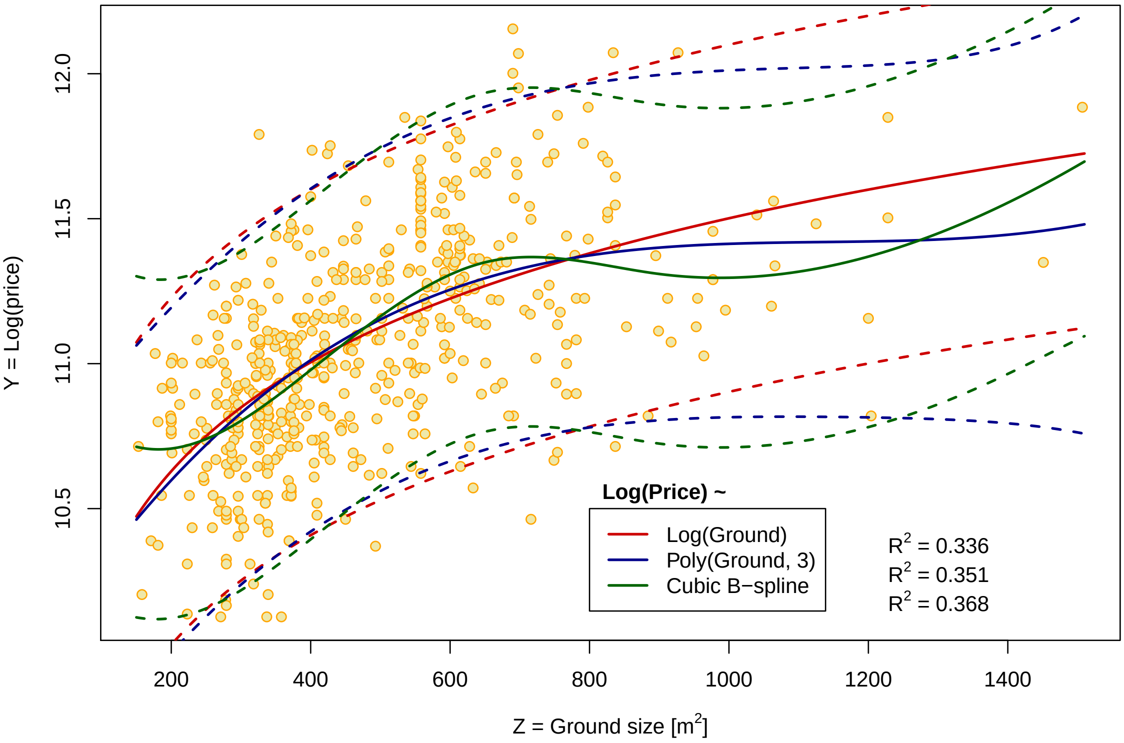
Spline better than a (global) cubic polynomial?

```
mB <- lm(log(price) ~ Bx - 1, data = Houses1987)
mpoly3 <- lm(log(price) ~ ground + I(ground^2) + I(ground^3), data = Houses1987)
anova(mpoly3, mB)
```

```
Analysis of Variance Table

Model 1: log(price) ~ ground + I(ground^2) + I(ground^3)
Model 2: log(price) ~ Bx - 1
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    542 48.968
2    539 47.663  3    1.3045 4.9174 0.002226 **
---
```

Illustrations

`Houses1987` ($n = 546$)
`log(price)` $\sim$ `log(ground)`, `log(price)` $\sim$ `poly(ground, 3)`,
`log(price)` $\sim$ `spline(ground, degree = 3)`, $\widehat{m}$ **with the 95% prediction band**



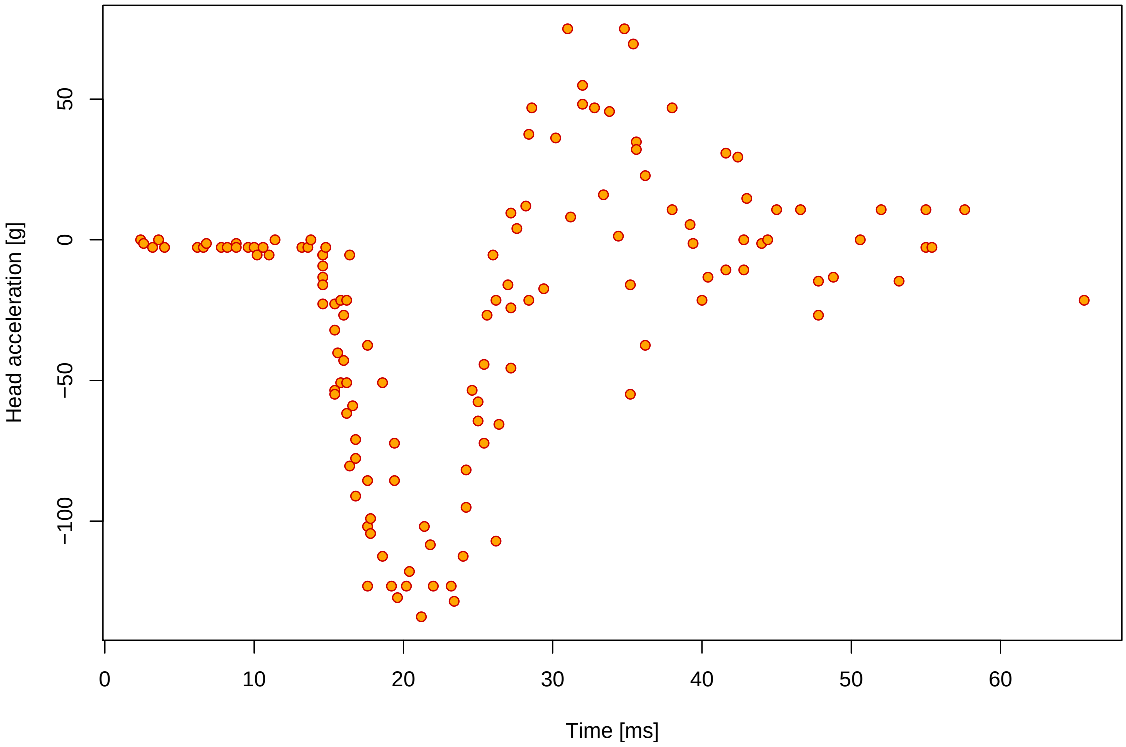
Interpretation of the regression coefficients

The single regression coefficients $\beta_1, \ldots, \beta_k$ do not usually have a direct reasonable interpretation.

Illustrations

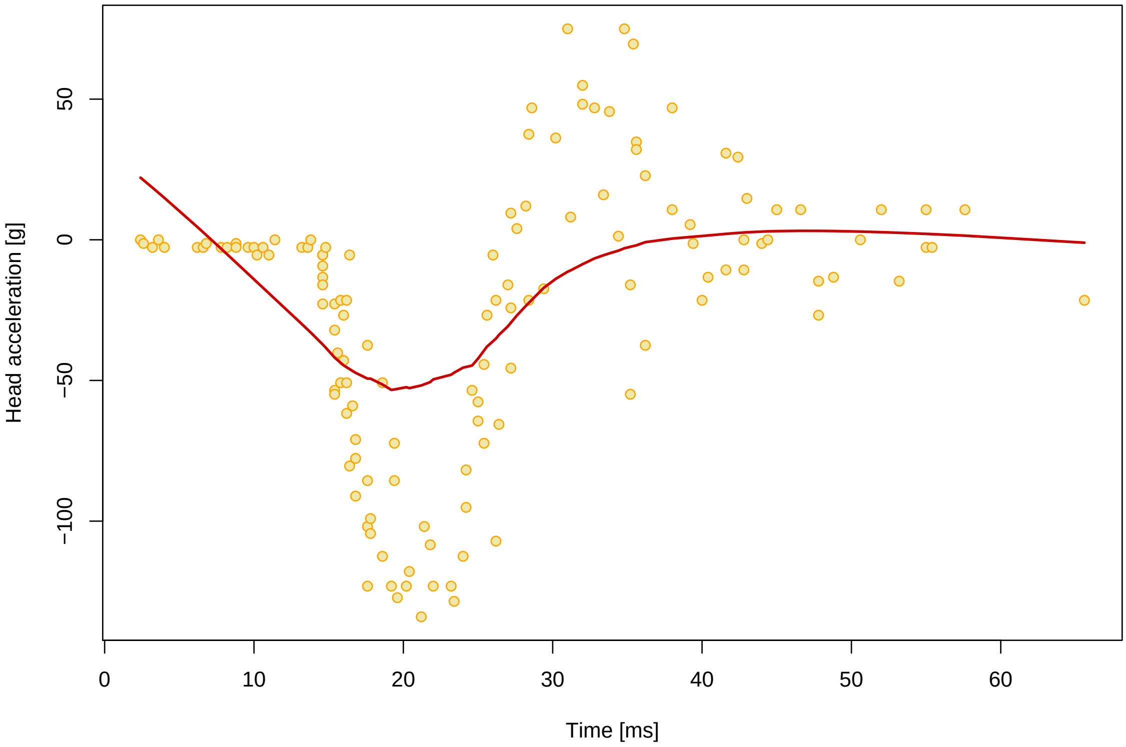
`Motorcycle` ($n = 133$)

`haccel` $\sim$ `time`



`Motorcycle` ($n = 133$)

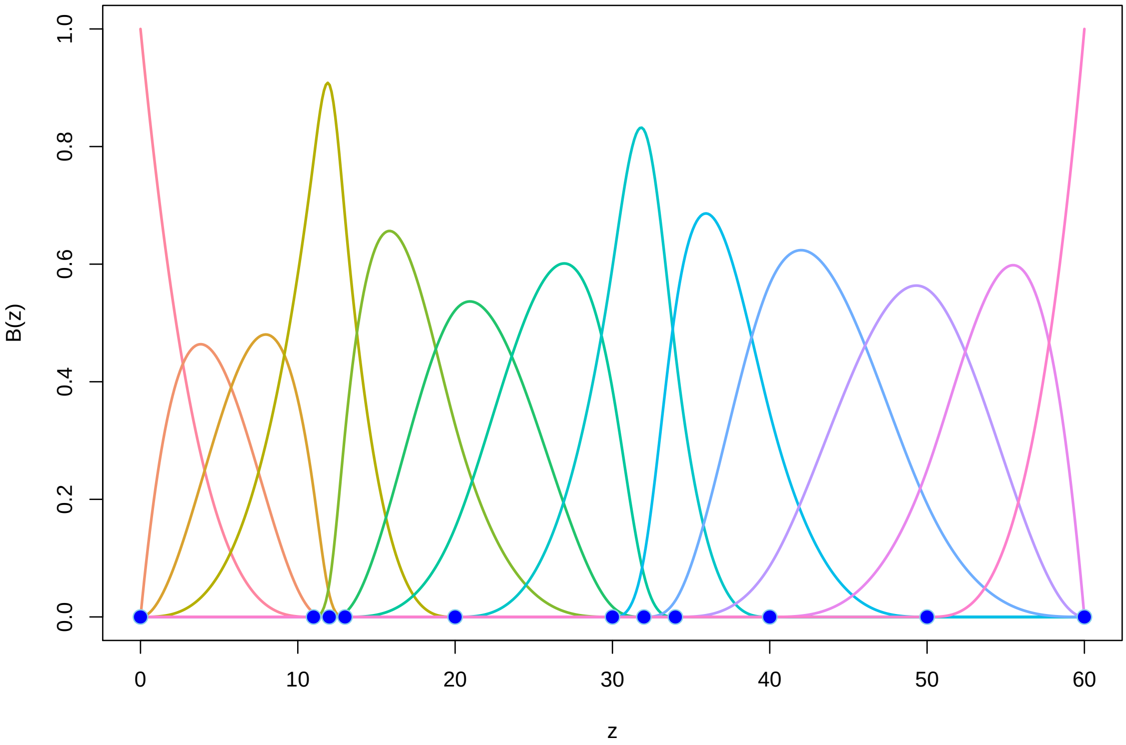
`haccel` $\sim$ `time`**, scatterplot with the LOWESS smoother**



Illustrations

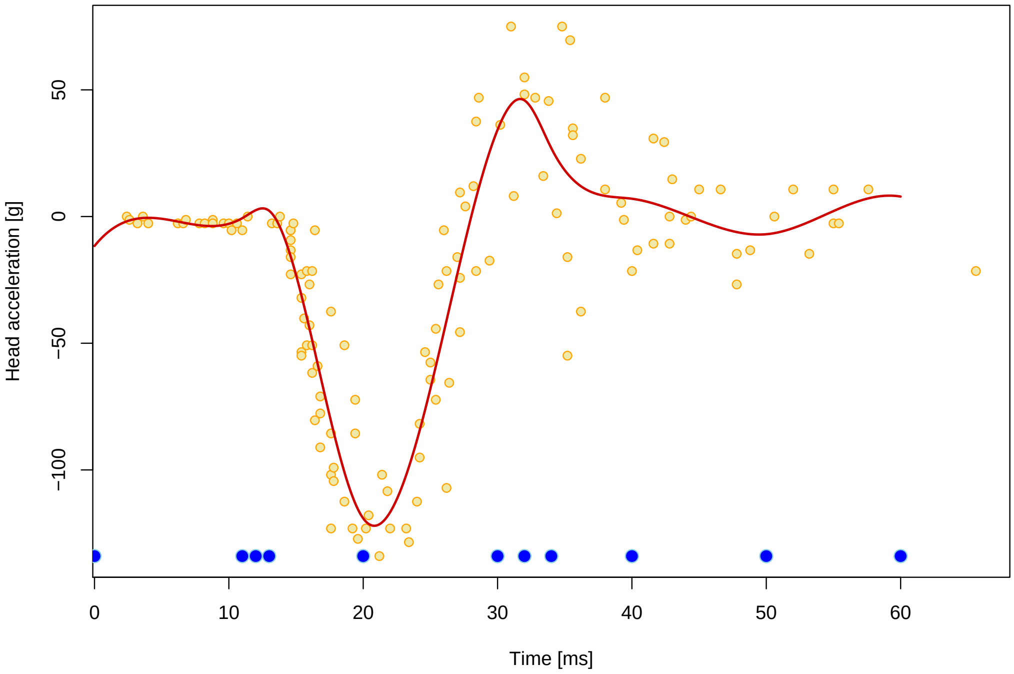
`Motorcycle` $(n = 133)$

B-spline basis (cubic, $d = 3$, $\boldsymbol{\lambda} = \bigl(0,\ 11,\ 12,\ 13,\ 20,\ 30,\ 32,\ 34,\ 40,\ 50,\ 60\bigr)^\top$)



`Motorcycle` $(n = 133)$

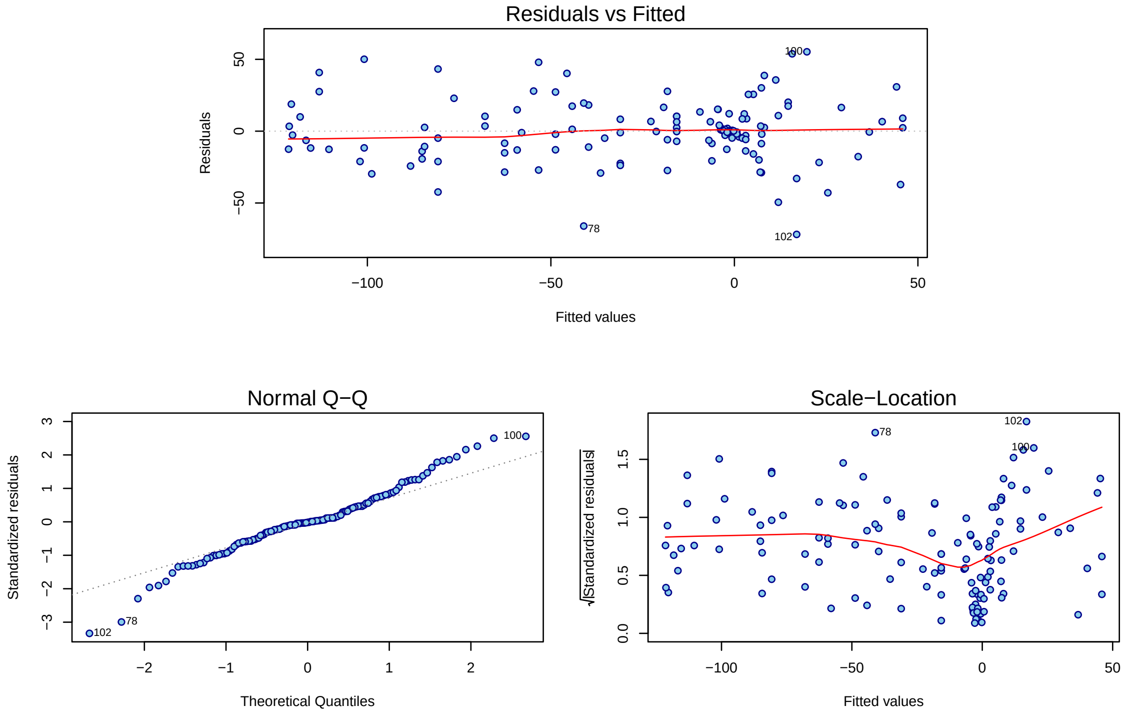
`haccel ~ spline(time)`, $\widehat{m}(z) = -11.62\, B_1(z) + 12.45\, B_2(z) - 13.99\, B_3(z) + 2.99\, B_4(z) + 6.11\, B_5(z) - 237.28\, B_6(z) + 17.34\, B_7(z) + 53.26\, B_8(z) + 5.07\, B_9(z) + 12.72\, B_{10}(z) - 22.00\, B_{11}(z) + 11.37\, B_{12}(z) + 6.97\, B_{13}(z)$



Illustrations

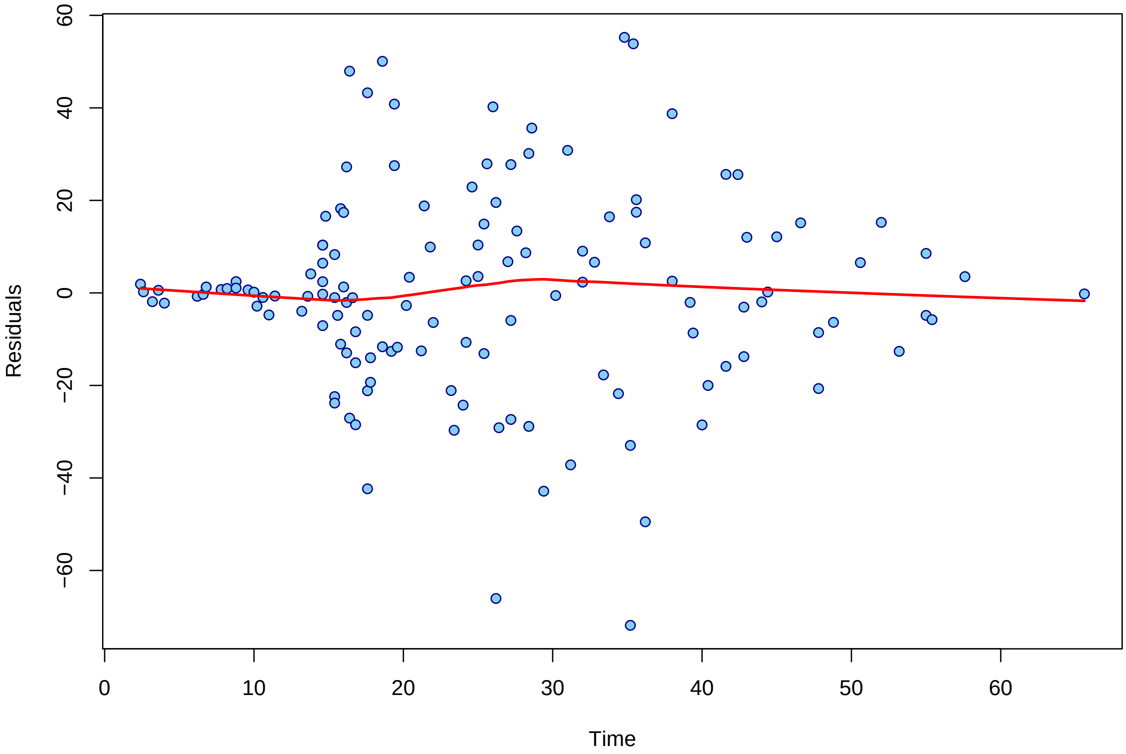
`Motorcycle` ($n = 133$)

`haccel` $\sim$ `spline(time)`, **residual plots**



`Motorcycle` ($n = 133$)

`haccel` $\sim$ `spline(time)`, **residuals versus covariate plot**



4.4 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.,

$$\mathcal{Z} = \bigl\{1, \ldots, G\bigr\}.$$

Unless explicitely stated (in Section 4.4.3), 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.

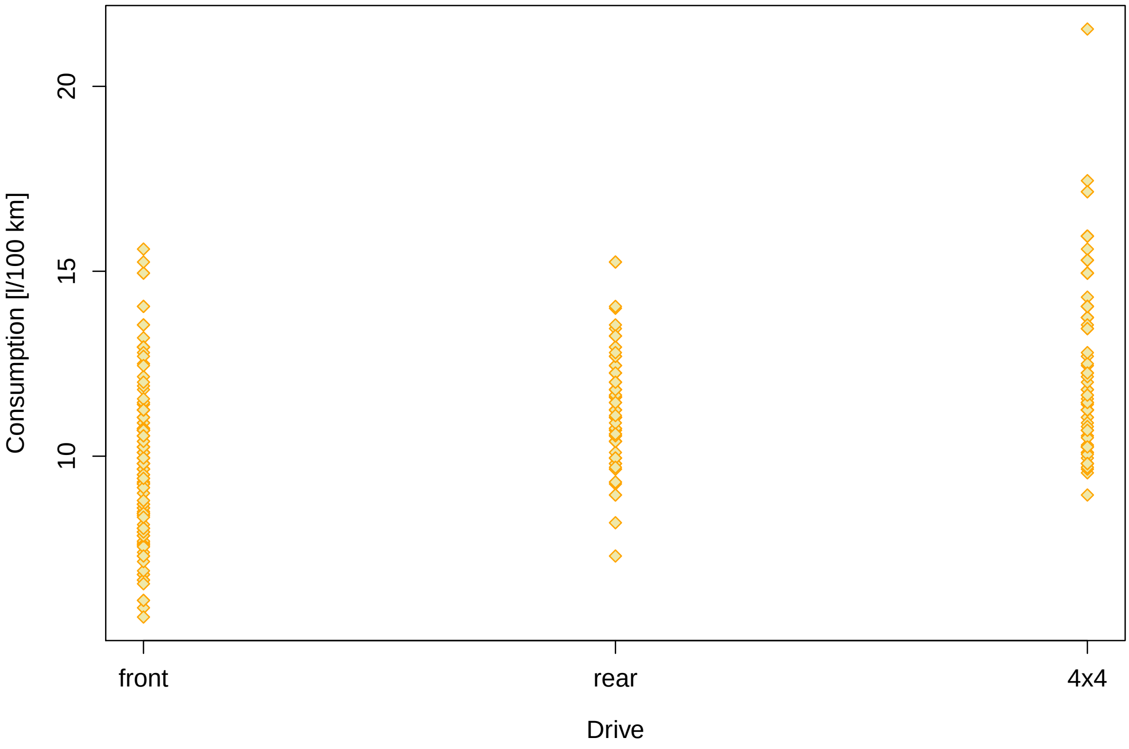
The regression function, $m : \mathcal{Z} \longrightarrow \mathbb{R}$ is now a function defined on a finite set aiming in parameterizing just G (conditional) response expectations $\mathbb{E}\bigl(Y \bigm| Z = 1\bigr), \ldots, \mathbb{E}\bigl(Y \bigm| Z = G\bigr)$. For some clarity in notation, we will also use symbols $m_1, \ldots, m_G$ for those expectations, i.e.,

$$\begin{aligned} m(1) &= \mathbb{E}\bigl(Y \bigm| Z = 1\bigr) &=:& \; m_1, \\ &\vdots & \vdots& \\ m(G) &= \mathbb{E}\bigl(Y \bigm| Z = G\bigr) &=:& \; m_G. \end{aligned}$$

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

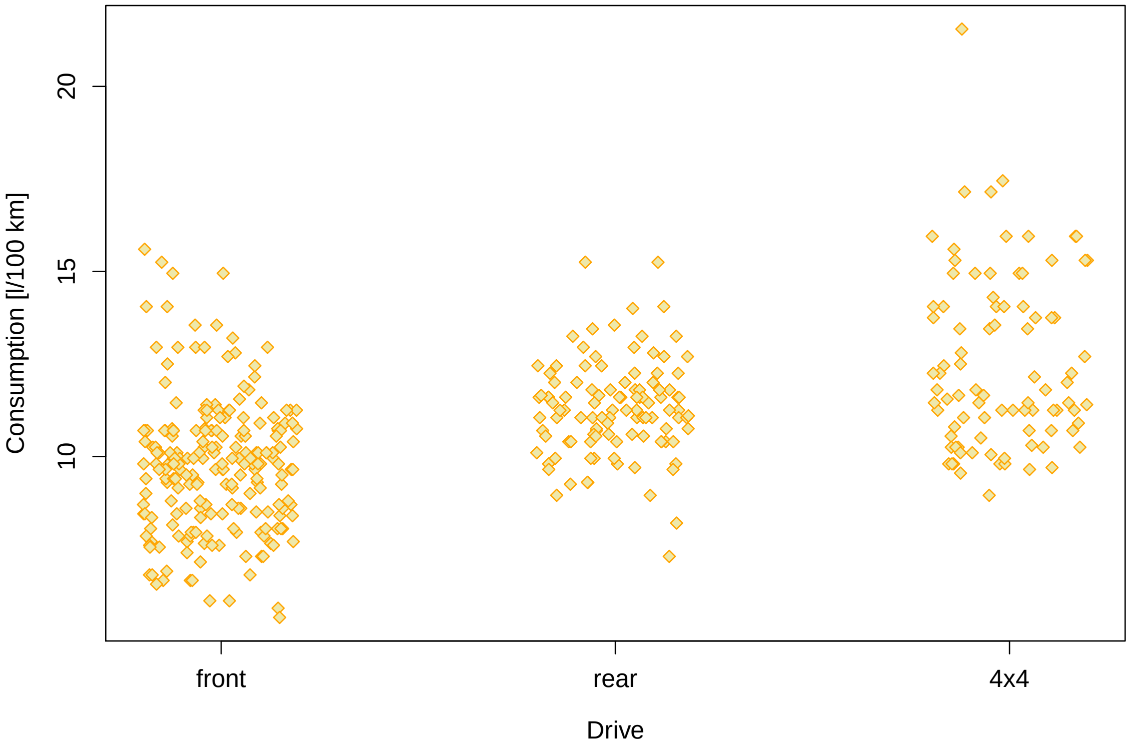
`consumption` $\sim$ `drive`



Illustrations

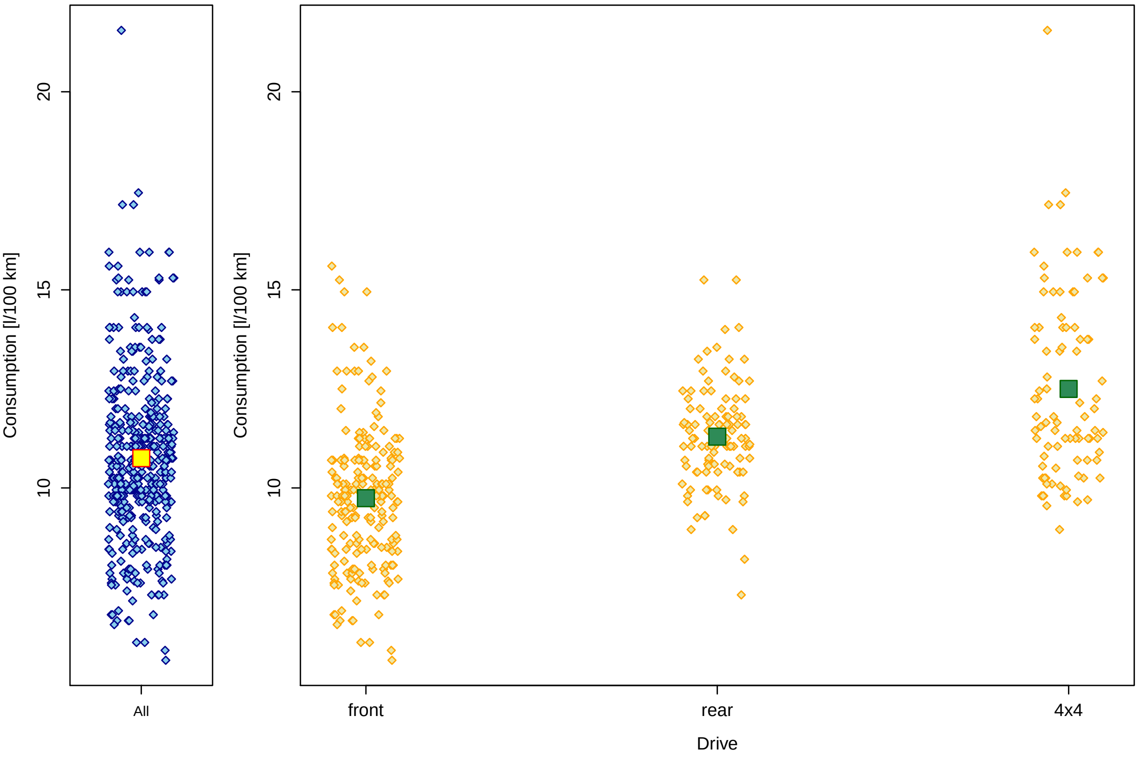
Cars2004nh **(subset,** $n = 409$**)**

`consumption` $\sim$ `drive`



Cars2004nh **(subset,** $n = 409$**)**

`consumption` $\sim$ `drive`



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

$$\boldsymbol{m} = \bigl(m_1,\, \ldots,\, m_G\bigr)^\top$$

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.

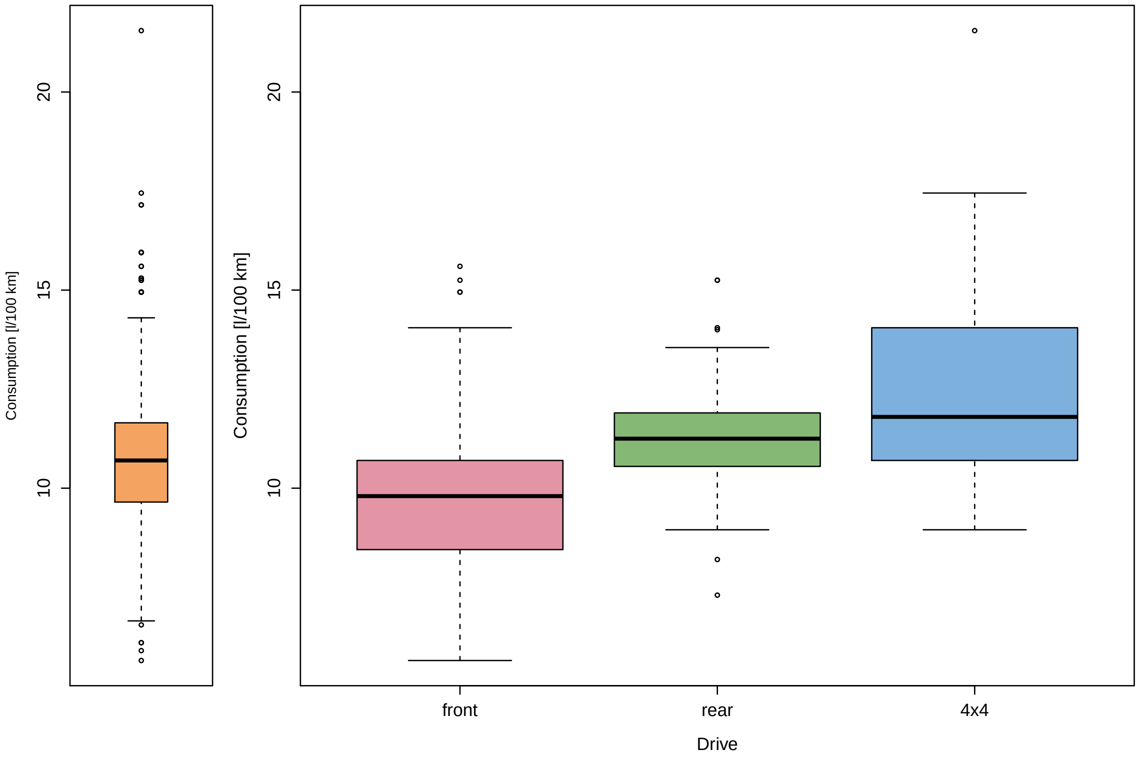
[11] *skupinové střední hodnoty* [12] *vektor skupinových středních hodnot*

4.4.1 Link to a G-sample problem

Illustrations

4.4.1 Link to a G-sample problem

`Cars2004nh` **(subset,** $n = 409$**)**



For following considerations, we will additionally assume (again without loss of generality) that the data $\bigl(Y_i,\, Z_i\bigr)^\top$, $i = 1, \ldots, n$, 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 will indicate the covariate value, i.e.,

$$\mathbb{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_{n_1} \\ - - - \\ \vdots \\ - - - \\ Z_{n-n_G+1} \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ - - \\ \vdots \\ - - \\ G \\ \vdots \\ G \end{pmatrix}, \quad \begin{array}{l} \left.\begin{array}{c} \\ \\ \end{array}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\begin{array}{c} \\ \\ \end{array}\right\} n_G\text{-times} \end{array} \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ - - - - \\ \vdots \\ - - - - \\ Y_{n-n_G+1} \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ - - - \\ \vdots \\ - - - \\ Y_{G,1} \\ \vdots \\ Y_{G,n_G} \end{pmatrix}.$$

Finally, let

$$\boldsymbol{Y}_g = \bigl(Y_{g,1},\, \ldots,\, Y_{g,n_g}\bigr)^\top, \qquad g = 1, \ldots, G,$$

denote a subvector of the response vector that corresponds to observations with the covariate value being equal to g. That is,

$$\boldsymbol{Y} = \bigl(Y_1,\,\ldots,\,Y_n\bigr)^\top = \bigl(\boldsymbol{Y}_1^\top,\,\ldots,\,\boldsymbol{Y}_G^\top\bigr)^\top.$$

Notes.

- Suppose that it can be assumed that $\bigl(Y_i,\,Z_i\bigr)^\top \overset{\text{i.i.d.}}{\sim} \bigl(Y,\,Z\bigr)^\top$, where $\mathbb{E}\bigl(Y \,\big|\, Z = g\bigr) =: m_g$, $\mathsf{var}\bigl(Y \,\big|\, Z = g\bigr) = \sigma^2$, $g = 1,\ldots,G$. 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 $Y \mid Z = g$ whose mean is m_g and the variance is σ^2 (which does not depend on a "group"). That is,

$$\begin{aligned} Y_{1,1},\,\ldots,\,Y_{1,n_1} &\overset{\text{i.i.d.}}{\sim} \bigl(m_1,\,\sigma^2\bigr), \\ &\ \vdots \\ Y_{G,1},\,\ldots,\,Y_{G,n_G} &\overset{\text{i.i.d.}}{\sim} \bigl(m_G,\,\sigma^2\bigr). \end{aligned} \tag{4.12}$$

  Note that (4.12) describes a classical G sample problem where the samples are assumed to be homoscedastic (having the same variance).

- 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).

4.4.2 Linear model parameterization of one-way classified group means

As usual, let $\boldsymbol{\mu}$ be the (conditional) response expectation, i.e.,

$$\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z}\bigr) = \boldsymbol{\mu} := \begin{pmatrix} \mu_{1,1}\\ \vdots \\ \mu_{1,n_1} \\ -- \\ \vdots \\ -- \\ \mu_{G,1}\\ \vdots\\ \mu_{G,n_G}\end{pmatrix} = \begin{pmatrix} m_1 \\ \vdots \\ m_1 \\ -- \\ \vdots \\ -- \\ m_G \\ \vdots \\ m_G\end{pmatrix} \begin{matrix} \left.\begin{matrix}\\ \\ \end{matrix}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\begin{matrix}\\ \\ \end{matrix}\right\} n_G\text{-times}\end{matrix} = \begin{pmatrix} m_1\,\mathbf{1}_{n_1}\\ \vdots \\ m_G\,\mathbf{1}_{n_G}\end{pmatrix}. \tag{4.13}$$

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

$$\mathbb{Q} = \begin{pmatrix} 1 & \ldots & 0\\ \vdots & \vdots & \vdots \\ 1 & \ldots & 0 \\ - & -- & - \\ \vdots & \vdots & \vdots \\ - & -- & - \\ 0 & \ldots & 1 \\ \vdots & \vdots & \vdots \\ 0 & \ldots & 1\end{pmatrix} \begin{matrix} \left.\begin{matrix}\\ \\ \end{matrix}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\begin{matrix}\\ \\ \end{matrix}\right\} n_G\text{-times}\end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \bigl(1,\,\ldots,\,0\bigr)\\ \vdots \\ \mathbf{1}_{n_G} \otimes \bigl(0,\,\ldots,\,1\bigr)\end{pmatrix}. \tag{4.14}$$

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

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_1^\top \\ -- \\ \vdots \\ -- \\ \boldsymbol{x}_G^\top \\ \vdots \\ \boldsymbol{x}_G^\top \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} n_1\text{-times} \\ \\ \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{x}_1^\top \\ \vdots \\ \mathbf{1}_{n_G} \otimes \boldsymbol{x}_G^\top \end{pmatrix}, \tag{4.15}$$

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,

$$\mathsf{rank}\bigl(\mathbb{X}\bigr) = \mathsf{rank}\bigl(\widetilde{\mathbb{X}}\bigr).$$

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

$$\mathsf{rank}\bigl(\widetilde{\mathbb{X}}\bigr) = G.$$

The group means then depend on a vector $\boldsymbol{\beta} = \bigl(\beta_0, \ldots, \beta_{k-1}\bigr)^\top$ 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 (4.14) as a model matrix $\mathbb{X}$. In that case, $\widetilde{\mathbb{X}} = \mathbf{I}_G$ and we have

$$\begin{aligned} \boldsymbol{\mu} &= \mathbb{Q}\,\boldsymbol{\beta}, \\ \boldsymbol{m} &= \boldsymbol{\beta}. \end{aligned} \tag{4.16}$$

Even though parameterization (4.16) 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.

4.4.3 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

$$\mathbb{X} = \begin{pmatrix} 1 & \boldsymbol{c}_1^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{c}_1^\top \\ - & - \,- \\ \vdots & \vdots \\ - & - \,- \\ 1 & \boldsymbol{c}_G^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{c}_G^\top \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \boldsymbol{1}_{n_1} \otimes \bigl(1,\,\boldsymbol{c}_1^\top\bigr) \\ \vdots \\ \boldsymbol{1}_{n_G} \otimes \bigl(1,\,\boldsymbol{c}_G^\top\bigr) \end{pmatrix},$$

where $\boldsymbol{c}_1, \dots, \boldsymbol{c}_G \in \mathbb{R}^{G-1}$ 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

$$\widetilde{\mathbb{X}} = \bigl(\boldsymbol{1}_G,\,\mathbb{C}\bigr).$$

If $\boldsymbol{\beta} = \bigl(\beta_0, \dots, \beta_{G-1}\bigr)^\top \in \mathbb{R}^G$ denote, as usual, a vector of regression coefficients, the group means $\boldsymbol{m}$ are parameterized as

$$\begin{aligned} m_g &= \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, \qquad\qquad g = 1, \dots, G, \\ \boldsymbol{m} = \widetilde{\mathbb{X}}\boldsymbol{\beta} = \bigl(\boldsymbol{1}_G,\,\mathbb{C}\bigr)\boldsymbol{\beta} &= \beta_0\,\boldsymbol{1}_G + \mathbb{C}\boldsymbol{\beta}^Z, \end{aligned} \tag{4.17}$$

where $\boldsymbol{\beta}^Z = \bigl(\beta_1, \dots, \beta_{G-1}\bigr)^\top$ is a non-intercept subvector of the regression coefficients. We will also refer to it as *effects of the covariate* Z. As we know,

$$\mathsf{rank}\bigl(\mathbb{X}\bigr) = \mathsf{rank}\bigl(\widetilde{\mathbb{X}}\bigr) = \mathsf{rank}\bigl((\boldsymbol{1}_G,\,\mathbb{C})\bigr).$$

Hence, to get the model matrix $\mathbb{X}$ of a full-rank ($\mathsf{rank}\bigl(\mathbb{X}\bigr) = G$), the matrix $\mathbb{C}$ must satisfy $\mathsf{rank}\bigl(\mathbb{C}\bigr) = G - 1$ and $\boldsymbol{1}_G \notin \mathcal{M}\bigl(\mathbb{C}\bigr)$. 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 $\boldsymbol{1}_G$.

Definition 4.5 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

$$\mathsf{rank}\bigl(\mathbb{C}\bigr) = G - 1, \qquad \boldsymbol{1}_G \notin \mathcal{M}\bigl(\mathbb{C}\bigr).$$

Terminology *((Pseudo)contrast matrix).*

Columns of matrix $\mathbb{C}$ are often chosen to form a set of $G-1$ contrasts (vectors which elements sum up to zero) from $\mathbb{R}^G$. In this case, we will call the matrix $\mathbb{C}$ as a *contrast matrix.*[13] In other cases, the matrix $\mathbb{C}$ will be called as a *pseudocontrast matrix.*[14]

Note. The (pseudo)contrast matrix $\mathbb{C}$ also determines parameterization of a categorical covariate according to Definition 4.1. Corresponding function $\boldsymbol{s}:\ \mathcal{Z} \longrightarrow \mathbb{R}^{G-1}$ is

$$\boldsymbol{s}(z) = \boldsymbol{c}_z^\top, \qquad z = 1, \dots, G,$$

and the reparameterizing matrix $\mathbb{S}$ is an $n \times (G-1)$ matrix

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{c}_1^\top \\ -- \\ \vdots \\ -- \\ \boldsymbol{c}_G^\top \\ \vdots \\ \boldsymbol{c}_G^\top \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ b \\ c \end{matrix}}\right\} n_1\text{-times} \\ \\ \\ \\ \left.\vphantom{\begin{matrix} a \\ b \\ c \end{matrix}}\right\} n_G\text{-times} \end{matrix} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \mathbf{1}_{n_G} \otimes \boldsymbol{c}_G^\top \end{pmatrix}.$$

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

$$\mathrm{H}_0:\ \beta_1 = 0 \ \ \& \ \ \cdots \ \ \& \ \ \beta_{G-1} = 0, \tag{4.18}$$

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 (4.18) is indeed equivalent to the hypothesis of equality of the group means

$$\mathrm{H}_0:\ m_1 \ = \ \cdots \ = \ m_G. \tag{4.19}$$

If normality of the response is assumed, equivalently an F-test on a submodel (Theorem 8.1) 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 6.2) can be used.

[13] *kontrastová matice* [14] *pseudokontrastová matice*

Notes. The following can be shown with only a little algebra:

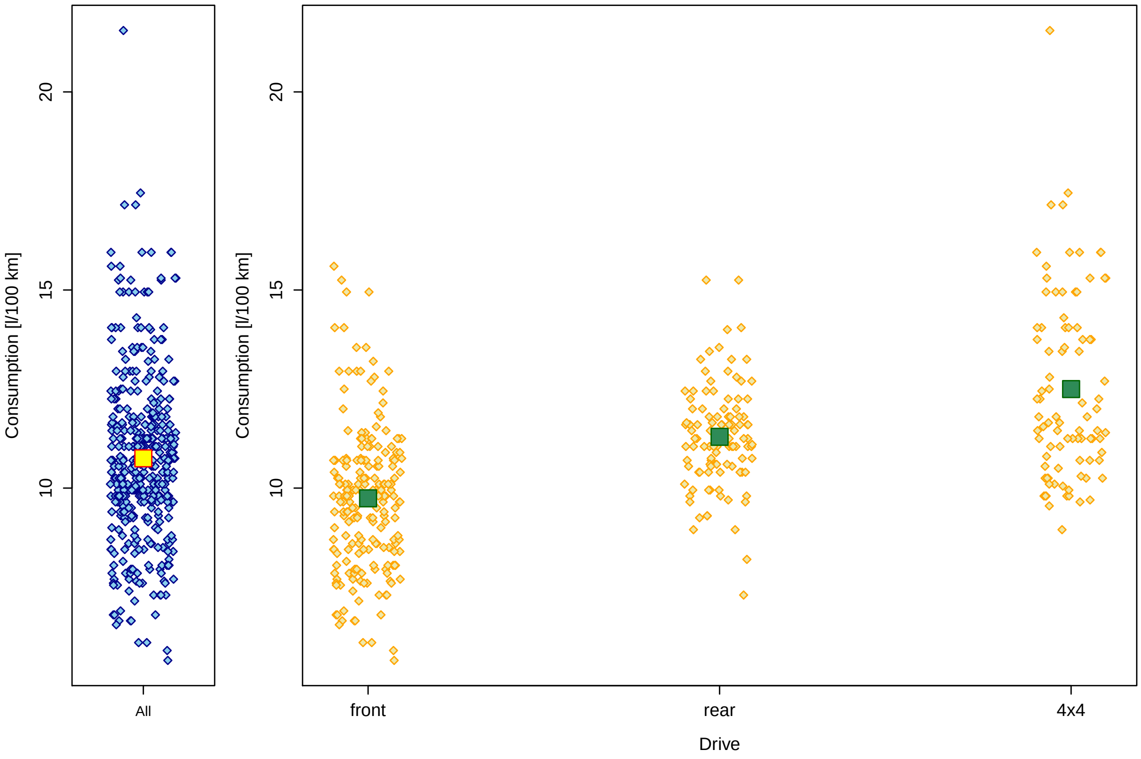
- If $G = 2$, $\boldsymbol{\beta} = \bigl(\beta_0,\, \beta_1\bigr)^\top$. The (usual) t-statistic to test the hypothesis $\mathrm{H}_0 : \beta_1 = 0$ using point (viii) of Theorem 6.2, i.e., the statistic based on the LSE of $\boldsymbol{\beta}$, is the same as a statistic of a standard two-sample t-test.
- If $G \geq 2$, the (usual) F-statistic to test the null hypothesis (4.18) using point (x) of Theorem 6.2 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 (4.19).

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.

Illustrations

`Cars2004nh` **(subset,** $n = 409$**,** $n_{front} = 212$, $n_{rear} = 108$, $n_{4x4} = 89$**)**

$\overline{Y} = 10.75$**,** $\overline{Y}_{front} = 9.74$**,** $\overline{Y}_{rear} = 11.29$**,** $\overline{Y}_{4x4} = 12.50$



Reference group pseudocontrasts (dummy variables)

$$\mathbb{C} = \begin{pmatrix} 0 & \dots & 0 \\ 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{G-1}^\top \\ \mathbf{I}_{G-1} \end{pmatrix} \tag{4.20}$$

Hence, the group means are parameterized as follows and the regression coefficients have the following interpretation

$$\begin{aligned} m_1 &= \beta_0, & \beta_0 &= m_1, \\ m_2 &= \beta_0 + \beta_1, & \beta_1 &= m_2 - m_1, \\ &\vdots & &\vdots \\ m_G &= \beta_0 + \beta_{G-1}, & \beta_{G-1} &= m_G - m_1. \end{aligned} \tag{4.21}$$

The intercept β_0 is equal to the mean of the first (reference) group, the elements of $\boldsymbol{\beta}^Z = \bigl(\beta_1, \dots, \beta_{G-1}\bigr)^\top$ (the effects of Z) provide differences between the means of the remaining groups and the reference one.

The regression function can be written as

$$m(z) = \beta_0 \; + \; \beta_1 \, \mathbb{I}(z = 2) \; + \; \cdots \; + \; \beta_{G-1} \, \mathbb{I}(z = G), \qquad z = 1, \dots, G.$$

That is, the related vector of regressors for each unit in a sample, $\boldsymbol{X} = \bigl(X_0, X_1, \dots, X_{G-1}\bigr)^\top$, is such that $X_j = \mathbb{I}(Z = j + 1)$, $j = 1, \dots, G - 1$. The regressors X_j are also called as *dummy variables*[15] (or shortly *dummies*) in this context.

Notes.

- With the pseudocontrast matrix $\mathbb{C}$ given by (4.20), 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 (4.20) 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.

[15] *umělé proměnné*

Illustrations

`Cars2004nh` **(subset,** $n = 409$**,** $n_{front} = 212$, $n_{rear} = 108$, $n_{4x4} = 89$)
$\overline{Y} = 10.75$, $\overline{Y}_{front} = 9.74$, $\overline{Y}_{rear} = 11.29$, $\overline{Y}_{4x4} = 12.50$

```
CarsNow <- subset(Cars2004nh,
       complete.cases(Cars2004nh[, c("consumption", "lweight", "engine.size")]))
mTrt <- lm(consumption ~ fdrive, data = CarsNow)
summary(mTrt)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.0913 -1.2489 -0.0440  0.9587  9.0511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7413     0.1247  78.149  < 2e-16 ***
fdriverear    1.5527     0.2146   7.237 2.32e-12 ***
fdrive4x4     2.7576     0.2292  12.030  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 406 degrees of freedom
Multiple R-squared:  0.2799,        Adjusted R-squared:  0.2764
F-statistic: 78.91 on 2 and 406 DF,  p-value: < 2.2e-16
```

`Cars2004nh` **(subset,** $n = 409$**,** $n_{front} = 212$, $n_{rear} = 108$, $n_{4x4} = 89$)
$\overline{Y} = 10.75$, $\overline{Y}_{front} = 9.74$, $\overline{Y}_{rear} = 11.29$, $\overline{Y}_{4x4} = 12.50$

```
mSAS <- lm(consumption ~ fdrive, data = CarsNow, contrasts = list(fdrive = contr.SAS))
summary(mSAS)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.0913 -1.2489 -0.0440  0.9587  9.0511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.4989     0.1924  64.969  < 2e-16 ***
fdrive1      -2.7576     0.2292 -12.030  < 2e-16 ***
fdrive2      -1.2049     0.2598  -4.637 4.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 406 degrees of freedom
Multiple R-squared:  0.2799,        Adjusted R-squared:  0.2764
F-statistic: 78.91 on 2 and 406 DF,  p-value: < 2.2e-16
```

Sum contrasts

$$\mathbb{C} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ -1 & \dots & -1 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{G-1} \\ -\mathbf{1}_{G-1}^\top \end{pmatrix} \tag{4.22}$$

In the following, let

$$\overline{m} = \frac{1}{G}\sum_{g=1}^{G} m_g$$

denote the mean of the group means. Those are then parameterized and the regression coefficients have the following interpretation

$$\begin{aligned} & & \beta_0 &= \overline{m}, \\ m_1 &= \beta_0 + \beta_1, & \beta_1 &= m_1 - \overline{m}, \\ &\vdots & &\vdots \\ m_{G-1} &= \beta_0 + \beta_{G-1}, & \beta_{G-1} &= m_{G-1} - \overline{m}. \\ m_G &= \beta_0 - \sum_{g=1}^{G-1}\beta_g, & & \end{aligned} \tag{4.23}$$

The regression function can be written as

$$m(z) = \beta_0 \,+\, \beta_1\,\mathbb{I}(z=1) \,+\, \cdots \,+\, \beta_{G-1}\,\mathbb{I}(z=G-1) \,-\, \Bigl(\sum_{g=1}^{G-1}\beta_g\Bigr)\,\mathbb{I}(z=G),$$
$$z = 1,\,\dots,\,G,$$

which however, is not that interesting now as in the case of previously discussed *reference groups* pseudocontrasts. Much better insight into the *sum contrasts* parameterization is obtained if we write each group mean as

$$m_g \;=\; \alpha_0 \;+\; \alpha_g, \qquad g = 1,\dots,G,$$

with a vector of parameters being $\boldsymbol{\alpha} = \bigl(\alpha_0,\,\alpha_1,\,\dots,\,\alpha_G\bigr)^\top$. This parameterization of G means by $G+1$ parameters would lead to a model matrix with $G+1$ columns whose rank, however, would be only G, i.e., less-than-full-rank model would have been obtained. On the other hand, this less-than-full-rank parameterization can be so called identified by a suitable constraint. In this case, one such constraint is to require

$$\sum_{g=1}^{G}\alpha_g = 0. \tag{4.24}$$

Expression (4.23) shows that the constraint (4.24) is satisfied if we define the α coefficients as follows

$$\begin{aligned} \alpha_0 &= \beta_0 && = \overline{m}, \\ \alpha_1 &= \beta_1 && = \mu_1 - \overline{m}, \\ &\vdots && \vdots \\ \alpha_{G-1} &= \beta_{G-1} && = \mu_{G-1} - \overline{m}, \\ \alpha_G &= -\sum_{g=1}^{G-1} \beta_g && = \mu_G - \overline{m}. \end{aligned} \tag{4.25}$$

This is also the reason why we talk about the *sum constraints* now.

In summary, in this case, the intercept $\alpha_0 = \beta_0$ equals to the mean of the group means and the elements of $\boldsymbol{\beta}^Z = \bigl(\beta_1, \ldots, \beta_{G-1}\bigr)^\top = \bigl(\alpha_1, \ldots, \alpha_{G-1}\bigr)^\top$ 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 (4.22) can be used by the mean of the function `contr.sum`.

Illustrations

`Cars2004nh` (subset, $n = 409$, $n_{front} = 212$, $n_{rear} = 108$, $n_{4x4} = 89$)
$\overline{Y} = 10.75$, $\overline{Y}_{front} = 9.74$, $\overline{Y}_{rear} = 11.29$, $\overline{Y}_{4x4} = 12.50$

```
mSum <- lm(consumption ~ fdrive, data = CarsNow, contrasts = list(fdrive = contr.sum))
summary(mSum)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.17804    0.09606 116.365   <2e-16 ***
fdrive1     -1.43677    0.12003 -11.970   <2e-16 ***
fdrive2      0.11594    0.13926   0.833    0.406

Residual standard error: 1.815 on 406 degrees of freedom
Multiple R-squared:  0.2799,	Adjusted R-squared:  0.2764
F-statistic: 78.91 on 2 and 406 DF,  p-value: < 2.2e-16
```

Values of $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$

```
alphaSum <- as.numeric(contr.sum(3) %*% coef(mSum)[-1])
names(alphaSum) <- levels(CarsNow[, "fdrive"])
print(alphaSum)
```

```
     front       rear        4x4
-1.4367702  0.1159377  1.3208326
```

Weighted sum contrasts

This section of the text contains materials which will not be examined.

$$\mathbb{C} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ -\dfrac{n_1}{n_G} & \dots & -\dfrac{n_{G-1}}{n_G} \end{pmatrix} \tag{4.26}$$

Let

$$\overline{m}_W = \frac{1}{n}\sum_{g=1}^{G} n_g\, m_g.$$

The group means are then parameterized and the regression coefficients have the following interpretation

$$\begin{array}{rclcrcl} & & & & \beta_0 & = & \overline{m}_W, \\ m_1 & = & \beta_0 + \beta_1, & & \beta_1 & = & m_1 - \overline{m}_W, \\ & \vdots & & & & \vdots & \\ m_{G-1} & = & \beta_0 + \beta_{G-1}, & & \beta_{G-1} & = & m_{G-1} - \overline{m}_W. \\ m_G & = & \beta_0 - \displaystyle\sum_{g=1}^{G-1}\frac{n_g}{n_G}\beta_g, & & & & \end{array} \tag{4.27}$$

The regression function can be written as

$$m(z) = \beta_0 \;+\; \beta_1\,\mathbb{I}(z=1) \;+\; \cdots \;+\; \beta_{G-1}\,\mathbb{I}(z=G-1) \;-\; \Bigl(\sum_{g=1}^{G-1}\frac{n_g}{n_G}\beta_g\Bigr)\,\mathbb{I}(z=G), \qquad z = 1,\,\dots,\,G.$$

If we consider the less-than-full-rank parameterization $m_g = \alpha_0 + \alpha_g$, $g = 1,\dots,G$, it is seen from (4.27) that the full-rank parameterization using the contrast matrix (4.26) links the regression coefficients of the two models as

$$\begin{array}{rclcl} \alpha_0 & = & \beta_0 & = & \overline{m}_W, \\ \alpha_1 & = & \beta_1 & = & m_1 - \overline{m}_W, \\ & \vdots & & \vdots & \\ \alpha_{G-1} & = & \beta_{G-1} & = & m_{G-1} - \overline{m}_W, \\ \alpha_G & = & -\displaystyle\sum_{g=1}^{G-1}\frac{n_g}{n_G}\beta_g & = & m_G - \overline{m}_W. \end{array}$$

At the same time, the vector $\boldsymbol{\alpha} = \bigl(\alpha_0,\,\alpha_1,\,\dots,\,\alpha_G\bigr)^\top$ satisfies

$$\sum_{g=1}^{G} n_g\,\alpha_g = 0. \tag{4.28}$$

That is, the full-rank parameterization using the weighted sum pseudocontrasts (4.27) is equivalent to the less-than-full-rank parameterization, where the regression coefficients are identified by the weighted sum constraint (4.28). The intercepts $\alpha_0 = \beta_0$ equal to the *weighted* mean of the group means and the elements of $\boldsymbol{\beta}^Z = \bigl(\beta_1, \ldots, \beta_{G-1}\bigr)^\top = \bigl(\alpha_1, \ldots, \alpha_{G-1}\bigr)^\top$ 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 $\alpha_G = -\sum_{g=1}^{G-1} \frac{n_g}{n_G}\beta_g$.

Illustrations

Cars2004nh (subset, $n = 409$, $n_{front} = 212$, $n_{rear} = 108$, $n_{4x4} = 89$)
$\overline{Y} = 10.75$, $\overline{Y}_{front} = 9.74$, $\overline{Y}_{rear} = 11.29$, $\overline{Y}_{4x4} = 12.50$

```
ng <- with(CarsNow, table(fdrive))
CwSum <- rbind(diag(G - 1), - ng[-G] / ng[G])
rownames(CwSum) <- levels(CarsNow[["fdrive"]])
mwSum <- lm(consumption ~ fdrive, data = CarsNow, contrasts = list(fdrive = CwSum))
summary(mwSum)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.75134    0.08974 119.802  < 2e-16 ***
fdrivefront -1.01007    0.08651 -11.676  < 2e-16 ***
fdriverear   0.54264    0.14982   3.622 0.000329 ***

Residual standard error: 1.815 on 406 degrees of freedom
Multiple R-squared:  0.2799,	Adjusted R-squared:  0.2764
F-statistic: 78.91 on 2 and 406 DF,  p-value: < 2.2e-16
```

Values of $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$

```
alphawSum <- as.numeric(CwSum %*% coef(mwSum)[-1])
names(alphawSum) <- levels(CarsNow[, "fdrive"])
print(alphawSum)
```

```
     front       rear        4x4
-1.0100712  0.5426367  1.7475317
```

Helmert contrasts

This section of the text contains materials which will not be examined.

$$\mathbb{C} = \begin{pmatrix} -1 & -1 & \dots & -1 \\ 1 & -1 & \dots & -1 \\ 0 & 2 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & G-1 \end{pmatrix} \tag{4.29}$$

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

$$\begin{aligned} \beta_0 &= \frac{1}{G}\sum_{g=1}^{G} m_g & &=: \overline{m}, \\ \beta_1 &= \frac{1}{2}(m_2 - m_1), \\ \beta_2 &= \frac{1}{3}\Bigl\{m_3 - \frac{1}{2}(m_1 + m_2)\Bigr\}, \\ \beta_3 &= \frac{1}{4}\Bigl\{m_4 - \frac{1}{3}(m_1 + m_2 + m_3)\Bigr\}, \\ &\vdots \\ \beta_{G-1} &= \frac{1}{G}\Bigl\{m_G - \frac{1}{G-1}\sum_{g=1}^{G-1} m_g\Bigr\}. \end{aligned}$$

which provide their (slightly awkward) interpretation: β_g, $g = 1, \dots, G-1$, is $1/(g+1)$ times the difference between the mean of group $g+1$ and the mean of the means of the previous groups $1, \dots, g$.

Note. In the R software, the Helmert contrasts with the $\mathbb{C}$ matrix being of the form (4.29) can be used by the mean of the function `contr.helmert`.

Illustrations

`Cars2004nh` **(subset,** $n = 409$, $n_{front} = 212$, $n_{rear} = 108$, $n_{4x4} = 89$)
$\overline{Y} = 10.75$, $\overline{Y}_{front} = 9.74$, $\overline{Y}_{rear} = 11.29$, $\overline{Y}_{4x4} = 12.50$

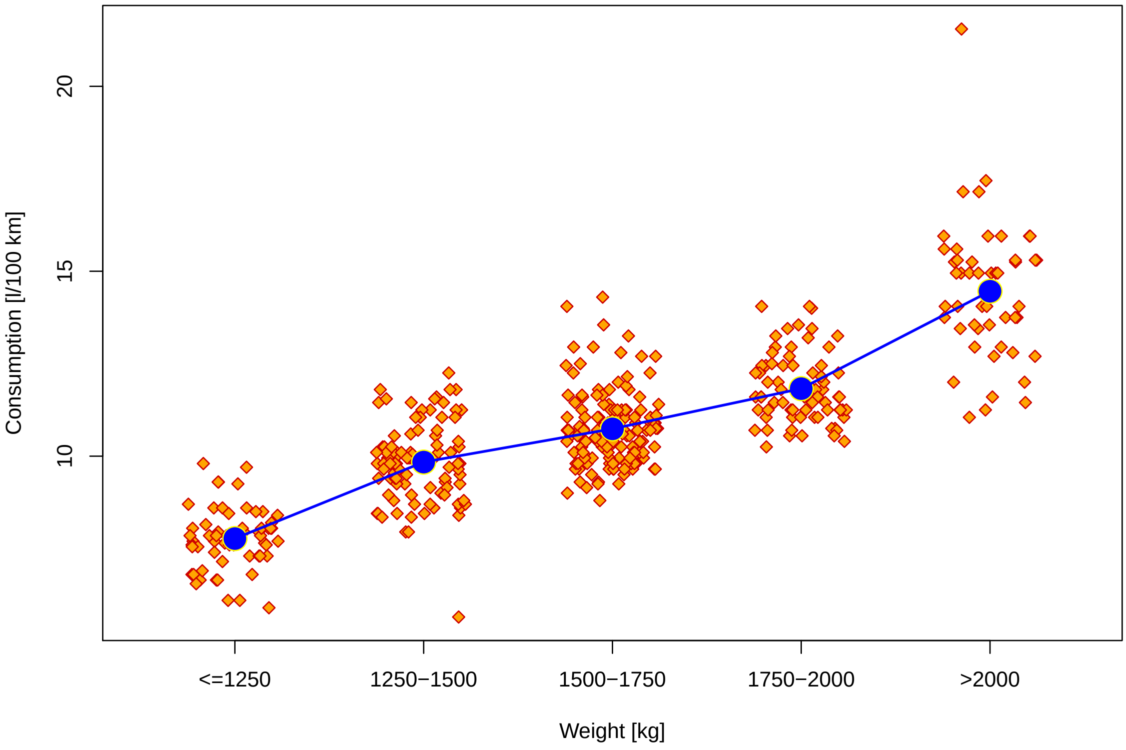
```
mHelmert <- lm(consumption ~ fdrive, data = CarsNow,
               contrasts = list(fdrive = contr.helmert))
summary(mHelmert)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.17804    0.09606 116.365  < 2e-16 ***
fdrive1      0.77635    0.10728   7.237 2.32e-12 ***
fdrive2      0.66042    0.07342   8.995  < 2e-16 ***

Residual standard error: 1.815 on 406 degrees of freedom
Multiple R-squared:  0.2799,	Adjusted R-squared:  0.2764
F-statistic: 78.91 on 2 and 406 DF,  p-value: < 2.2e-16
```

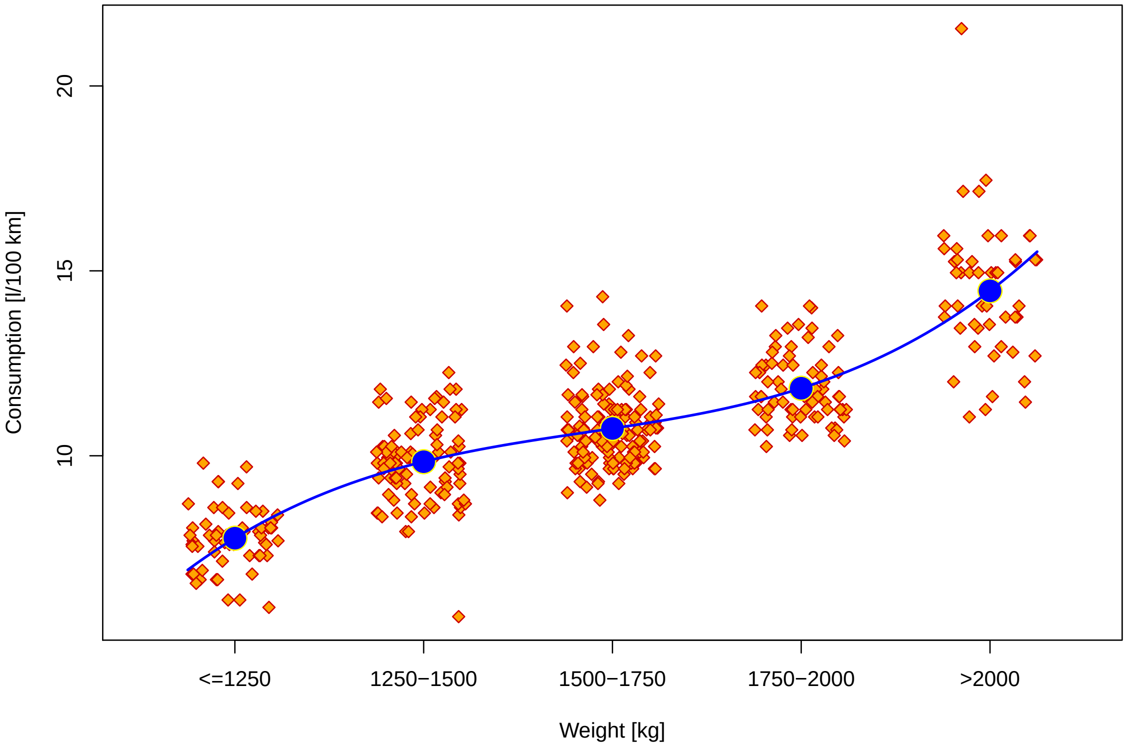
Orthonormal polynomial contrasts

Illustrations

`Cars2004nh` **(subset,** $n = 409$**,** n**'s** $= 57,\ 95,\ 137,\ 71,\ 49$**)**
$\overline{Y} = 10.75$**,** $\overline{Y}_1 = 7.77$**,** $\overline{Y}_2 = 9.84$**,** $\overline{Y}_3 = 10.74$**,** $\overline{Y}_4 = 11.83$**,** $\overline{Y}_5 = 14.46$



`Cars2004nh` **(subset,** $n = 409$**,** n**'s** $= 57,\ 95,\ 137,\ 71,\ 49$**)**
$\overline{Y} = 10.75$**,** $\overline{Y}_1 = 7.77$**,** $\overline{Y}_2 = 9.84$**,** $\overline{Y}_3 = 10.74$**,** $\overline{Y}_4 = 11.83$**,** $\overline{Y}_5 = 14.46$



$$\mathbb{C} = \begin{pmatrix} P^1(\omega_1) & P^2(\omega_1) & \dots & P^{G-1}(\omega_1) \\ P^1(\omega_2) & P^2(\omega_2) & \dots & P^{G-1}(\omega_2) \\ \vdots & \vdots & \vdots & \vdots \\ P^1(\omega_G) & P^2(\omega_G) & \dots & P^{G-1}(\omega_G) \end{pmatrix}, \tag{4.30}$$

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 columns of the $\mathbb{C}$ matrix (4.30) are 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

$$\mathbb{C} = \begin{pmatrix} -\dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} \end{pmatrix},$$

G = 3

$$\mathbb{C} = \begin{pmatrix} -\dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{6}} \\ 0 & -\dfrac{2}{\sqrt{6}} \\ \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{6}} \end{pmatrix},$$

G = 4

$$\mathbb{C} = \begin{pmatrix} -\dfrac{3}{2\sqrt{5}} & \dfrac{1}{2} & -\dfrac{1}{2\sqrt{5}} \\ -\dfrac{1}{2\sqrt{5}} & -\dfrac{1}{2} & \dfrac{3}{2\sqrt{5}} \\ \dfrac{1}{2\sqrt{5}} & -\dfrac{1}{2} & -\dfrac{3}{2\sqrt{5}} \\ \dfrac{3}{2\sqrt{5}} & \dfrac{1}{2} & \dfrac{1}{2\sqrt{5}} \end{pmatrix}.$$

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 (4.30) 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 (4.30) 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

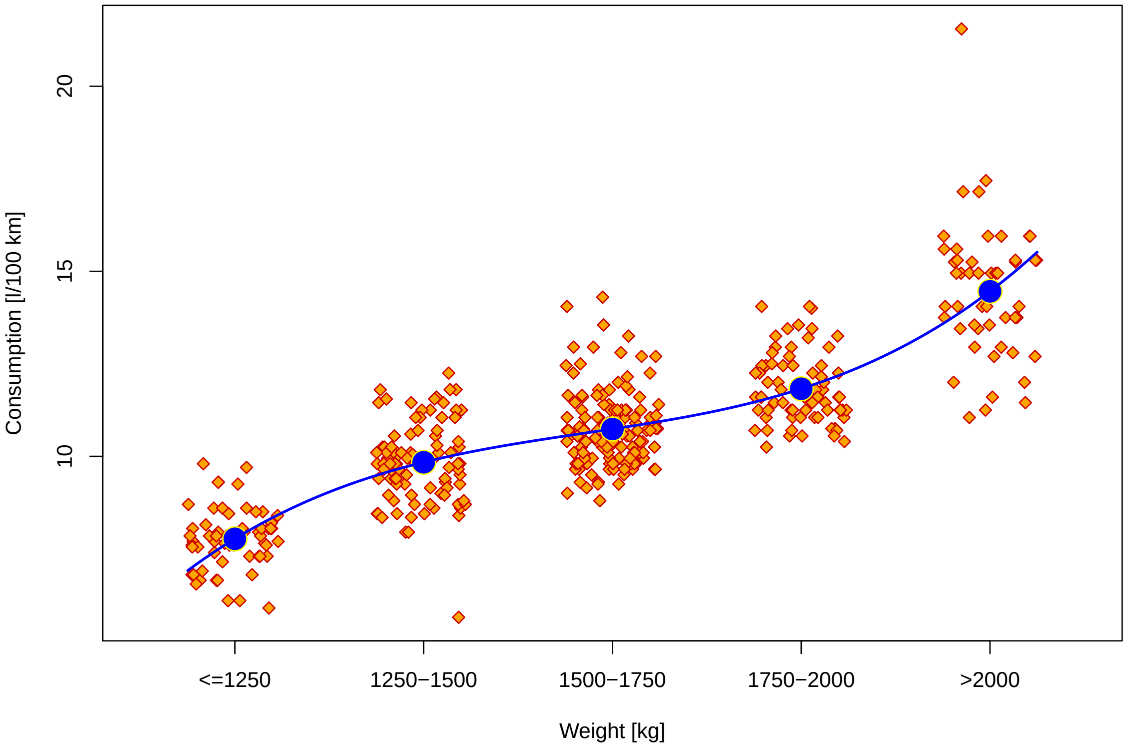
$$\mathrm{H}_0: \; \beta_d = 0 \;\; \& \;\; \ldots \;\; \& \;\; \beta_{G-1} = 0$$

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 (4.30) 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`).

Illustrations

Cars2004nh (subset, $n = 409$, n's $= 57,\ 95,\ 137,\ 71,\ 49$)
$\overline{Y} = 10.75$, $\overline{Y}_1 = 7.77$, $\overline{Y}_2 = 9.84$, $\overline{Y}_3 = 10.74$, $\overline{Y}_4 = 11.83$, $\overline{Y}_5 = 14.46$



Cars2004nh (subset, $n = 409$, n's $= 57,\ 95,\ 137,\ 71,\ 49$)
$\overline{Y} = 10.75$, $\overline{Y}_1 = 7.77$, $\overline{Y}_2 = 9.84$, $\overline{Y}_3 = 10.74$, $\overline{Y}_4 = 11.83$, $\overline{Y}_5 = 14.46$

```
mTrt <- lm(consumption ~ fweight, data = CarsNow)
summary(mTrt)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1900 -0.7102 -0.0400  0.6232  7.0898

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.7719     0.1497   51.91   <2e-16 ***
fweight1250-1500   2.0681     0.1894   10.92   <2e-16 ***
fweight1500-1750   2.9671     0.1782   16.65   <2e-16 ***
fweight1750-2000   4.0548     0.2010   20.17   <2e-16 ***
fweight>2000       6.6883     0.2202   30.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.13 on 404 degrees of freedom
Multiple R-squared:  0.7221,       Adjusted R-squared:  0.7193
F-statistic: 262.4 on 4 and 404 DF,  p-value: < 2.2e-16
```

```
summary(aov(consumption ~ fweight, data = CarsNow))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
fweight       4 1341.0   335.3   262.4 <2e-16 ***
Residuals   404  516.2     1.3
```

Illustrations

`Cars2004nh` **(subset,** $n = 409$**,** n**'s** $= 57,\ 95,\ 137,\ 71,\ 49$**)**
$\overline{Y} = 10.75,\quad \overline{Y}_1 = 7.77$**,** $\overline{Y}_2 = 9.84$**,** $\overline{Y}_3 = 10.74$**,** $\overline{Y}_4 = 11.83$**,** $\overline{Y}_5 = 14.46$

```
mPoly <- lm(consumption ~ fweight, data = CarsNow,
            contrasts = list(fweight = contr.poly))
summary(mPoly)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1900 -0.7102 -0.0400  0.6232  7.0898

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.093e+01  5.975e-02 182.876  < 2e-16 ***
fweight.L    4.858e+00  1.501e-01  32.359  < 2e-16 ***
fweight.Q    3.526e-01  1.370e-01   2.574   0.0104 *
fweight.C    8.585e-01  1.320e-01   6.503 2.33e-10 ***
fweight^4   -7.193e-05  1.126e-01  -0.001   0.9995

Residual standard error: 1.13 on 404 degrees of freedom
Multiple R-squared:  0.7221,        Adjusted R-squared:  0.7193
F-statistic: 262.4 on 4 and 404 DF,  p-value: < 2.2e-16
```

```
summary(aov(consumption ~ fweight, data = CarsNow))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
fweight       4 1341.0   335.3   262.4 <2e-16 ***
Residuals   404  516.2     1.3
```

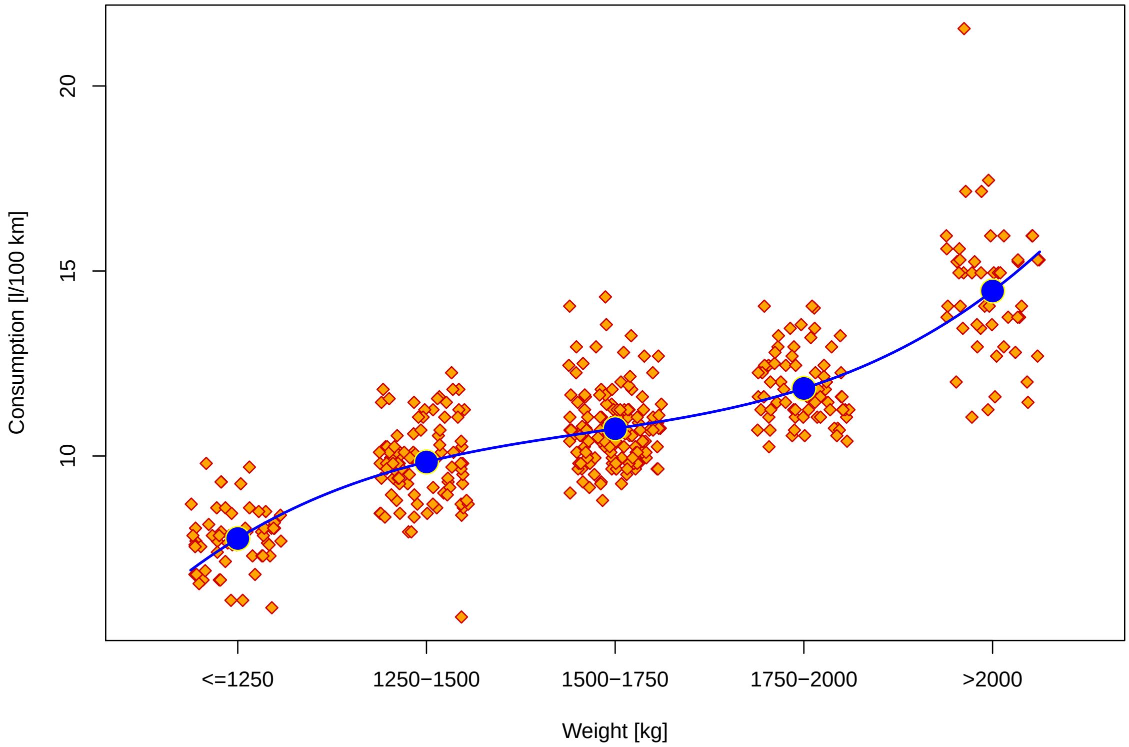
`Cars2004nh` **(subset,** $n = 409$**)**
Polynomial of degree 4 based on representation of the covariate values by numbers 1, 2, 3, 4, 5, $m_g = \beta_0 + \beta_1\, g + \beta_2\, g^2 + \beta_3\, g^3 + \beta_4\, g^4$**,** $g = 1, \ldots, 5$

```
CarsNow <- transform(CarsNow, nweight = as.numeric(fweight))
p4 <- lm(consumption ~ nweight + I(nweight^2) + I(nweight^3) + I(nweight^4),
         data = CarsNow)
summary(p4)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1900 -0.7102 -0.0400  0.6232  7.0898

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.177e+00  1.820e+00   1.745   0.0818 .
nweight       6.312e+00  3.274e+00   1.928   0.0546 .
I(nweight^2) -1.943e+00  1.947e+00  -0.998   0.3190
I(nweight^3)  2.265e-01  4.687e-01   0.483   0.6292
I(nweight^4) -2.507e-05  3.925e-02  -0.001   0.9995

Residual standard error: 1.13 on 404 degrees of freedom
Multiple R-squared:  0.7221,        Adjusted R-squared:  0.7193
F-statistic: 262.4 on 4 and 404 DF,  p-value: < 2.2e-16
```

Illustrations

Cars2004nh (subset, $n = 409$)
Is a linear trend adequate?



Cars2004nh (subset, $n = 409$)
Is a linear trend adequate?

```
p1 <- lm(consumption ~ nweight, data = CarsNow)
anova(p1, p4)
```

```
Analysis of Variance Table

Model 1: consumption ~ nweight
Model 2: consumption ~ nweight + I(nweight^2) + I(nweight^3) + I(nweight^4)
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    407 577.49
2    404 516.20  3    61.291 15.99 7.667e-10 ***
```

```
anova(p1, mPoly)
```

```
Analysis of Variance Table

Model 1: consumption ~ nweight
Model 2: consumption ~ fweight
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    407 577.49
2    404 516.20  3    61.291 15.99 7.667e-10 ***
```

Chapter 5

Multiple Regression

5.1 Multiple covariates in a linear model

This section is not complete in the notes.

5.1.1 Additivity

Definition 5.1 Additivity of the covariate effect.

We say that a covariate Z_1 *acts additively in the regression model with covariates* $\boldsymbol{Z} = \bigl(Z_1,\,\ldots,\,Z_p\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$ *if the regression function is of the form*

$$\mathbb{E}\bigl(Y \bigm| Z_1 = z_1,\, Z_2 = z_2, \ldots, Z_p = z_p\bigr) = m_1(z_1) + m_2(\boldsymbol{z}_{(-1)}), \tag{5.1}$$

where $\boldsymbol{z}_{(-1)} = \bigl(z_2,\,\ldots,\,z_p\bigr)^\top$, $m_1 : \mathbb{R} \longrightarrow \mathbb{R}$ *and* $m_2 : \mathbb{R}^{p-1} \longrightarrow \mathbb{R}$ *are some measurable functions.*

5.1.2 Interactions

Definition 5.2 Interaction terms.

Let $\bigl(Z,\,W\bigr)^\top \in \mathcal{Z} \times \mathcal{W} \subseteq \mathbb{R}^2$ *be two covariates being parameterized using parameterizations* $\boldsymbol{s}_Z : \mathcal{Z} \longrightarrow \mathbb{R}^{k-1}$ *(*$\boldsymbol{s}_Z = \bigl(s_Z^1,\,\ldots,\,s_Z^{k-1}\bigr)^\top$*) and* $\boldsymbol{s}_W : \mathcal{W} \longrightarrow \mathbb{R}^{l-1}$ *(*$\boldsymbol{s}_W = \bigl(s_W^1,\,\ldots,\,s_W^{l-1}\bigr)^\top$*). By interaction terms based on those two parameterizations we mean elements of a vector*

$$\begin{aligned}\boldsymbol{s}_{ZW}(Z,\,W) &:= \boldsymbol{s}_W^\top(W) \otimes \boldsymbol{s}_Z^\top(Z)\\ &= \bigl(s_Z^1(Z)\cdot s_W^1(W),\ldots,s_Z^{k-1}(Z)\cdot s_W^1(W),\,\ldots,\,s_Z^1(Z)\cdot s_W^{l-1}(W),\ldots,s_Z^{k-1}(Z)\cdot s_W^{l-1}(W)\bigr)^\top.\end{aligned}$$

5.2 Numeric and categorical covariate

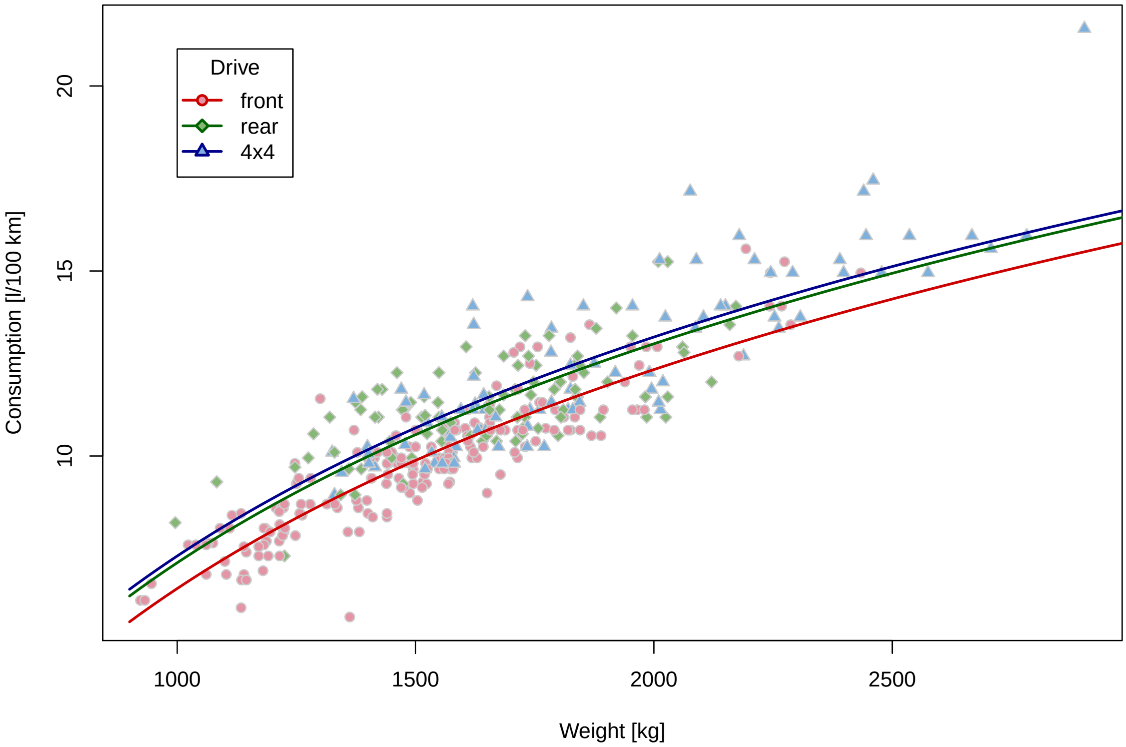
This section is not complete in the notes.

5.2.1 Additivity

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

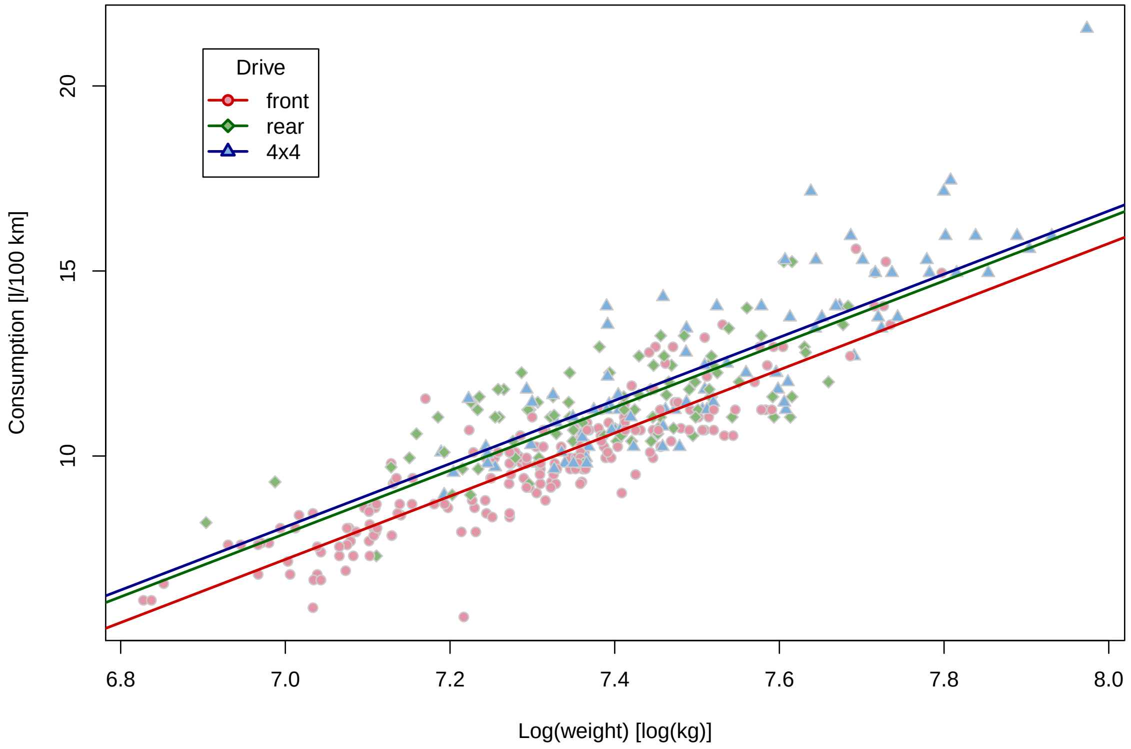
`consumption ~ drive + log(weight)`, $\widehat{m}(z,\,w) = -52.56 + 0.70\,\mathbb{I}[z = \textbf{rear}] + 0.88\mathbb{I}[z = \textbf{4x4}] + 8.54\,\log(w)$



Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption ~ drive + log(weight),` $\widehat{m}(z,\,w) = -52.56 + 0.70\,\mathbb{I}[z = \textbf{rear}] + 0.88\mathbb{I}[z = \textbf{4x4}] + 8.54\,\log(w)$



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption ~ drive + log(weight),` `contr.treatment` **param. of** `drive`

Y**: consumption** ***[l/100 km]***, Z**: drive,** W**: weight** ***[kg]***

$$m(z,\,w) = \beta_0 + \beta_1^Z\,\mathbb{I}[z = \text{rear}] + \beta_2^Z\,\mathbb{I}[z = \text{4x4}] + \beta^W\,\log(w)$$

```
lm(consumption ~ fdrive + lweight, data = CarsNow)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4064 -0.6649 -0.1323  0.5747  5.1533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.5605     1.9627 -26.780  < 2e-16 ***
fdriverear    0.6964     0.1181   5.897 7.83e-09 ***
fdrive4x4     0.8787     0.1363   6.445 3.29e-10 ***
lweight       8.5381     0.2688  31.762  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.9726 on 405 degrees of freedom
Multiple R-squared:  0.7937, Adjusted R-squared:  0.7922
F-statistic: 519.5 on 3 and 405 DF,  p-value: < 2.2e-16
```

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `drive + log(weight)`, <u>`contr.sum`</u> **param. of** `drive`

Y: **consumption** ***[l/100 km]***, Z: **drive,** W: **weight** ***[kg]***

$$m(z,\,w) = \beta_0 + \beta_1^Z\,\mathbb{I}[z = \text{front}] + \beta_2^Z\,\mathbb{I}[z = \text{rear}] - (\beta_1^Z + \beta_2^Z)\,\mathbb{I}[z = \text{4x4}] + \beta^W\,\log(w)$$

```
lm(consumption ~ fdrive + lweight, data = CarsNow,
   contrasts = list(fdrive = "contr.sum"))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4064 -0.6649 -0.1323  0.5747  5.1533

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.03547    1.99090 -26.137  < 2e-16 ***
fdrive1      -0.52504    0.07044  -7.454 5.53e-13 ***
fdrive2       0.17134    0.07465   2.295   0.0222 *
lweight       8.53810    0.26882  31.762  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9726 on 405 degrees of freedom
Multiple R-squared:  0.7937, Adjusted R-squared:  0.7922
F-statistic: 519.5 on 3 and 405 DF,  p-value: < 2.2e-16
```

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `drive + log(weight)`, <u>`contr.sum`</u> **param. of** `drive`

Y: **consumption** ***[l/100 km]***, Z: **drive,** W: **weight** ***[kg]***

$$m(z,\,w) = \beta_0 + \beta_1^Z\,\mathbb{I}[z = \text{front}] + \beta_2^Z\,\mathbb{I}[z = \text{rear}] - (\beta_1^Z + \beta_2^Z)\,\mathbb{I}[z = \text{4x4}] + \beta^W\,\log(w)$$

Estimates of parameters $\alpha_1^Z = \beta_1^Z$, $\alpha_2^Z = \beta_2^Z$, $\alpha_3^Z = -\beta_1^Z - \beta_2^Z$

```
        Estimate Std. Error   t value    P value        Lower      Upper
front -0.5250404 0.07043545 -7.454206 5.5325e-13 -0.66350509 -0.3865756
rear   0.1713353 0.07464863  2.295224   0.022231  0.02458813  0.3180824
4x4    0.3537051 0.08437896  4.191864 3.3999e-05  0.18782965  0.5195805
```

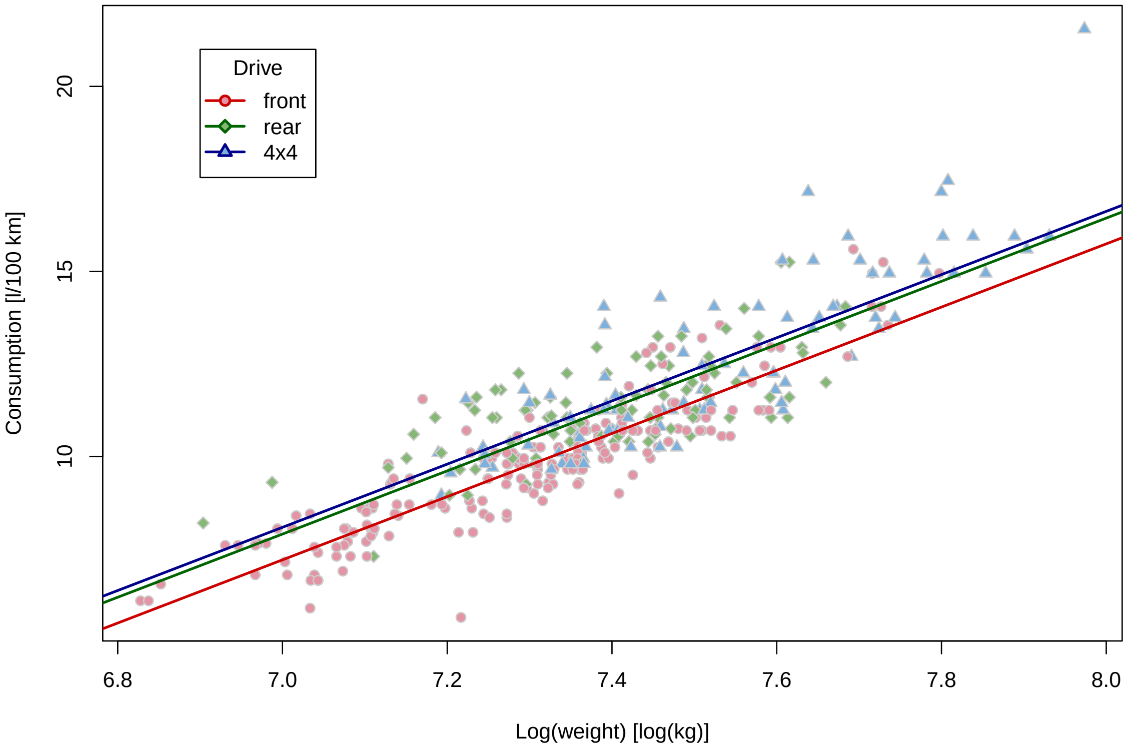
5.2.2 Partial effects

Note that the following tests and estimated effects make only sense if it can be assumed that the *additivity* model holds.

Illustrations

Cars2004nh **(subset,** $n = 409$)

consumption ∼ drive + log(weight), **partial effect of log(**weight)**?**



Cars2004nh **(subset,** $n = 409$)

consumption ∼ drive + log(weight)

For a given** drive**, does the log(**weight**) have an effect on the mean consumption? Partial effect of log(**weight**)

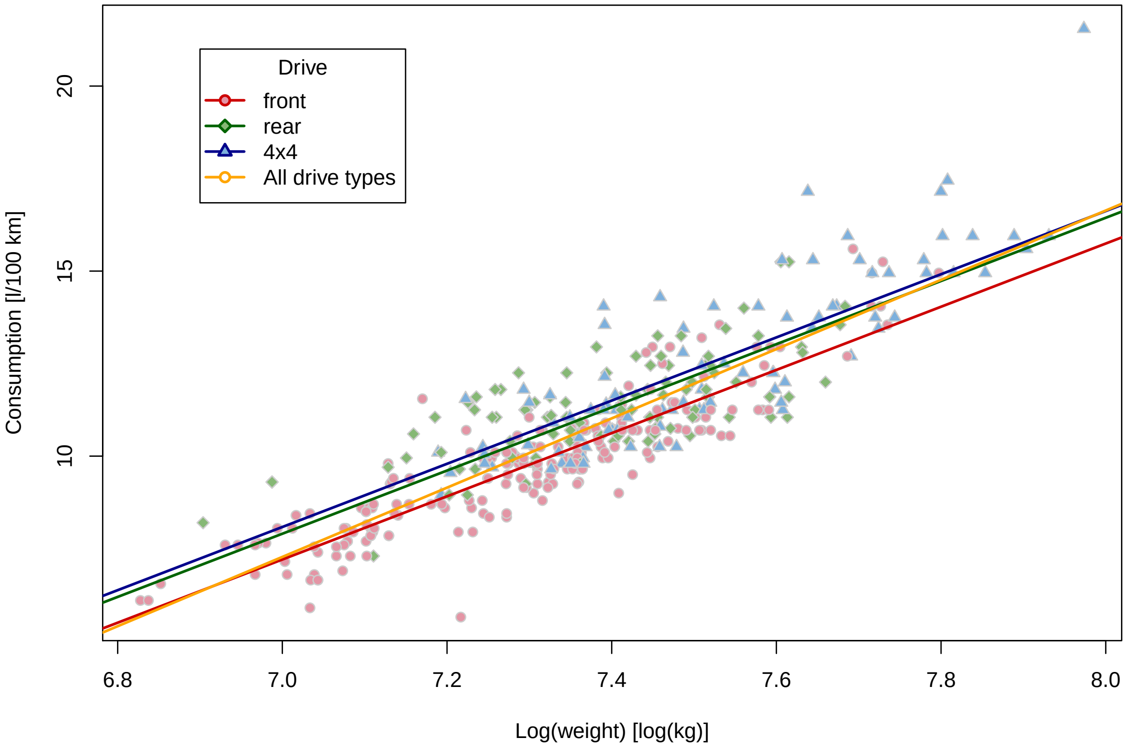
```
lm(consumption ~ fdrive + lweight, data = CarsNow)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4064 -0.6649 -0.1323  0.5747  5.1533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.5605     1.9627 -26.780  < 2e-16 ***
fdriverear    0.6964     0.1181   5.897 7.83e-09 ***
fdrive4x4     0.8787     0.1363   6.445 3.29e-10 ***
lweight       8.5381     0.2688  31.762  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9726 on 405 degrees of freedom
Multiple R-squared:  0.7937, Adjusted R-squared:  0.7922
F-statistic: 519.5 on 3 and 405 DF,  p-value: < 2.2e-16
```

Illustrations

Cars2004nh **(subset,** $n = 409$**)**

consumption ∼ drive + log(weight), **partial effect of** drive**?**



Cars2004nh **(subset,** $n = 409$**)**

consumption ∼ drive + log(weight)

Analysis of covariance to evaluate effect of *drive* ***given log(****weight****)***

```
mAddit <- lm(consumption ~ fdrive + lweight, data = CarsNow)
mOneLine <- lm(consumption ~ lweight, data = CarsNow)
anova(mOneLine, mAddit)
```

```
Analysis of Variance Table

Model 1: consumption ~ lweight
Model 2: consumption ~ fdrive + lweight
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    407 435.68
2    405 383.10  2    52.577 27.791 4.896e-12 ***
```

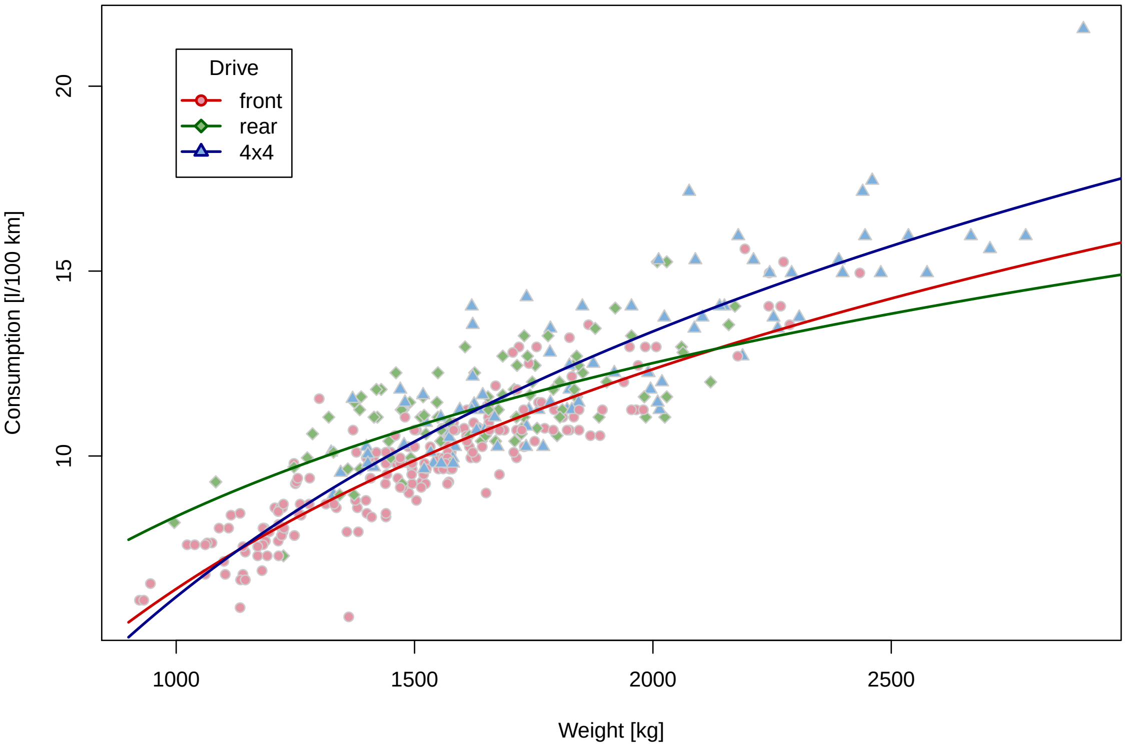
5.2.3 Interactions

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption ~ drive + log(weight) + drive:log(weight),`

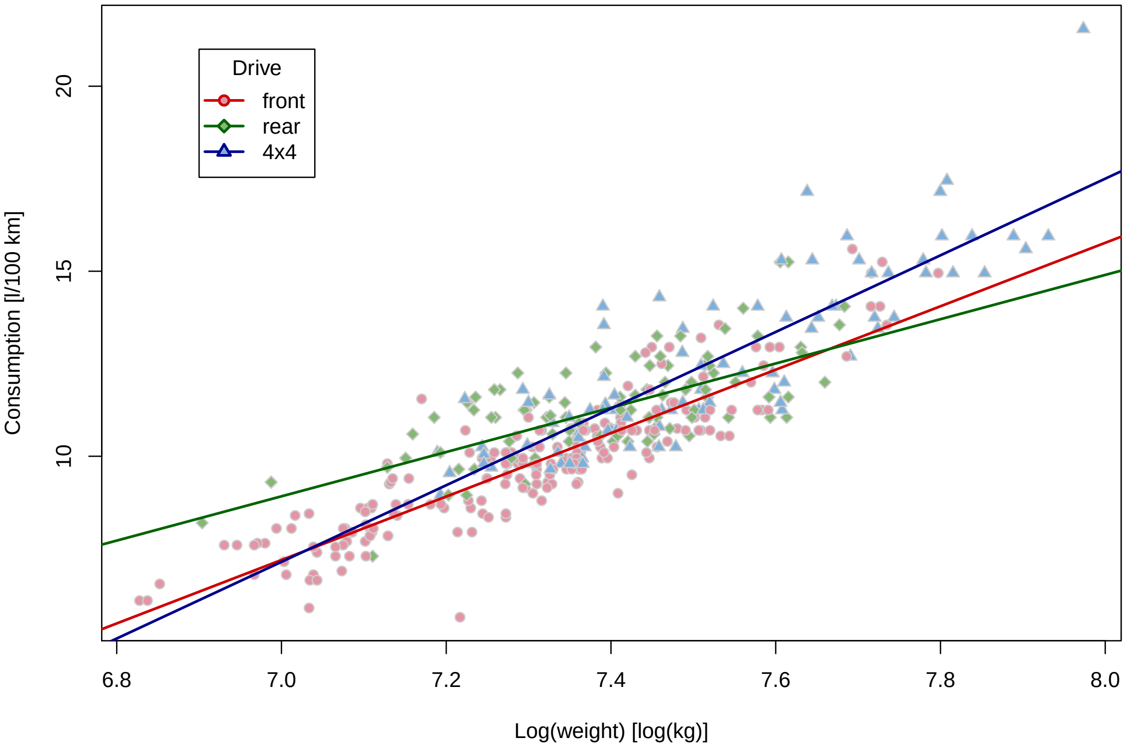
$\widehat{m}(z, w) = -52.80 + 19.84\,\mathbb{I}[z = \textbf{rear}] - 12.54\mathbb{I}[z = \textbf{4x4}] + 8.57\,\log(w) - 2.59\,\mathbb{I}[z = \textbf{rear}]\,\log(w) + 1.78\,\mathbb{I}[z = \textbf{4x4}]\,\log(w)$



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption ~ drive + log(weight) + drive:log(weight),`

$\widehat{m}(z, w) = -52.80 + 19.84\,\mathbb{I}[z = \textbf{rear}] - 12.54\mathbb{I}[z = \textbf{4x4}] + 8.57\,\log(w) - 2.59\,\mathbb{I}[z = \textbf{rear}]\,\log(w) + 1.78\,\mathbb{I}[z = \textbf{4x4}]\,\log(w)$



Illustrations

Cars2004nh **(subset,** $n = 409$)

consumption ~ drive + log(weight) + drive:log(weight), contr.treatment **param. of** drive

Reference group pseudocontrasts for drive

$$m(z,\,w) = \beta_0 + \beta_1^Z\,\mathbb{I}[z = \text{rear}] + \beta_2^Z\,\mathbb{I}[z = \text{4x4}] + \beta^W\ \log(w)$$
$$+\,\beta_1^{ZW}\,\mathbb{I}[z = \text{rear}]\ \log(w) + \beta_2^{ZW}\,\mathbb{I}[z = \text{4x4}]\ \log(w)$$

```
lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -52.8047     2.5266 -20.900  < 2e-16 ***
fdriverear          19.8445     5.1297   3.869 0.000128 ***
fdrive4x4          -12.5366     4.6506  -2.696 0.007319 **
lweight              8.5716     0.3461  24.763  < 2e-16 ***
fdriverear:lweight  -2.5890     0.6956  -3.722 0.000226 ***
fdrive4x4:lweight    1.7837     0.6240   2.858 0.004480 **
---

Residual standard error: 0.9404 on 403 degrees of freedom
Multiple R-squared:  0.8081, Adjusted R-squared:  0.8057
F-statistic: 339.4 on 5 and 403 DF,  p-value: < 2.2e-16
```

Cars2004nh **(subset,** $n = 409$)

consumption ~ drive + log(weight) + drive:log(weight), contr.sum **param. of** drive

Sum contrasts for drive

$$m(z,\,w) = \beta_0 + \beta_1^Z\,\mathbb{I}[z = \text{front}] + \beta_2^Z\,\mathbb{I}[z = \text{rear}] - (\beta_1^Z + \beta_2^Z)\,\mathbb{I}[z = \text{4x4}] + \beta^W\ \log(w)$$
$$+\,\beta_1^{ZW}\,\mathbb{I}[z = \text{front}]\ \log(w) + \beta_2^{ZW}\,\mathbb{I}[z = \text{rear}]\ \log(w) - (\beta_1^{ZW} + \beta_2^{ZW})\,\mathbb{I}[z = \text{4x4}]\ \log(w)$$

```
lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow,
   contrasts = list(fdrive = contr.sum))
```

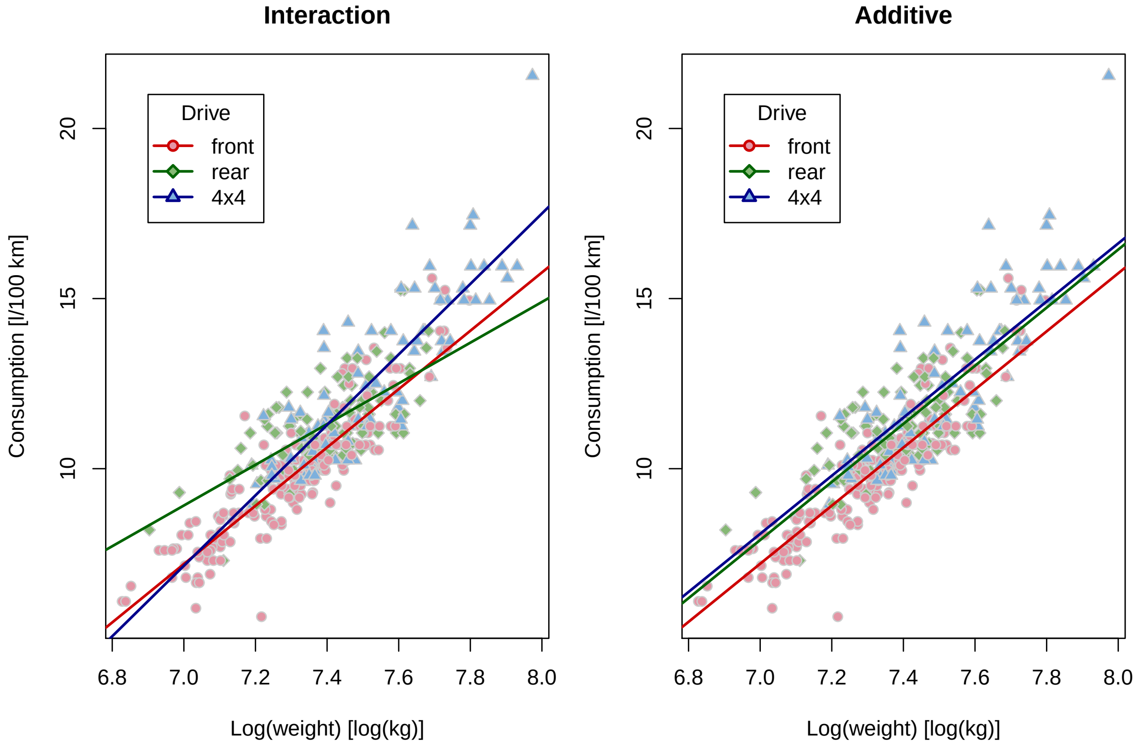
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -50.3688     2.1489 -23.440  < 2e-16 ***
fdrive1          -2.4360     2.5972  -0.938    0.349
fdrive2          17.4085     3.3558   5.188 3.38e-07 ***
lweight           8.3031     0.2894  28.696  < 2e-16 ***
fdrive1:lweight   0.2684     0.3517   0.763    0.446
fdrive2:lweight  -2.3206     0.4529  -5.124 4.64e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9404 on 403 degrees of freedom
Multiple R-squared:  0.8081, Adjusted R-squared:  0.8057
F-statistic: 339.4 on 5 and 403 DF,  p-value: < 2.2e-16
```

5.2.4 Additivity or interactions?

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` ∼ `drive, log(weight)`, **additivity or interactions?**



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` ∼ `drive, log(weight)`, **additivity or interactions?**

Does the log(`weight`) have different effect on the mean consumption depending on the `drive` type?

```
mInter <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow)
mAddit <- lm(consumption ~ fdrive + lweight, data = CarsNow)
anova(mAddit, mInter)
```

```
Analysis of Variance Table

Model 1: consumption ~ fdrive + lweight
Model 2: consumption ~ fdrive + lweight + fdrive:lweight
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    405 383.1
2    403 356.4  2    26.702 15.097 4.758e-07 ***
```

5.2.5 More complex parameterizations of a numeric covariate

5.3 Two numeric covariates

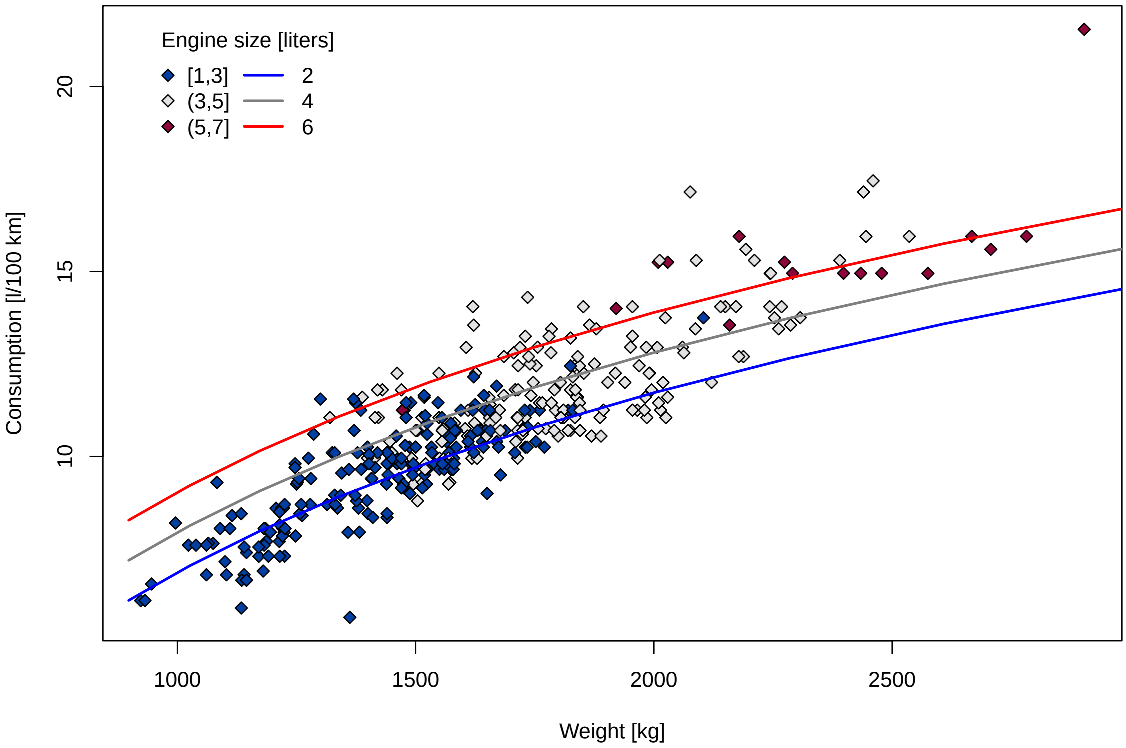
This section is not complete in the notes.

5.3.1 Additivity

Illustrations

`Cars2004nh` **(subset, $n = 409$)**

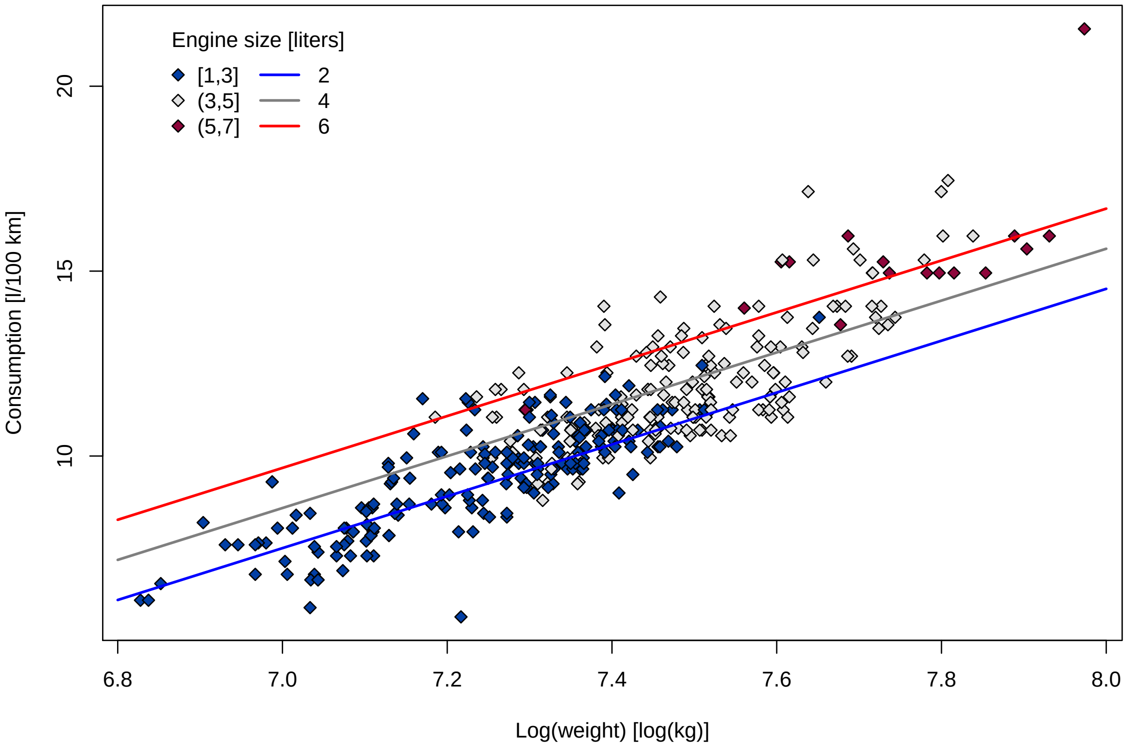
`consumption` $\sim$ `engine.size + log(weight),` $\widehat{m}(z, w) = -42.65 + 0.54\, z + 7.01\, \log(w)$



Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight),` $\widehat{m}(z,\,w) = -42.65 + 0.54\,z + 7.01\,\log(w)$



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight)`

Y: **consumption** ***[l/100 km]***, Z: **engine size** ***[l]***, W: **weight** ***[kg]***

$$m(z,\,w) = \beta_0 + \beta^Z\,z + \beta^W\,\log(w)$$

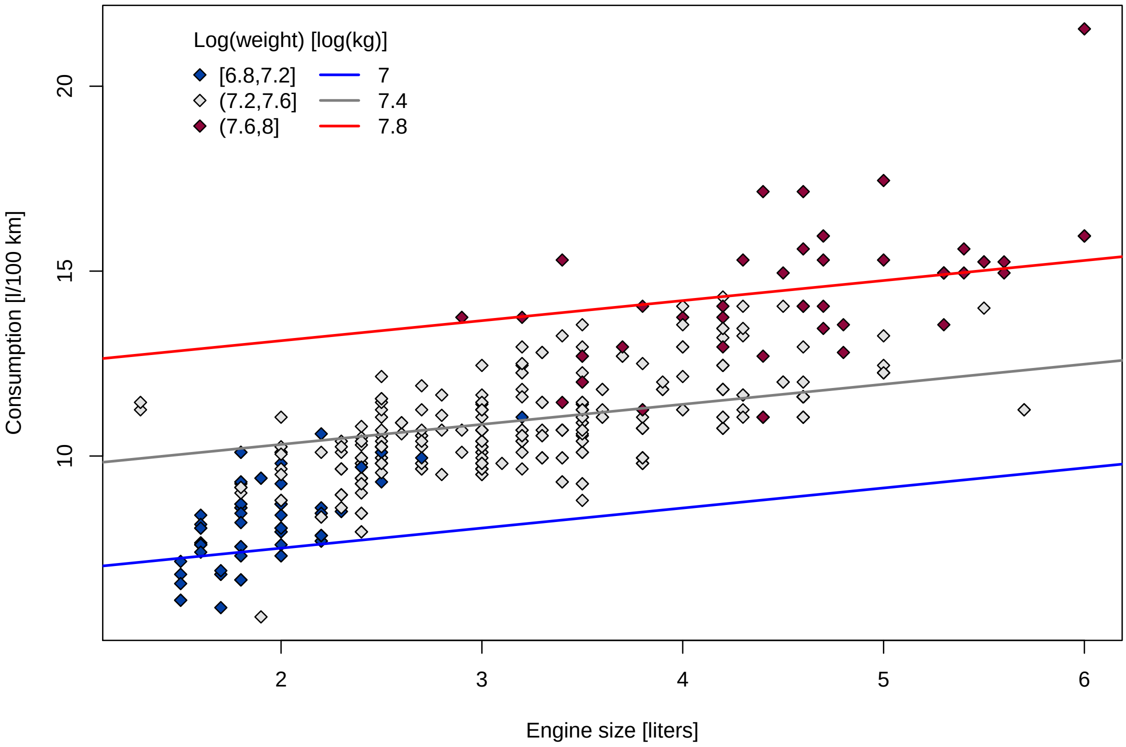
```
lm(consumption ~ engine.size + lweight, data = CarsNow)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3243 -0.6741 -0.1286  0.5270  5.0459

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.65641    2.99243 -14.255  < 2e-16 ***
engine.size   0.54231    0.08304   6.531 1.96e-10 ***
lweight       7.01155    0.43501  16.118  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9854 on 406 degrees of freedom
Multiple R-squared:  0.7877, Adjusted R-squared:  0.7867
F-statistic: 753.3 on 2 and 406 DF,  p-value: < 2.2e-16
```

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight)`, $\quad \widehat{m}(z,\,w) = -42.65 + 0.54\,z + 7.01\,\log(w)$



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight)`

Y**: consumption** ***[l/100 km]***, Z**: engine size** ***[l]***, W**: weight** ***[kg]***

$$m(z,\,w) = \beta_0 + \beta^Z\,z + \beta^W\,\log(w)$$

```
lm(consumption ~ engine.size + lweight, data = CarsNow)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.3243 -0.6741 -0.1286  0.5270  5.0459 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -42.65641    2.99243 -14.255  < 2e-16 ***
engine.size   0.54231    0.08304   6.531 1.96e-10 ***
lweight       7.01155    0.43501  16.118  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9854 on 406 degrees of freedom
Multiple R-squared:  0.7877, Adjusted R-squared:  0.7867 
F-statistic: 753.3 on 2 and 406 DF,  p-value: < 2.2e-16
```

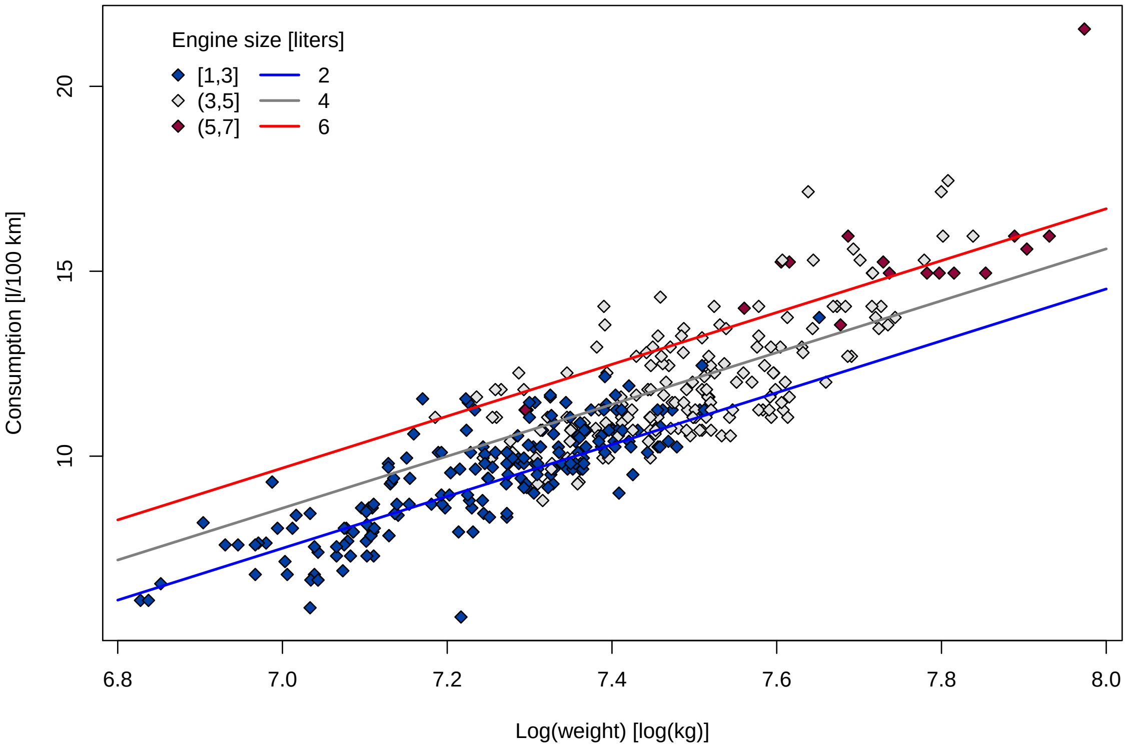
5.3.2 Partial effects

Note that the following tests and estimated effects make only sense if it can be assumed that the *additivity* model holds.

Illustrations

`Cars2004nh` **(subset,** $n = 409$)

`consumption ~ engine.size + log(weight)`, **partial effect of log**`(weight)`**?**



`Cars2004nh` **(subset,** $n = 409$)

`consumption ~ engine.size + log(weight)`

Y**: consumption** ***[l/100 km]***, Z**: engine size** ***[l]***, W**: weight** ***[kg]***

$$m(z, w) = \beta_0 + \beta^Z z + \beta^W \log(w)$$

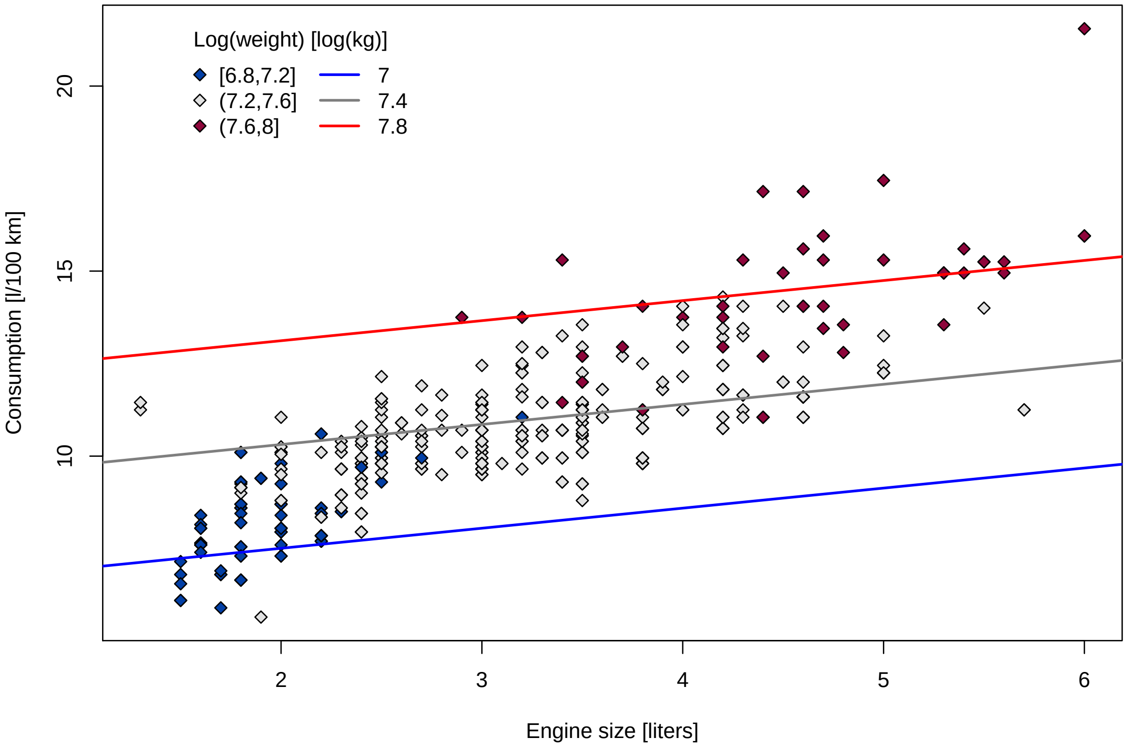
```
lm(consumption ~ engine.size + lweight, data = CarsNow)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.3243 -0.6741 -0.1286  0.5270  5.0459 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -42.65641    2.99243 -14.255  < 2e-16 ***
engine.size   0.54231    0.08304   6.531 1.96e-10 ***
lweight       7.01155    0.43501  16.118  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.9854 on 406 degrees of freedom
Multiple R-squared:  0.7877, Adjusted R-squared:  0.7867 
F-statistic: 753.3 on 2 and 406 DF,  p-value: < 2.2e-16
```

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight)`, **partial effect of engine.size?**



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight)`

Y**: consumption** ***[l/100 km]***, Z**: engine size** ***[l]***, W**: weight** ***[kg]***

$$m(z,\,w) = \beta_0 + \beta^Z\,z + \beta^W\,\log(w)$$

```
lm(consumption ~ engine.size + lweight, data = CarsNow)
```

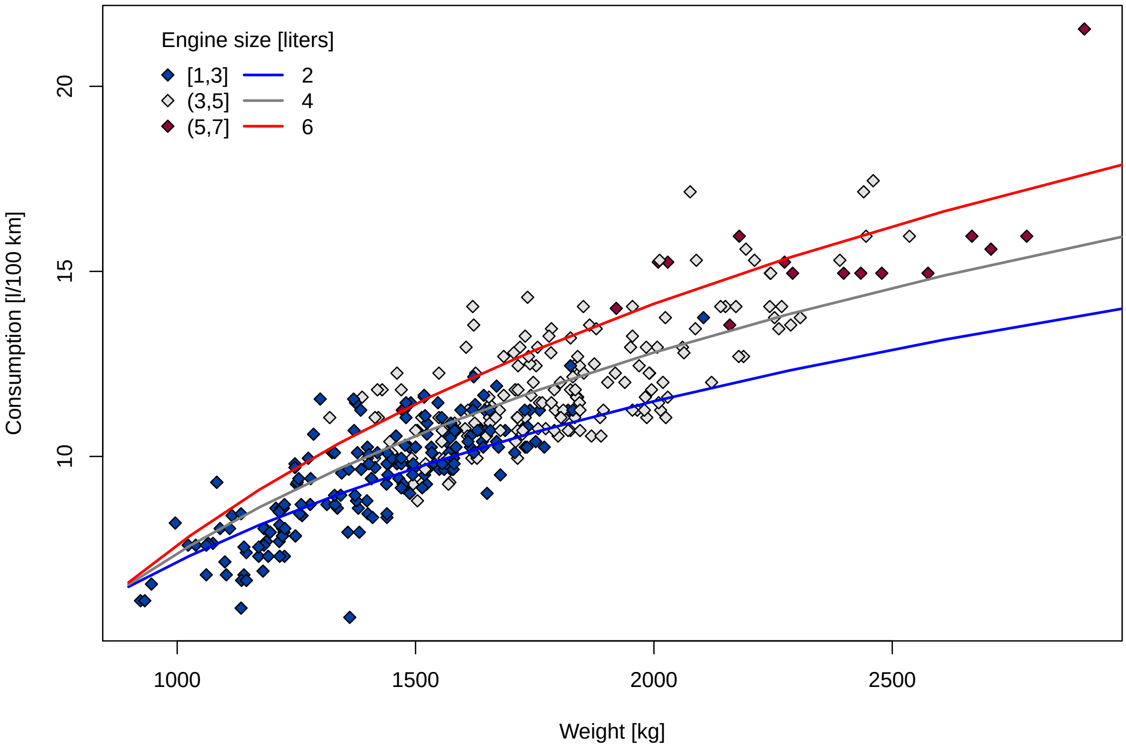
```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3243 -0.6741 -0.1286  0.5270  5.0459

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.65641    2.99243 -14.255  < 2e-16 ***
engine.size   0.54231    0.08304   6.531 1.96e-10 ***
lweight       7.01155    0.43501  16.118  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.9854 on 406 degrees of freedom
Multiple R-squared:  0.7877, Adjusted R-squared:  0.7867
F-statistic: 753.3 on 2 and 406 DF,  p-value: < 2.2e-16
```

5.3.3 Interactions

Illustrations

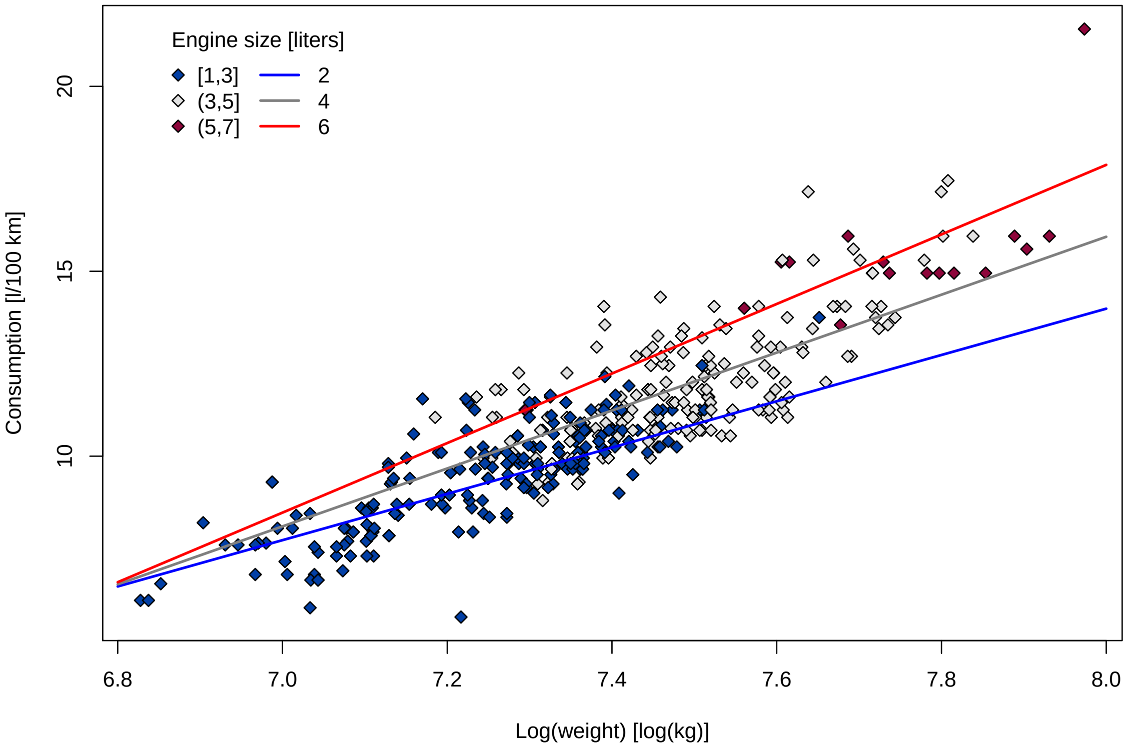
`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `engine.size + log(weight) + engine.size:log(weight),` $\widehat{m}(z, w) = -25.46 - 5.32\,z + 4.69\,\log(w) + 0.79\,z\,\log(w)$



`Cars2004nh` **(subset,** $n = 409$**)**

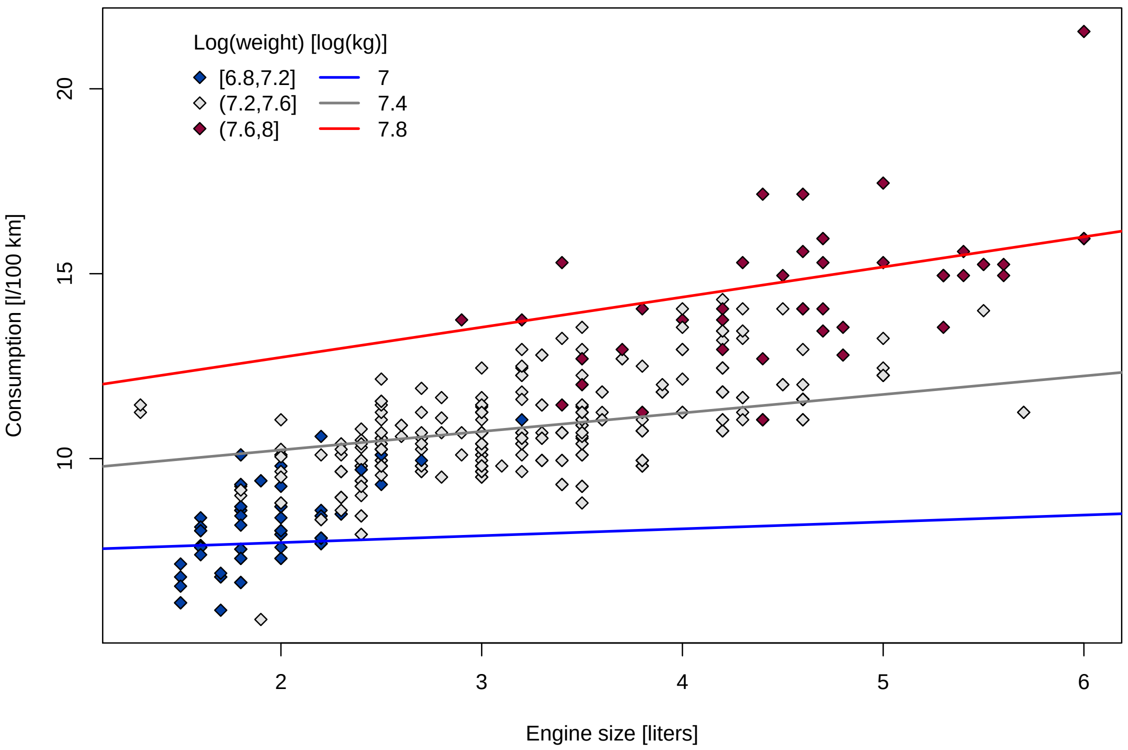
`consumption` $\sim$ `engine.size + log(weight) + engine.size:log(weight),` $\widehat{m}(z, w) = -25.46 - 5.32\,z + 4.69\,\log(w) + 0.79\,z\,\log(w)$



Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption ~ engine.size + log(weight) + engine.size:log(weight)`, $\widehat{m}(z, w) = -25.46 - 5.32\,z + 4.69\,\log(w) + 0.79\,z\,\log(w)$



`Cars2004nh` **(subset,** $n = 409$**)**

`consumption ~ engine.size + log(weight) + engine.size:log(weight)`

Y**: consumption** ***[l/100 km]***, Z**: engine size** ***[l]***, W**: weight** ***[kg]***

$$m(z, w) = \beta_0 + \beta^Z\,z + \beta^W\,\log(w) + \beta^{ZW}\,z\,\log(w)$$

```
lm(consumption ~ engine.size + lweight + engine.size:lweight, data = CarsNow)
```

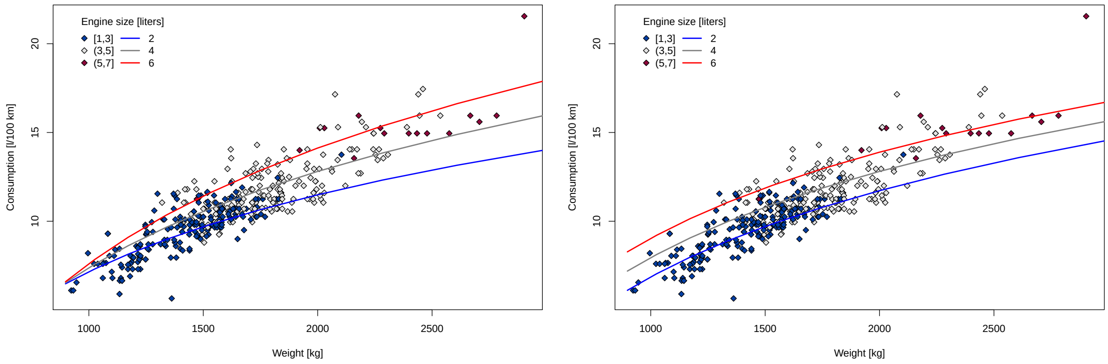
```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3999 -0.6538 -0.1407  0.4779  3.9219

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -25.4574     5.1267  -4.966 1.01e-06 ***
engine.size          -5.3160     1.4338  -3.708 0.000238 ***
lweight               4.6877     0.7104   6.599 1.30e-10 ***
engine.size:lweight   0.7860     0.1921   4.092 5.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9669 on 405 degrees of freedom
Multiple R-squared:  0.7961,        Adjusted R-squared:  0.7946
F-statistic: 527.2 on 3 and 405 DF,  p-value: < 2.2e-16
```

5.3.4 Additivity or interactions?

Illustrations

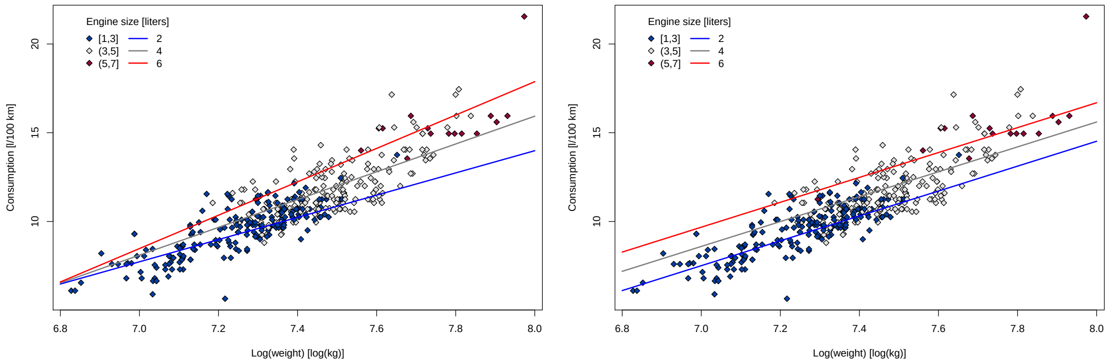
`Cars2004nh` **(subset,** $n = 409$)

`consumption` $\sim$ `engine.size, log(weight)`, **additivity or interactions?**



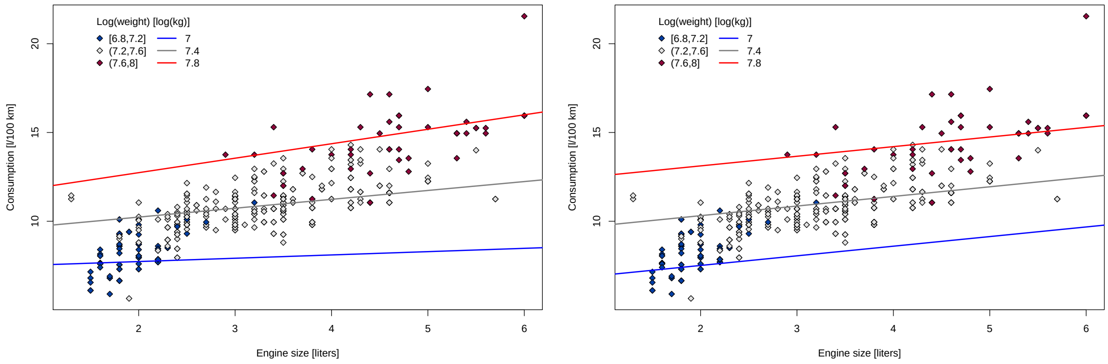
`Cars2004nh` **(subset,** $n = 409$)

`consumption` $\sim$ `engine.size, log(weight)`, **additivity or interactions?**



`Cars2004nh` **(subset,** $n = 409$)

`consumption` $\sim$ `engine.size, log(weight)`, **additivity or interactions?**



Illustrations

`Cars2004nh` (subset, $n = 409$)

`consumption` ~ `engine.size + log(weight) + engine.size:log(weight)`

Y: consumption *[l/100 km]*, Z: engine size *[l]*, W: weight *[kg]*

$$m(z,\,w) = \beta_0 + \beta^Z\,z + \beta^W\,\log(w) + \beta^{ZW}\,z\,\log(w)$$

Does the [log]`weight` have different effect on the mean consumption depending on the `engine size`?

Does the `engine size` have different effect on the mean consumption depending on the [log]`weight`?

```
lm(consumption ~ engine.size + lweight + engine.size:lweight, data = CarsNow)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3999 -0.6538 -0.1407  0.4779  3.9219

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -25.4574     5.1267  -4.966 1.01e-06 ***
engine.size          -5.3160     1.4338  -3.708 0.000238 ***
lweight               4.6877     0.7104   6.599 1.30e-10 ***
engine.size:lweight   0.7860     0.1921   4.092 5.15e-05 ***

...
```

`Cars2004nh` (subset, $n = 409$)

`consumption` ~ `engine.size + log(weight) + engine.size:log(weight)`

Y: consumption *[l/100 km]*, Z: engine size *[l]*, W: weight *[kg]*

$$m(z,\,w) = \beta_0 + \beta^Z\,z + \beta^W\,\log(w) + \beta^{ZW}\,z\,\log(w)$$

Does the [log]`weight` have different effect on the mean consumption depending on the `engine size`?

Does the `engine size` have different effect on the mean consumption depending on the [log]`weight`?

```
mAddit  <- lm(consumption ~ engine.size + lweight, data = CarsNow)
mInter  <- lm(consumption ~ engine.size*lweight, data = CarsNow)
anova(mAddit, mInter)
```

```
Analysis of Variance Table

Model 1: consumption ~ engine.size + lweight
Model 2: consumption ~ engine.size * lweight
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    406 394.26
2    405 378.60  1    15.656 16.748 5.154e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.3.5 More complex parameterization of either covariate

5.4 Two categorical covariates

This section is not complete in the notes.

Illustrations

`HowelsAll` **(subset,** $n = 289$**)**
Covariates: gender ($G = 2$) **and population** ($H = 3$)

```
data(HowellsAll, package = "mffSM")
```

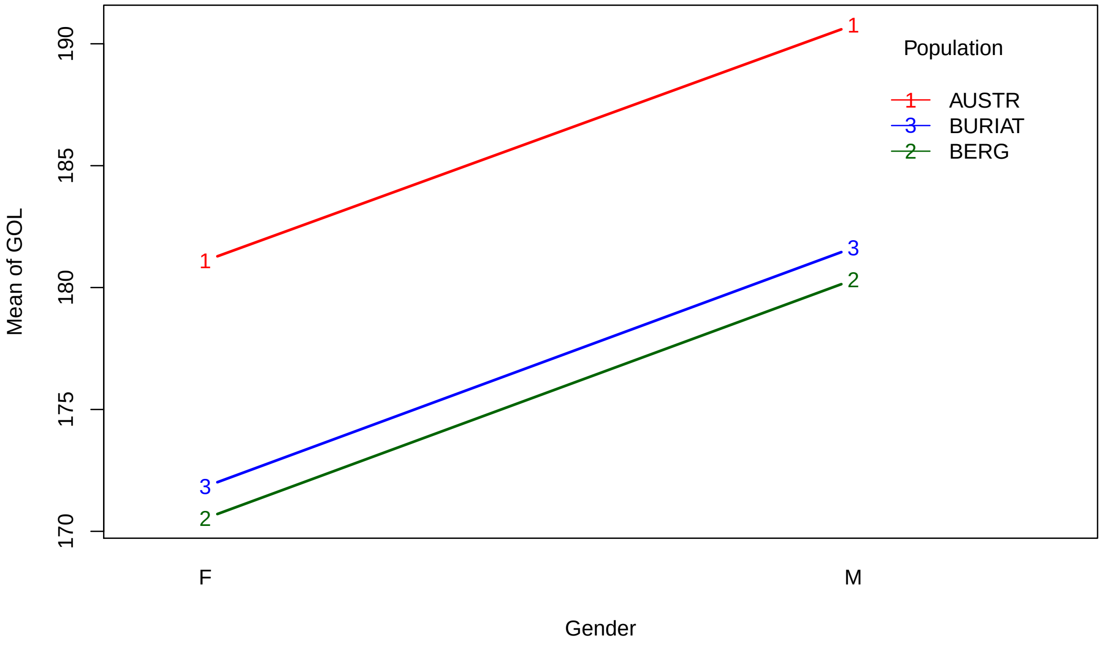
```
    gender popul oca gol fgender fpopul fgen.pop fpop.gen
1        1     1 123 176       M   BERG   M:BERG   BERG:M
2        1     1 115 173       M   BERG   M:BERG   BERG:M
3        1     1 117 176       M   BERG   M:BERG   BERG:M
4        1     1 113 185       M   BERG   M:BERG   BERG:M
...
57       0     1 125 171       F   BERG   F:BERG   BERG:F
58       0     1 103 178       F   BERG   F:BERG   BERG:F
59       0     1 115 165       F   BERG   F:BERG   BERG:F
60       0     1 117 169       F   BERG   F:BERG   BERG:F
...
110      1     0 109 194       M  AUSTR  M:AUSTR  AUSTR:M
112      1     0 115 188       M  AUSTR  M:AUSTR  AUSTR:M
116      1     0 115 187       M  AUSTR  M:AUSTR  AUSTR:M
117      1     0 109 196       M  AUSTR  M:AUSTR  AUSTR:M
...
192      0     0 109 186       F  AUSTR  F:AUSTR  AUSTR:F
193      0     0 115 175       F  AUSTR  F:AUSTR  AUSTR:F
194      0     0 111 185       F  AUSTR  F:AUSTR  AUSTR:F
195      0     0 113 184       F  AUSTR  F:AUSTR  AUSTR:F
...
241      1     2 118 180       M BURIAT M:BURIAT BURIAT:M
242      1     2 124 180       M BURIAT M:BURIAT BURIAT:M
243      1     2 117 183       M BURIAT M:BURIAT BURIAT:M
244      1     2 116 174       M BURIAT M:BURIAT BURIAT:M
...
295      0     2 116 175       F BURIAT F:BURIAT BURIAT:F
296      0     2 122 174       F BURIAT F:BURIAT BURIAT:F
297      0     2 113 174       F BURIAT F:BURIAT BURIAT:F
298      0     2 123 168       F BURIAT F:BURIAT BURIAT:F
...
```

5.4.1 Additivity

Illustrations

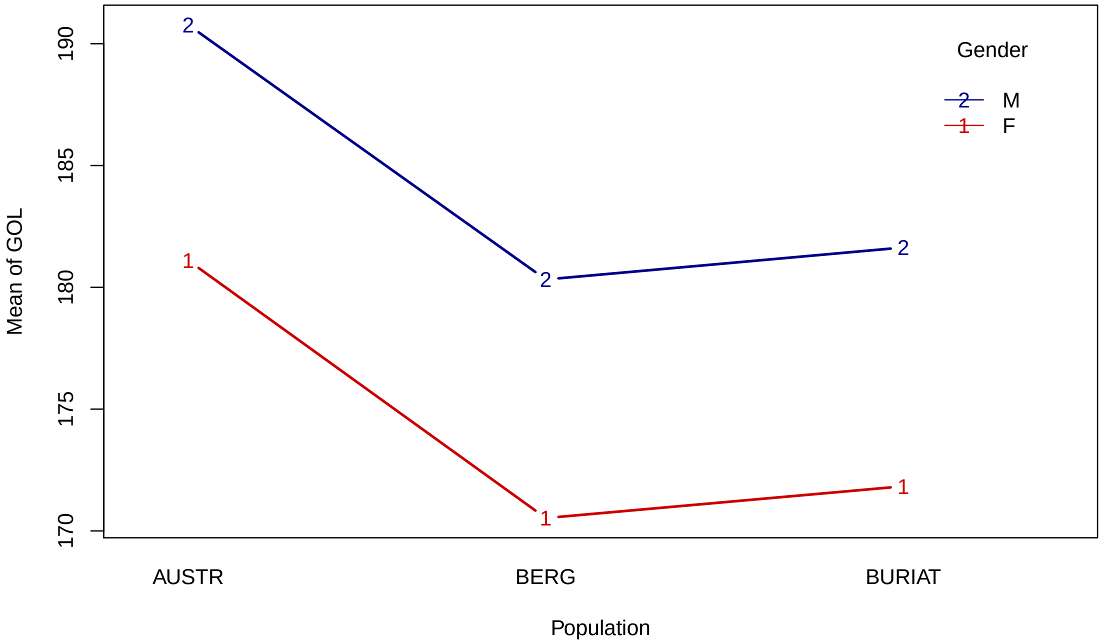
`HowellsAll` **($n = 289$)**

`gol` **(glabell-occipital length) $\sim$ gender ($G = 2$) and population ($H = 3$)**



`HowellsAll` **($n = 289$)**

`gol` **(glabell-occipital length) $\sim$ gender ($G = 2$) and population ($H = 3$)**



Illustrations

`HowelsAll` (subset, $n = 289$)

`gol` $\sim$ `gender + popul`, `contr.treatment` **parameterisation**

***Z*: gender (*Female, Male*), *W*: population (*Australia, Berg, Burjati*)**

$$m(z,\,w) = \beta_0 + \beta^Z\,\mathbb{I}[z = \text{male}] + \beta^W_1\,\mathbb{I}[w = \text{Berg}] + \beta^W_2\,\mathbb{I}[w = \text{Burjati}]$$

```
lm(gol ~ fgender + fpopul, data = HowellsAll)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-15.5400  -4.3103  -0.3103   4.4600  17.6897 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  181.0712     0.7814 231.724   <2e-16 ***
fgenderM       9.7703     0.7529  12.977   <2e-16 ***
fpopulBERG   -10.5311     0.9706 -10.850   <2e-16 ***
fpopulBURIAT  -9.2213     0.9695  -9.511   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.284 on 285 degrees of freedom
Multiple R-squared:  0.4729,	Adjusted R-squared:  0.4674 
F-statistic: 85.24 on 3 and 285 DF,  p-value: < 2.2e-16
```

`HowelsAll` (subset, $n = 289$)

`gol` $\sim$ `gender + popul`, `contr.sum` **parameterisation**

***Z*: gender (*Female, Male*), *W*: population (*Australia, Berg, Burjati*)**

$$\begin{aligned} m(z,\,w) &= \beta_0 + \beta^Z\,\mathbb{I}[z = \text{female}] - \beta^Z\,\mathbb{I}[z = \text{male}] \\ &\quad + \beta^W_1\,\mathbb{I}[w = \text{Austr}] + \beta^W_2\,\mathbb{I}[w = \text{Berg}] + (-\beta^W_1 - \beta^W_2)\,\mathbb{I}[w = \text{Burjati}] \end{aligned}$$

```
options(contrasts = c("contr.sum", "contr.sum"))
lm(gol ~ fgender + fpopul, data = HowellsAll)
```

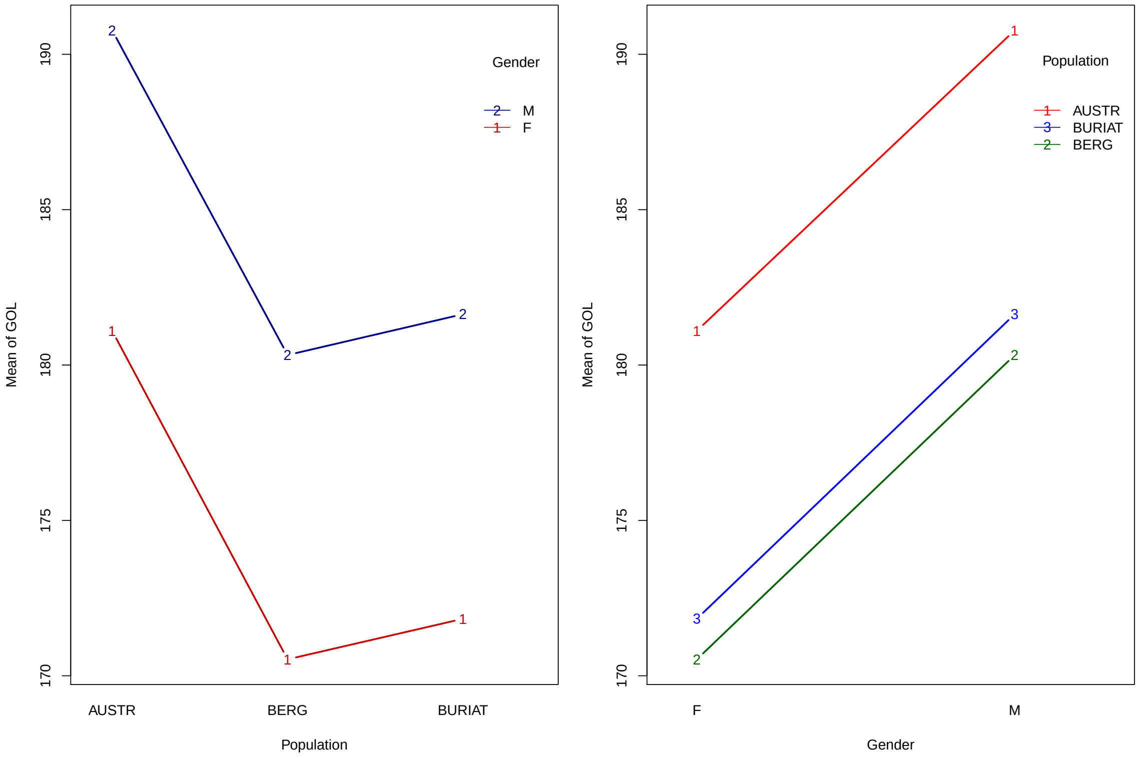
```
Residuals:
     Min       1Q   Median       3Q      Max 
-15.5400  -4.3103  -0.3103   4.4600  17.6897 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 179.3722     0.3797 472.421  < 2e-16 ***
fgender1     -4.8852     0.3765 -12.977  < 2e-16 ***
fpopul1       6.5842     0.5811  11.330  < 2e-16 ***
fpopul2      -3.9470     0.5157  -7.654 3.03e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.284 on 285 degrees of freedom
Multiple R-squared:  0.4729,	Adjusted R-squared:  0.4674 
F-statistic: 85.24 on 3 and 285 DF,  p-value: < 2.2e-16
```

5.4.2 Partial effects

Note that the following tests and estimated effects make only sense if it can be assumed that the *additivity* model holds.

Illustrations

`HowellsAll` **($n = 289$)**
`gol` **(glabell-occipital length) $\sim$ gender ($G = 2$) and population ($H = 3$),**
partial effect of gender, of population?



Illustrations

HowelsAll **(subset,** $n = 289$**)**

gol $\sim$ gender + popul

For a given *population****,***

does *gender* ***have an effect in the mean value of*** *gol****?***

Partial effect of *gender*

```
mgolAddit  <- lm(gol ~ fgender + fpopul, data = HowellsAll)
mgolPopul  <- lm(gol ~ fpopul, data = HowellsAll)
anova(mgolPopul, mgolAddit)
```

```
Analysis of Variance Table

Model 1: gol ~ fpopul
Model 2: gol ~ fgender + fpopul
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    286 17904
2    285 11254  1    6649.7 168.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HowelsAll **(subset,** $n = 289$**)**

gol $\sim$ gender + popul

For a given *gender****,***

does *population* ***have an effect in the mean value of*** *gol****?***

Partial effect of *population*

```
mgolAddit  <- lm(gol ~ fgender + fpopul, data = HowellsAll)
mgolGender  <- lm(gol ~ fgender, data = HowellsAll)
anova(mgolGender, mgolAddit)
```

```
Analysis of Variance Table

Model 1: gol ~ fgender
Model 2: gol ~ fgender + fpopul
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    287 16415
2    285 11254  2    5160.7 65.345 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HowelsAll **(subset,** $n = 289$**)**

gol $\sim$ gender + popul

F-tests of significance of both partial effects

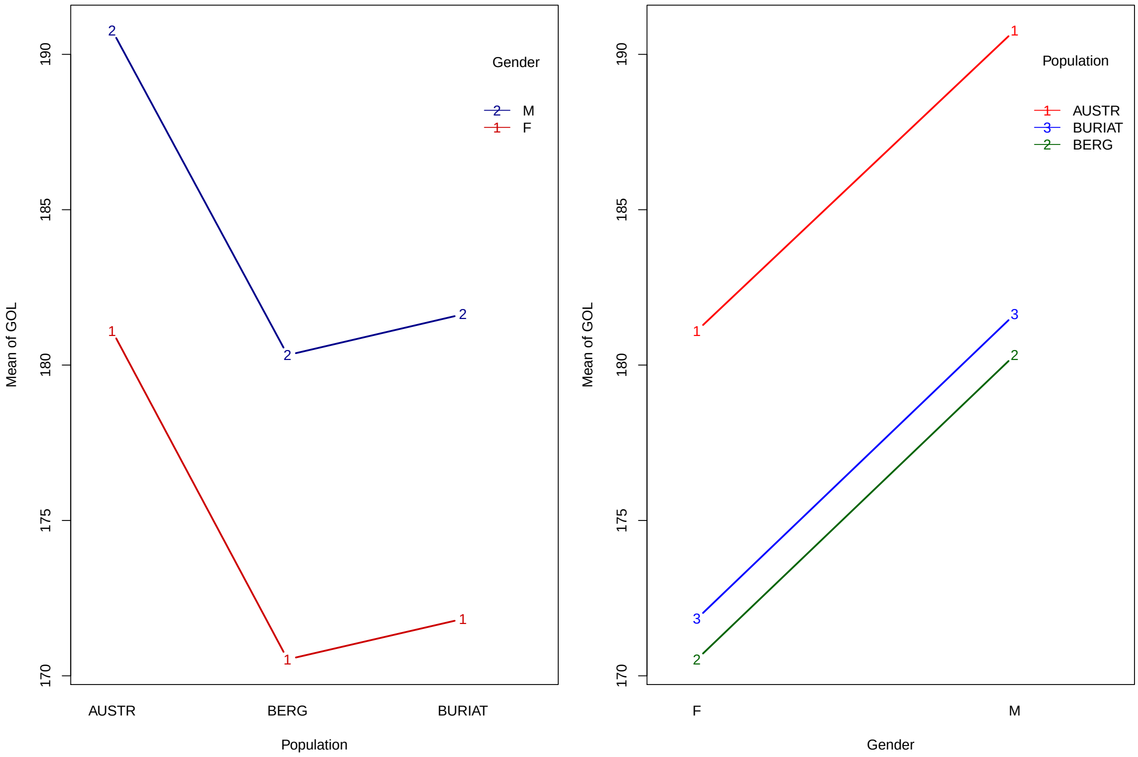
```
mgolAddit  <- lm(gol ~ fgender + fpopul, data = HowellsAll)
drop1(mgolAddit, test = "F")
```

```
Single term deletions

Model:
gol ~ fgender + fpopul
        Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>                11254 1066.3
fgender  1    6649.7 17904 1198.5 168.396 < 2.2e-16 ***
fpopul   2    5160.7 16415 1171.4  65.345 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Illustrations

`HowellsAll` ($n = 289$)

`gol` **(glabell-occipital length)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$), **quantification of both partial effects?**



`HowelsAll` **(subset,** $n = 289$**)**

`gol ~ gender + popul`, `contr.treatment` **parameterisation**

Z**: gender (*Female, Male*),** W**: population (*Australia, Berg, Burjati*)**

$$m(z,\,w) = \beta_0 + \beta^Z\,\mathbb{I}[z = \text{male}] + \beta_1^W\,\mathbb{I}[w = \text{Berg}] + \beta_2^W\,\mathbb{I}[w = \text{Burjati}]$$

```
lm(gol ~ fgender + fpopul, data = HowellsAll)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.5400  -4.3103  -0.3103   4.4600  17.6897

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.0712     0.7814 231.724   <2e-16 ***
fgenderM       9.7703     0.7529  12.977   <2e-16 ***
fpopulBERG   -10.5311     0.9706 -10.850   <2e-16 ***
fpopulBURIAT  -9.2213     0.9695  -9.511   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.284 on 285 degrees of freedom
Multiple R-squared:  0.4729,       Adjusted R-squared:  0.4674
F-statistic: 85.24 on 3 and 285 DF,  p-value: < 2.2e-16
```

Illustrations

`HowelsAll` **(subset,** $n = 289$**)**

`gol` $\sim$ `gender + popul`

LSE's of $\mathbb{E}\big(Y \,\big|\, Z = g_1, W = \star\big) - \mathbb{E}\big(Y \,\big|\, Z = g_2, W = \star\big)$
and $\mathbb{E}\big(Y \,\big|\, Z = \star, W = h_1\big) - \mathbb{E}\big(Y \,\big|\, Z = \star, W = h_2\big)$

```
mgolAddit  <- lm(gol ~ fgender + fpopul, data = HowellsAll)
L  <- matrix(c(0,1,0,0, 0,0,1,0, 0,0,0,1, 0,0,-1,1), ncol = 4, byrow = TRUE)
rownames(L)  <- c("Male-Female", "Berg-Austr", "Burjati-Austr", "Burjati-Berg")
colnames(L)  <- names(coef(mgolAddit))
print(L)
```

```
              (Intercept) fgenderM fpopulBERG fpopulBURIAT
Male-Female             0        1          0            0
Berg-Austr              0        0          1            0
Burjati-Austr           0        0          0            1
Burjati-Berg            0        0         -1            1
```

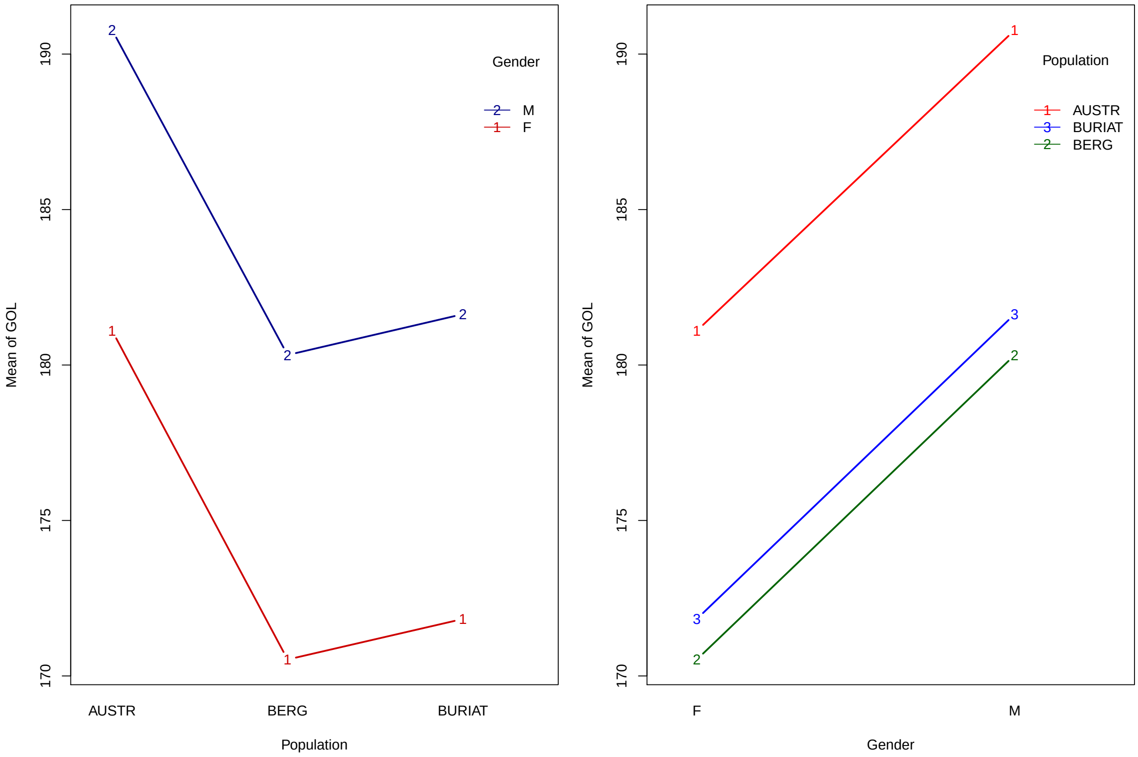
```
mffSM::LSest(mgolAddit, L = L)
```

```
                Estimate Std. Error    t value P value       Lower     Upper
Male-Female     9.770313  0.7529092  12.976750 < 2e-16   8.2883454 11.252282
Berg-Austr    -10.531148  0.9705782 -10.850385 < 2e-16 -12.4415591 -8.620737
Burjati-Austr  -9.221329  0.9695097  -9.511332 < 2e-16 -11.1296364 -7.313021
Burjati-Berg    1.309819  0.8512377   1.538723 0.12498  -0.3656911  2.985330
```

Illustrations

`HowellsAll` ($n = 289$)

`gol` (glabell-occipital length) $\sim$ gender ($G = 2$) and population ($H = 3$), alternative quantification of both partial effects?



`HowelsAll` **(subset,** $n = 289$**)**

`gol ~ gender + popul`, `contr.sum` **parameterisation**

Z: **gender (*Female, Male*),** W: **population (*Australia, Berg, Burjati*)**

$$m(z,\,w) = \beta_0 + \beta^Z\,\mathbb{I}[z = \text{female}] - \beta^Z\,\mathbb{I}[z = \text{male}] + \beta^W_1\,\mathbb{I}[w = \text{Austr}] + \beta^W_2\,\mathbb{I}[w = \text{Berg}] + (-\beta^W_1 - \beta^W_2)\,\mathbb{I}[w = \text{Burjati}]$$

```
options(contrasts = c("contr.sum", "contr.sum"))
lm(gol ~ fgender + fpopul, data = HowellsAll)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.5400  -4.3103  -0.3103   4.4600  17.6897

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 179.3722     0.3797 472.421  < 2e-16 ***
fgender1     -4.8852     0.3765 -12.977  < 2e-16 ***
fpopul1       6.5842     0.5811  11.330  < 2e-16 ***
fpopul2      -3.9470     0.5157  -7.654 3.03e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.284 on 285 degrees of freedom
Multiple R-squared:  0.4729,       Adjusted R-squared:  0.4674
F-statistic: 85.24 on 3 and 285 DF,  p-value: < 2.2e-16
```

Illustrations

`HowelsAll` **(subset,** $n = 289$**)**

`gol` $\sim$ `gender + popul`

LSE's of $\mathbb{E}\bigl(Y \,\big|\, Z = g, W = \star\bigr) - \frac{1}{G}\sum_{j=1}^{G}\mathbb{E}\bigl(Y \,\big|\, Z = j, W = \star\bigr)$

and $\mathbb{E}\bigl(Y \,\big|\, Z = \star, W = h\bigr) - \frac{1}{H}\sum_{j=1}^{H}\mathbb{E}\bigl(Y \,\big|\, Z = \star, W = j\bigr)$

```
options(contrasts = c("contr.sum", "contr.sum"))
mgolAdditSum  <- lm(gol ~ fgender + fpopul, data = HowellsAll)
L  <- matrix(c(0,1,0,0, 0,-1,0,0, 0,0,1,0, 0,0,0,1, 0,0,-1,-1), ncol = 4, byrow = TRUE)
rownames(L)  <- c("Female", "Male", "Australia", "Berg", "Burjati")
colnames(L)  <- names(coef(mgolAdditSum))
print(L)
```

```
          (Intercept) fgender1 fpopul1 fpopul2
Female              0        1       0       0
Male                0       -1       0       0
Australia           0        0       1       0
Berg                0        0       0       1
Burjati             0        0      -1      -1
```

```
mffSM::LSest(mgolAdditSum, L = L)
```

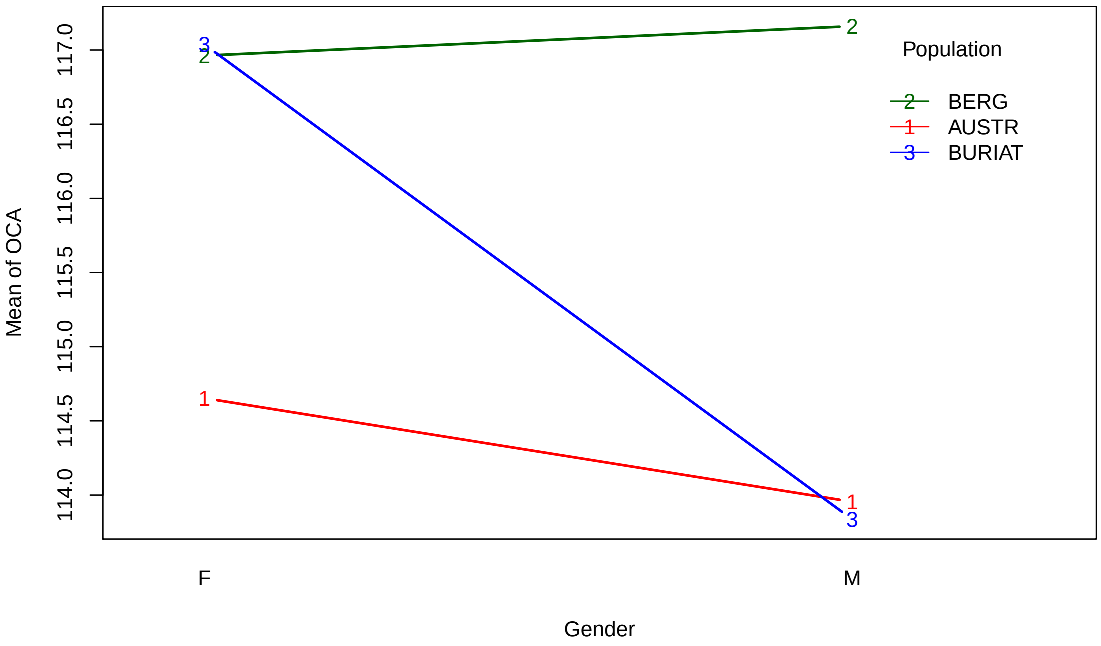
```
           Estimate Std. Error    t value    P value     Lower     Upper
Female    -4.885157  0.3764546 -12.976750 < 2.22e-16 -5.626141 -4.144173
Male       4.885157  0.3764546  12.976750 < 2.22e-16  4.144173  5.626141
Australia  6.584159  0.5811231  11.330059 < 2.22e-16  5.440321  7.727997
Berg      -3.946989  0.5156772  -7.653992 3.0336e-13 -4.962008 -2.931970
Burjati   -2.637170  0.5150067  -5.120651 5.6141e-07 -3.650869 -1.623470
```

5.4.3 Interactions

Illustrations

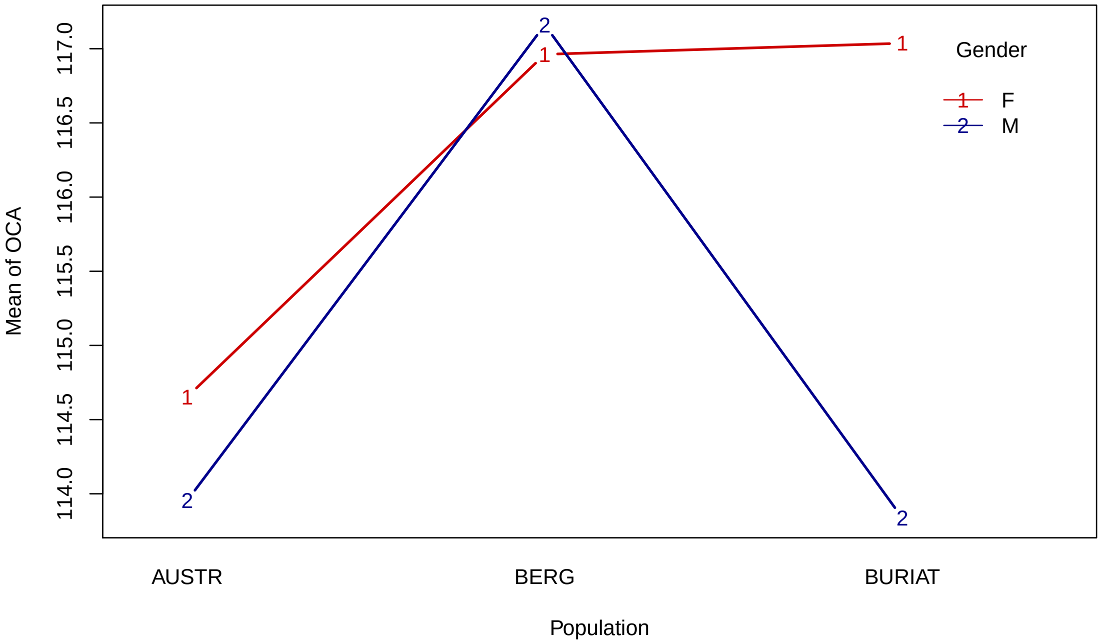
`HowellsAll` ($n = 289$)

`oca` **(occipital angle)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$)



`HowellsAll` ($n = 289$)

`oca` **(occipital angle)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$)



Illustrations

`HowelsAll` (subset, $n = 289$)

`oca` $\sim$ `gender + popul + gender:popul`, `contr.treatment` **parameterisation**

Z: **gender (*Female, Male*),** W: **population (*Australia, Berg, Burjati*)**

$$m(z,\,w) = \beta_0 + \beta^Z\,\mathbb{I}[z = \text{male}] + \beta^W_1\,\mathbb{I}[w = \text{Berg}] + \beta^W_2\,\mathbb{I}[w = \text{Burjati}]$$
$$+\,\beta^{ZW}_1\mathbb{I}[z = \text{male}, w = \text{Berg}] + \beta^{ZW}_2\mathbb{I}[z = \text{male}, w = \text{Burjati}]$$

```
lm(oca ~ fgender*fpopul, data = HowellsAll)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-15.1607  -3.1607   0.0455   3.1636  13.8393 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           114.6531     0.7186 159.548   <2e-16 ***
fgenderM               -0.6985     1.2910  -0.541   0.5889    
fpopulBERG              2.3092     0.9969   2.316   0.0213 *  
fpopulBURIAT            2.3840     0.9925   2.402   0.0169 *  
fgenderM:fpopulBERG     0.8970     1.6112   0.557   0.5782    
fgenderM:fpopulBURIAT  -2.5022     1.6110  -1.553   0.1215    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.03 on 283 degrees of freedom
Multiple R-squared:  0.07842,	Adjusted R-squared:  0.06214 
F-statistic: 4.816 on 5 and 283 DF,  p-value: 0.0003046
```

`HowelsAll` (subset, $n = 289$)

`oca` $\sim$ `gender + popul + gender:popul`, `contr.sum` **parameterisation**

Z: **gender (*Female, Male*),** W: **population (*Australia, Berg, Burjati*)**

$$\begin{aligned} m(z,\,w) \;=\; & \beta_0 + \beta^Z\,\mathbb{I}[z = \text{female}] - \beta^Z\,\mathbb{I}[z = \text{male}] \\ & + \beta^W_1\,\mathbb{I}[w = \text{Austr.}] + \beta^W_2\,\mathbb{I}[w = \text{Berg}] + \bigl(-\beta^W_1 - \beta^W_2\bigr)\,\mathbb{I}[w = \text{Burjati}] \\ & + \beta^{ZW}_1\mathbb{I}[z = \text{fem.}, w = \text{Aus.}] + \beta^{ZW}_2\mathbb{I}[z = \text{fem.}, w = \text{Berg}] + \bigl(-\beta^{ZW}_1 - \beta^{ZW}_2\bigr)\,\mathbb{I}[z = \text{fem.}, w = \text{Bur.}] \\ & - \beta^{ZW}_1\mathbb{I}[z = \text{male}, w = \text{Aus.}] - \beta^{ZW}_2\mathbb{I}[z = \text{male}, w = \text{Berg}] + \bigl(\beta^{ZW}_1 + \beta^{ZW}_2\bigr)\,\mathbb{I}[z = \text{male}, w = \text{Bur.}] \end{aligned}$$

```
options(contrasts = c("contr.sum", "contr.sum"))
lm(oca ~ fgender + fpopul, data = HowellsAll)
```

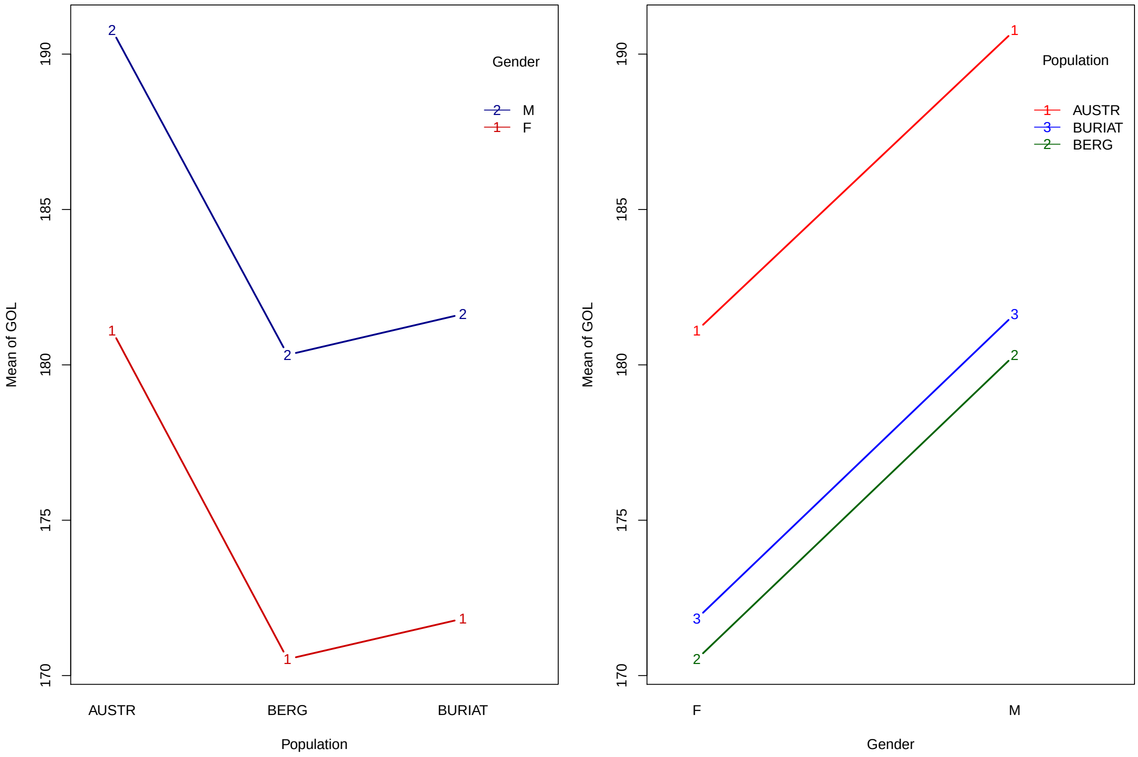
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      115.6007     0.3129 369.455  < 2e-16 ***
fgender1           0.6168     0.3129   1.971 0.049671 *  
fpopul1           -1.2969     0.4866  -2.665 0.008138 ** 
fpopul2            1.4608     0.4187   3.489 0.000563 ***
fgender1:fpopul1  -0.2675     0.4866  -0.550 0.582896    
fgender1:fpopul2  -0.7160     0.4187  -1.710 0.088376 .  
---

Residual standard error: 5.03 on 283 degrees of freedom
Multiple R-squared:  0.07842,	Adjusted R-squared:  0.06214 
F-statistic: 4.816 on 5 and 283 DF,  p-value: 0.0003046
```

5.4.4 Additivity or interactions?

Illustrations

HowellsAll ($n = 289$)

gol (glabell-occipital length) $\sim$ gender ($G = 2$) and population ($H = 3$), additivity or interactions?



HowelsAll **(subset,** $n = 289$**)**

gol (glabell-occipital length) $\sim$ gender ($G = 2$) and population ($H = 3$)

Do the mean gol differences between *male* and *female* depend on population?

Do the mean gol differences between populations depend on gender?

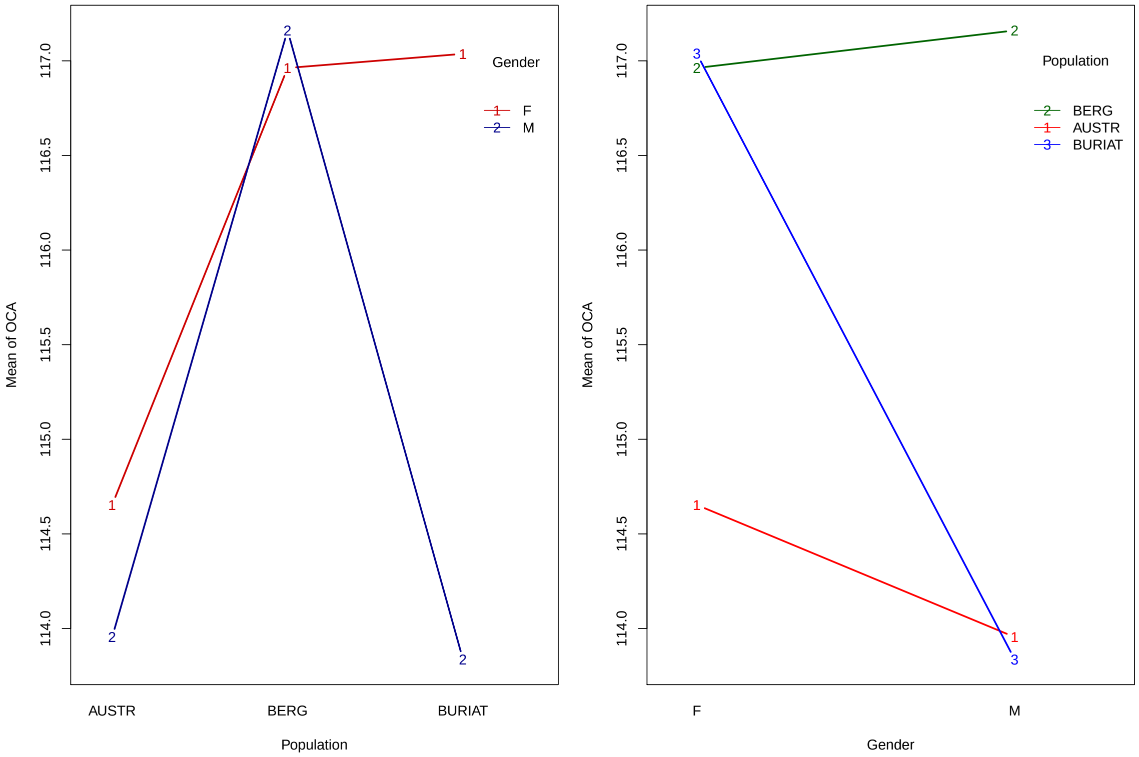
```
mgolAddit  <- lm(gol ~ fgender + fpopul, data = HowellsAll)
mgolInter  <- lm(gol ~ fgender*fpopul, data = HowellsAll)
anova(mgolAddit, mgolInter)
```

```
Analysis of Variance Table

Model 1: gol ~ fgender + fpopul
Model 2: gol ~ fgender * fpopul
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    285 11254
2    283 11254  2   0.19404 0.0024 0.9976
```

Illustrations

HowellsAll ($n = 289$)

oca (occipital angle) $\sim$ gender ($G = 2$) and population ($H = 3$), additivity or interactions?



HowelsAll (subset, $n = 289$)

oca **(occipital angle) $\sim$ gender ($G = 2$) and population ($H = 3$)**

Do the mean oca **differences between** *male* **and** *female* **depend on** population**?**

Do the mean oca **differences between** populations **depend on** gender**?**

```
mocaAddit  <- lm(oca ~ fgender + fpopul, data = HowellsAll)
mocaInter  <- lm(oca ~ fgender*fpopul, data = HowellsAll)
anova(mocaAddit, mocaInter)
```

```
Analysis of Variance Table

Model 1: oca ~ fgender + fpopul
Model 2: oca ~ fgender * fpopul
  Res.Df  RSS Df Sum of Sq      F  Pr(>F)
1    285 7326
2    283 7161  2    165.02 3.2607 0.03981 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.5 Multiple regression model

5.5.1 Model terms

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 4.3 and 4.4 or inside an interaction (see Defition 5.2) 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 4.3 and 4.4 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 $\boldsymbol{X}_1 = \boldsymbol{s}(Z_1)$, ..., $\boldsymbol{X}_n = \boldsymbol{s}(Z_n)$ 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

$$\mathbb{S} = \begin{pmatrix} s(Z_1) \\ \vdots \\ s(Z_n) \end{pmatrix} = \big(\boldsymbol{S}\big), \qquad \begin{aligned} \boldsymbol{X}_1 = X_1 &= s(Z_1), \\ &\vdots \\ \boldsymbol{X}_n = X_n &= s(Z_n). \end{aligned} \tag{5.2}$$

(ii) **Polynomial**: $\boldsymbol{s} = \big(s_1, \ldots, s_{k-1}\big)^\top$ such that $s_j(z) = P^j(z)$ is polynomial in z of degree j, $j = 1, \ldots, k-1$. This leads to

$$\mathbb{S} = \begin{pmatrix} P^1(Z_1) & \ldots & P^{k-1}(Z_1) \\ \vdots & \vdots & \vdots \\ P^1(Z_n) & \ldots & P^{k-1}(Z_n) \end{pmatrix} = \Big(\boldsymbol{P}^1, \quad \ldots, \quad \boldsymbol{P}^{k-1}\Big), \tag{5.3}$$

$$\begin{aligned} \boldsymbol{X}_1 &= \big(P^1(Z_1), \ldots, P^{k-1}(Z_1)\big)^\top, \\ &\vdots \\ \boldsymbol{X}_n &= \big(P^1(Z_n), \ldots, P^{k-1}(Z_n)\big)^\top. \end{aligned}$$

For a particular form of the basis polynomials $P^1, \ldots, P^{k-1}$, raw or orthonormal polynomials have been suggested in Sections 4.3.2 and 4.3.3. Other choices are possible as well.

(iii) **Regression spline**: $\boldsymbol{s} = \big(s_1, \ldots, s_k\big)^\top$ 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 $\boldsymbol{\lambda} = \big(\lambda_1, \ldots, \lambda_{k-d+1}\big)^\top$. This leads to

$$\mathbb{S} = \mathbb{B} = \begin{pmatrix} B_1(Z_1) & \ldots & B_k(Z_1) \\ \vdots & \vdots & \vdots \\ B_1(Z_n) & \ldots & B_k(Z_n) \end{pmatrix} = \Big(\boldsymbol{B}^1, \quad \ldots, \quad \boldsymbol{B}^k\Big), \tag{5.4}$$

$$\begin{aligned} \boldsymbol{X}_1 &= \big(B_1(Z_1), \ldots, B_k(Z_1)\big)^\top, \\ &\vdots \\ \boldsymbol{X}_n &= \big(B_1(Z_n), \ldots, B^k(Z_n)\big)^\top. \end{aligned}$$

Categorical covariate with $\mathcal{Z} = \big\{1, \ldots, G\big\}$. 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

$$\mathbb{S} = \begin{pmatrix} \boldsymbol{c}_{Z_1}^\top \\ \vdots \\ \boldsymbol{c}_{Z_n}^\top \end{pmatrix} = \Big(\boldsymbol{C}^1, \quad \ldots, \quad \boldsymbol{C}^{G-1}\Big), \qquad \begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{c}_{Z_1}, \\ &\vdots \\ \boldsymbol{X}_n &= \boldsymbol{c}_{Z_n}. \end{aligned} \tag{5.5}$$

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 elements of a chosen parameterization $\boldsymbol{s}$ and 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 5.3 The main effect model term.

Depending on a chosen parameterization $\boldsymbol{s}$, the main effect model term[1] *(of order one) of a given covariate Z is defined as a transformation $\boldsymbol{t}$ with elements as follows and a matrix $\mathbb{T}$ with columns as follows:*

Numeric covariate

(i) ***Simple transformation*** *with $s: \mathcal{Z} \longrightarrow \mathbb{R}$.*
$\boldsymbol{t} = s$ and $\mathbb{T}$ is (the only) column $\boldsymbol{S}$ of the reparameterizing matrix $\mathbb{S}$ given by (5.2), i.e.,

$$\mathbb{T} = \big(\boldsymbol{S}\big).$$

(ii) ***Polynomial*** *with $\boldsymbol{s} = \big(s_1, \ldots, s_{k-1}\big)^\top$, $s_j(z) = P^j(z)$, $j = 1, \ldots, k-1$.*
$\boldsymbol{t} = s_1 = P^1$ (linear polynomial) and $\mathbb{T}$ is the first column $\boldsymbol{P}^1$ of the reparameterizing matrix $\mathbb{S}$ (given by Eq. 5.3) that corresponds to the linear transformation of the covariate Z, i.e.,

$$\mathbb{T} = \big(\boldsymbol{P}^1\big).$$

(iii) ***Regression spline*** *with $\boldsymbol{s} = \big(s_1, \ldots, s_k\big)^\top$, $s_j(z) = B_j(z)$, $j = 1, \ldots, k$.*
$\boldsymbol{t} = \boldsymbol{s}$ (all basis splines) and the matrix $\mathbb{T}$ contains (all) columns $\boldsymbol{B}^1, \ldots, \boldsymbol{B}^k$ of the reparameterizing matrix $\mathbb{S} = \mathbb{B}$ given by (5.4), i.e.,

$$\mathbb{T} = \big(\boldsymbol{B}^1, \ldots, \boldsymbol{B}^k\big).$$

Categorical covariate *with $\boldsymbol{s}(z) = \boldsymbol{c}(z)$.*
$\boldsymbol{t} = \boldsymbol{s}$ (row of a chosen (pseudo)contrast matrix) and the matrix $\mathbb{T}$ contains (all) columns $\boldsymbol{C}^1, \ldots, \boldsymbol{C}^{G-1}$ of the reparameterizing matrix $\mathbb{S}$ given by (5.5), i.e.,

$$\mathbb{T} = \big(\boldsymbol{C}^1, \ldots, \boldsymbol{C}^{G-1}\big).$$

Definition 5.4 The main effect model term of order j.

If a numeric *covariate Z is parameterized using the polynomial of degree $k-1$, i.e., $\boldsymbol{s} = \big(s_1, \ldots, s_{k-1}\big)^\top$, $s_j(z) = P^j(z)$, $j = 1, \ldots, k-1$, then* the main effect model term of order j, $j = 2, \ldots, k-1$, *means the element $s_j(z) = P^j(z)$ of the polynomial parameterization and a matrix $\mathbb{T}^j$ whose the only column is the jth column $\boldsymbol{P}^j$ of the reparameterizing matrix $\mathbb{S}$ (given by Eq. 5.3) that corresponds to the polynomial of degree j, i.e.,*

$$\mathbb{T}^j = \big(\boldsymbol{P}^j\big).$$

Note. The terms $\mathbb{T}, \ldots, \mathbb{T}^{j-1}$ are called as *lower order* terms included in the term $\mathbb{T}^j$.

[1] *hlavní efekt*

Two-way interaction model terms

In the following, consider two covariates Z and W and their main effect model terms $\boldsymbol{t}_Z$, $\mathbb{T}_Z$ and $\boldsymbol{t}_W$, $\mathbb{T}_W$.

Definition 5.5 The two-way interaction model term.

The two-way interaction[2] *model term means elements of a vector* $\boldsymbol{t}_W \otimes \boldsymbol{t}_Z$ *and 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 enters 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 $\boldsymbol{t}_Z$, $\mathbb{T}_Z$ and $\boldsymbol{t}_W$, $\mathbb{T}_W$ and $\boldsymbol{t}_V$, $\mathbb{T}_V$.

Definition 5.6 The three-way interaction model term.

The three-way interaction[3] *model term means a vector* $\boldsymbol{t}_V \otimes (\boldsymbol{t}_W \otimes \boldsymbol{t}_Z)$ *and a matrix* $\mathbb{T}^{ZWV}$*, where*

$$\mathbb{T}^{ZWV} := \bigl(\mathbb{T}_Z : \mathbb{T}_W\bigr) : \mathbb{T}_V.$$

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.

[2] *dvojná interakce* [3] *trojná interakce*

5.5.2 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 **−1** among the model terms.

5.5.3 Hierarchically well formulated model

Definition 5.7 Hierarchically well formulated model.

Hierarchically well formulated (HWF) model[4] *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 5.1.

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:

$$x = \delta\,(t - \varphi), \qquad t = \varphi + \frac{x}{\delta}, \tag{5.6}$$

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 (5.6) 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 ≡ 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 $m(x) = \beta_0 + \beta_2\,x^2$ ≡ 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 $m(x,\,z) = \beta_0 + \beta_1\,z + \beta_2\,x\,z$ ≡ it can be assumed that with $z = 0$, the response expectation does not depend on a value of x, i.e., $\mathbb{E}\bigl(Y \bigm| X = x,\ Z = 0\bigr) = \beta_0$ (a constant).

[4] *hierarchicky dobře formulovaný model*

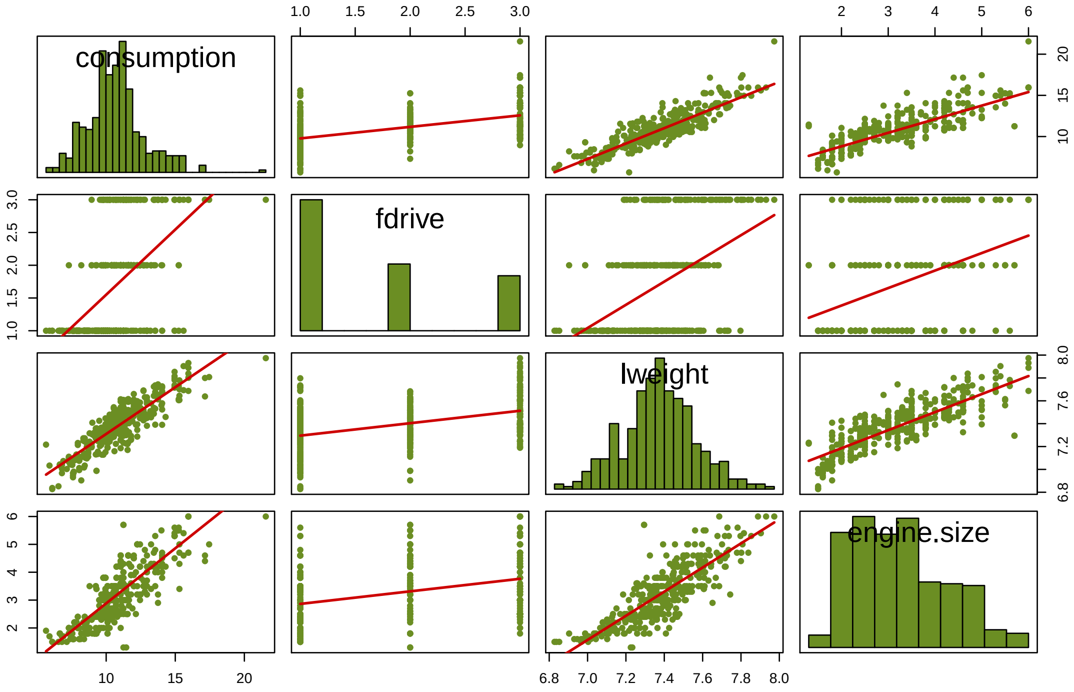
5.5.4 Usual strategy to specify a multiple regression model

This section is not complete in the notes.

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `drive, engine size, log(weight)`



Illustrations

Cars2004nh (subset, $n = 409$)

consumption ∼ drive + engine size + log(weight)

```
mAddit <- lm(consumption ~ fdrive + engine.size + lweight, data = CarsNow)
summary(mAddit)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -35.84930    3.08092 -11.636  < 2e-16 ***
fdriverear    0.46260    0.11715   3.949 9.26e-05 ***
fdrive4x4     0.98198    0.13019   7.543 3.07e-13 ***
engine.size   0.56908    0.08361   6.807 3.62e-11 ***
lweight       6.03099    0.44795  13.464  < 2e-16 ***

Residual standard error: 0.9223 on 404 degrees of freedom
Multiple R-squared:  0.8149,	Adjusted R-squared:  0.8131
F-statistic: 444.8 on 4 and 404 DF,  p-value: < 2.2e-16
```

```
drop1(mAddit, test = "F")
```

```
Single term deletions

Model:
consumption ~ fdrive + engine.size + lweight
            Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                   343.69 -61.161
fdrive       2    50.574 394.26  -9.012  29.725 9.046e-13 ***
engine.size  1    39.413 383.10 -18.758  46.330 3.625e-11 ***
lweight      1   154.205 497.89  88.436 181.267 < 2.2e-16 ***
```

Cars2004nh (subset, $n = 409$)

consumption ∼ drive + engine size + log(weight) + drive:log(weight)

```
mInter1 <- lm(consumption ~ fdrive + engine.size + lweight + fdrive:lweight, data = CarsNow)
summary(mInter1)
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -37.44459    3.22260 -11.619  < 2e-16 ***
fdriverear          22.90273    4.86163   4.711 3.40e-06 ***
fdrive4x4           -8.59853    4.42520  -1.943   0.0527 .
engine.size          0.57588    0.08125   7.088 6.16e-12 ***
lweight              6.24702    0.46296  13.494  < 2e-16 ***
fdriverear:lweight  -3.03731    0.65971  -4.604 5.57e-06 ***
fdrive4x4:lweight    1.26748    0.59358   2.135   0.0333 *

Residual standard error: 0.8877 on 402 degrees of freedom
Multiple R-squared:  0.8294,	Adjusted R-squared:  0.8269
F-statistic: 325.8 on 6 and 402 DF,  p-value: < 2.2e-16
```

```
drop1(mInter1, test = "F")
```

```
Single term deletions

Model:
consumption ~ fdrive + engine.size + lweight + fdrive:lweight
               Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                      316.81 -90.469
engine.size     1    39.590 356.40 -44.308  50.236 6.159e-12 ***
fdrive:lweight  2    26.879 343.69 -61.161  17.054 7.782e-08 ***
```

Illustrations

Cars2004nh (subset, $n = 409$)

consumption $\sim$ drive + engine size + log(weight) + drive:log(weight) + engine size:log(weight)

```
mInter2 <- lm(consumption ~ fdrive + engine.size + lweight + fdrive:lweight +
              engine.size:lweight, data = CarsNow)
summary(mInter2)
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -22.8398     4.9687  -4.597 5.76e-06 ***
fdriverear           27.3567     4.9219   5.558 4.98e-08 ***
fdrive4x4             4.3904     5.5249   0.795 0.427287
engine.size          -5.8845     1.6945  -3.473 0.000571 ***
lweight               4.2821     0.6873   6.230 1.18e-09 ***
fdriverear:lweight   -3.6356     0.6675  -5.446 8.98e-08 ***
fdrive4x4:lweight    -0.4836     0.7425  -0.651 0.515241
engine.size:lweight   0.8662     0.2270   3.817 0.000157 ***

Residual standard error: 0.8731 on 401 degrees of freedom
Multiple R-squared:  0.8354,        Adjusted R-squared:  0.8325
F-statistic: 290.7 on 7 and 401 DF,  p-value: < 2.2e-16
```

```
drop1(mInter2, test = "F")
```

```
consumption ~ fdrive + engine.size + lweight + fdrive:lweight +
    engine.size:lweight
                    Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                           305.70 -103.064
fdrive:lweight       2    24.150 329.85  -75.966  15.839 2.395e-07 ***
engine.size:lweight  1    11.105 316.81  -90.469  14.567 0.0001566 ***
```

Cars2004nh (subset, $n = 409$)

consumption $\sim$ (drive + engine size + log(weight))^2

```
mInter <- lm(consumption ~ (fdrive + engine.size + lweight)^2, data = CarsNow)
summary(mInter)
```

```
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -26.124609   5.776121  -4.523 8.06e-06 ***
fdriverear              26.875936   7.367167   3.648 0.000299 ***
fdrive4x4               13.308169   8.311915   1.601 0.110147
engine.size             -5.391862   1.746264  -3.088 0.002158 **
lweight                  4.757609   0.817131   5.822 1.19e-08 ***
fdriverear:engine.size   0.009665   0.182958   0.053 0.957895
fdrive4x4:engine.size    0.315489   0.216880   1.455 0.146547
fdriverear:lweight      -3.571144   1.061146  -3.365 0.000839 ***
fdrive4x4:lweight       -1.818723   1.189560  -1.529 0.127081
engine.size:lweight      0.790111   0.233312   3.386 0.000778 ***

Residual standard error: 0.8726 on 399 degrees of freedom
Multiple R-squared:  0.8364,        Adjusted R-squared:  0.8327
F-statistic: 226.7 on 9 and 399 DF,  p-value: < 2.2e-16
```

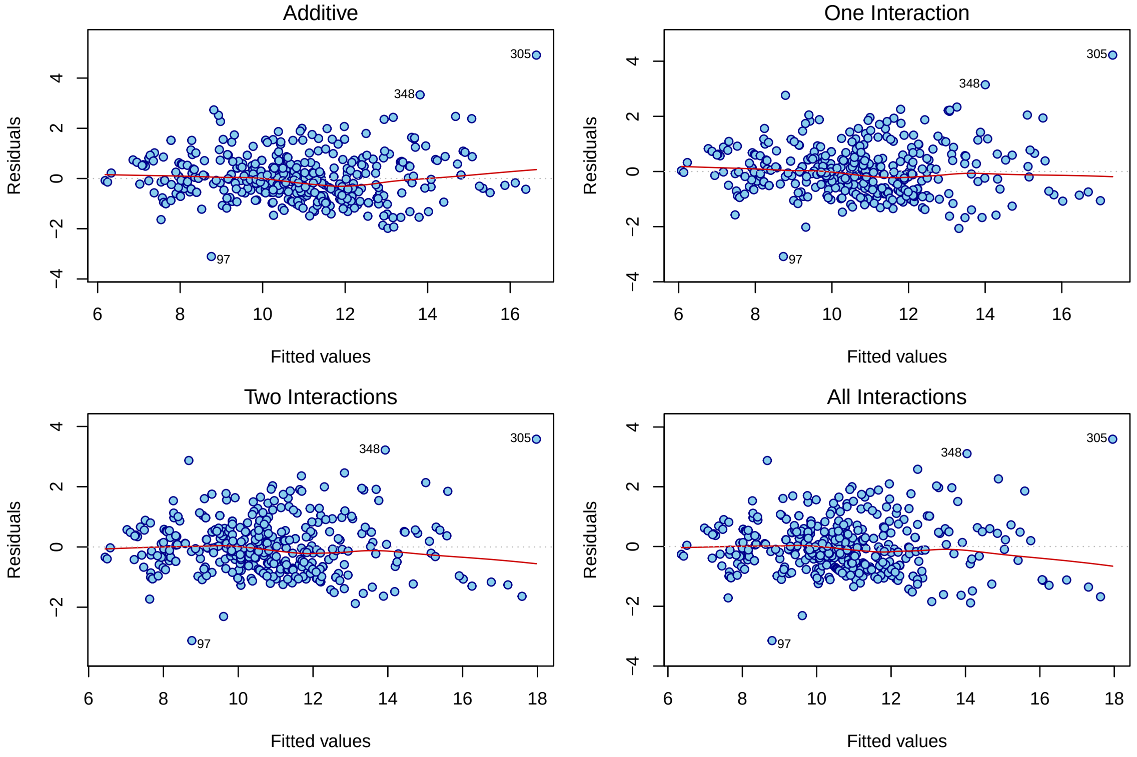
```
drop1(mInter, test = "F")
```

```
consumption ~ (fdrive + engine.size + lweight)^2
                    Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                           303.78 -101.642
fdrive:engine.size   2    1.9215 305.70 -103.064  1.2619 0.2842440
fdrive:lweight       2    8.6863 312.46  -94.112  5.7045 0.0036085 **
engine.size:lweight  1    8.7315 312.51  -92.052 11.4684 0.0007782 ***
```

Illustrations

`Cars2004nh` (subset, $n = 409$)

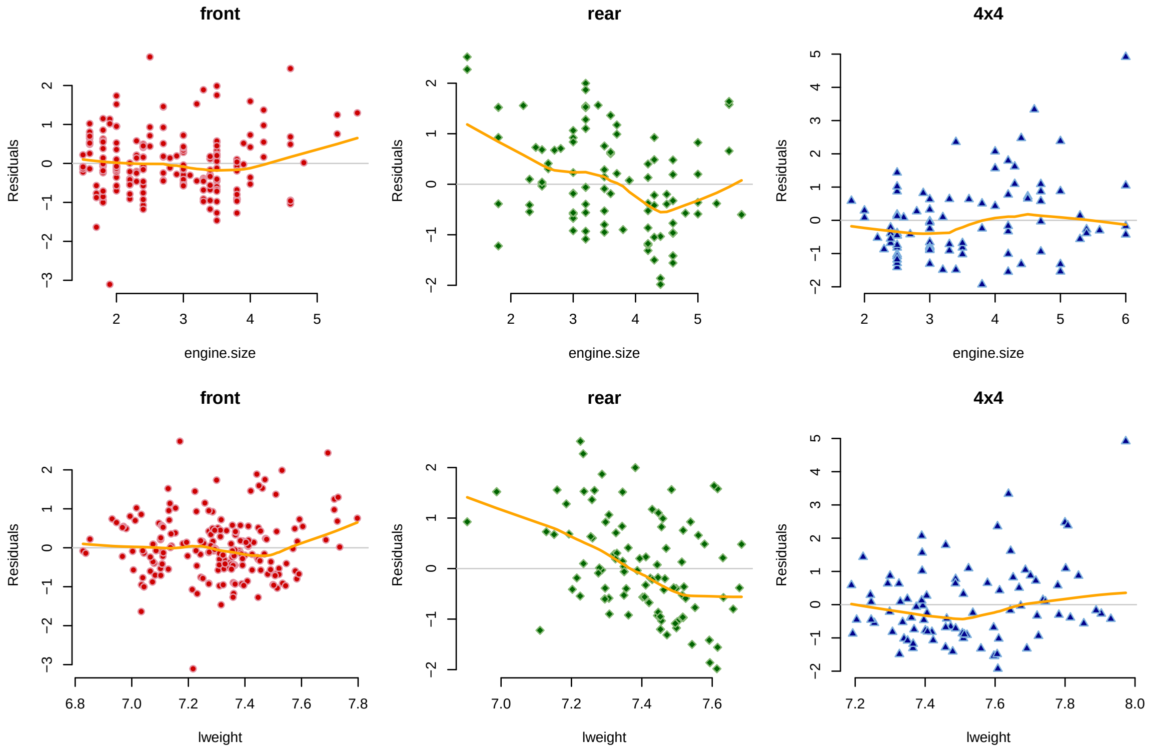
`consumption` $\sim$ `drive`, `engine size`, `log(weight)`



Illustrations

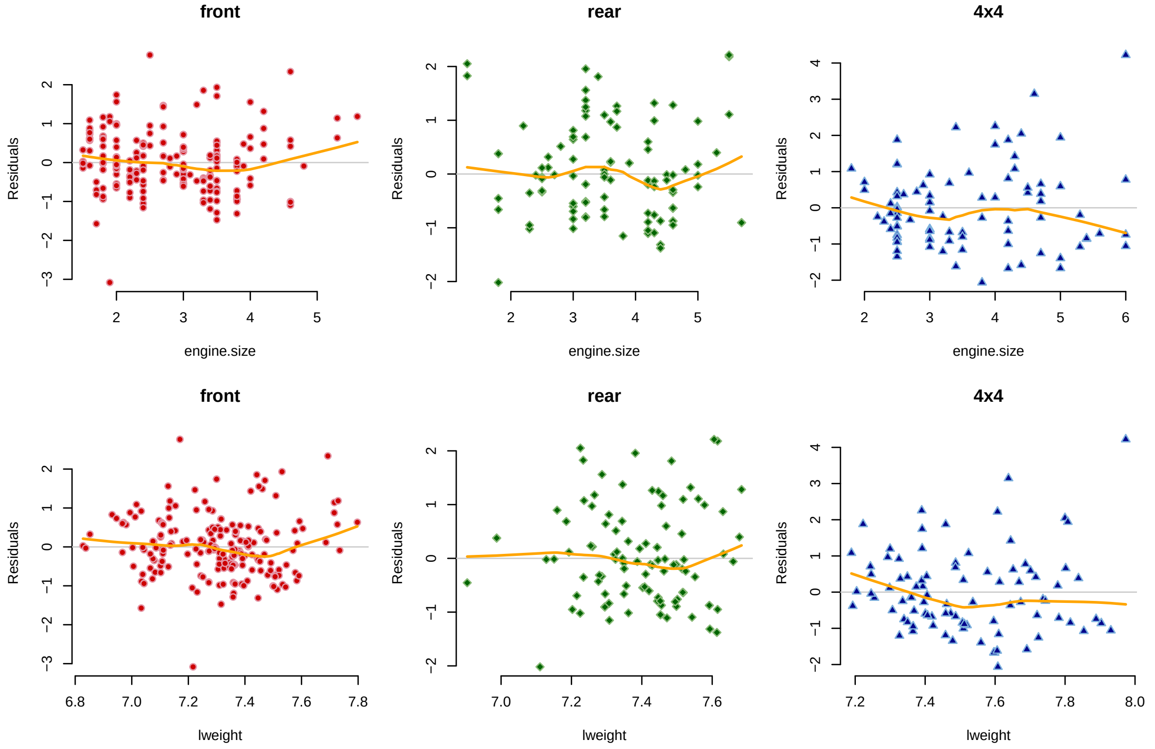
`Cars2004nh` (subset, $n = 409$)

`consumption` $\sim$ `drive` $+$ `engine size` $+$ `log(weight)`



`Cars2004nh` (subset, $n = 409$)

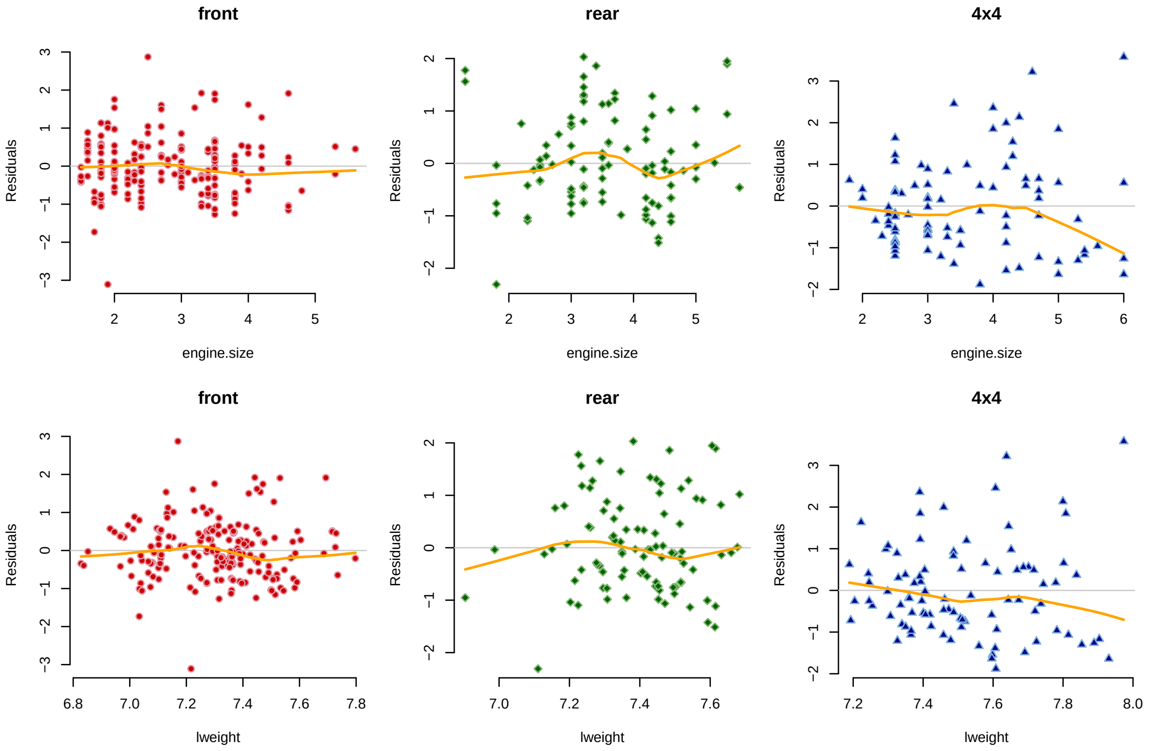
`consumption` $\sim$ `drive` $+$ `engine size` $+$ `log(weight)` $+$ `drive:log(weight)`



Illustrations

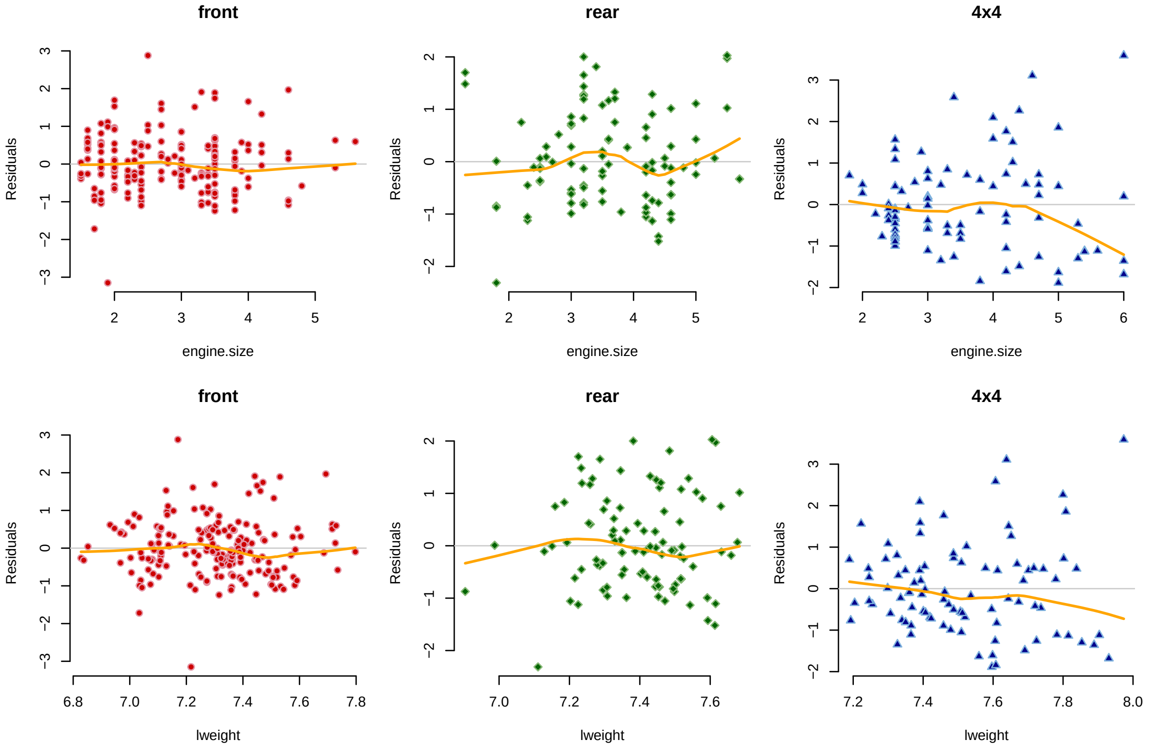
Cars2004nh (subset, $n = 409$)

consumption $\sim$ drive $+$ engine size $+$ log(weight) $+$ drive:log(weight) $+$ engine size:log(weight)



Cars2004nh (subset, $n = 409$)

consumption $\sim$ (drive $+$ engine size $+$ log(weight))^2



Illustrations

Cars2004nh (subset, $n = 409$)

consumption ∼ drive, engine size, log(weight)

```
anova(mAddit, mInter)
```

```
Model 1: consumption ~ fdrive + engine.size + lweight
Model 2: consumption ~ (fdrive + engine.size + lweight)^2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    404 343.69
2    399 303.78  5    39.906 10.483 1.813e-09 ***
```

```
anova(mInter1, mInter)
```

```
Model 1: consumption ~ fdrive + engine.size + lweight + fdrive:lweight
Model 2: consumption ~ (fdrive + engine.size + lweight)^2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    402 316.81
2    399 303.78  3    13.027 5.7034 0.0007864 ***
```

```
anova(mInter2, mInter)
```

```
Model 1: consumption ~ fdrive + engine.size + lweight + fdrive:lweight +
    engine.size:lweight
Model 2: consumption ~ (fdrive + engine.size + lweight)^2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    401 305.70
2    399 303.78  2    1.9215 1.2619 0.2842
```

5.5.5 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 13.1), ANOVA tables produced in a context of linear models provide on each row input of a certain F-statistic, now that based on Theorem 8.2. 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.

Illustrations

`consumption ∼ drive + log(weight) + drive:log(weight)`

Certain ANOVA table for the model:

$$\begin{aligned} m(z,\,w) = \beta_0 + \beta_1\,\mathbb{I}[z = \text{rear}] + \beta_2\,\mathbb{I}[z = \text{4x4}] + \beta_3\,\log(w) \\ + \beta_4\,\mathbb{I}[z = \text{rear}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{4x4}]\,\log(w) \end{aligned}$$

```
mInter1 <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow)
anova(mInter1)
```

```
Analysis of Variance Table

Response: consumption
                Df Sum Sq Mean Sq  F value    Pr(>F)    
fdrive           2 519.89  259.94  293.935 < 2.2e-16 ***
lweight          1 954.26  954.26 1079.040 < 2.2e-16 ***
fdrive:lweight   2  26.70   13.35   15.097 4.758e-07 ***
Residuals      403 356.40    0.88                      
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

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:

M_0: ~ 1,

M_A: $\sim \mathsf{A}$,

M_B: $\sim \mathsf{B}$,

M_{A+B}: $\sim \mathsf{A} + \mathsf{B}$,

M_{AB}: $\sim \mathsf{A} + \mathsf{B} + \mathsf{A}{:}\mathsf{B}$.

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 5.2 (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	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{A})$	$\star$	$\star$	$\star$
A: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	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{B})$	$\star$	$\star$	$\star$
A: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		

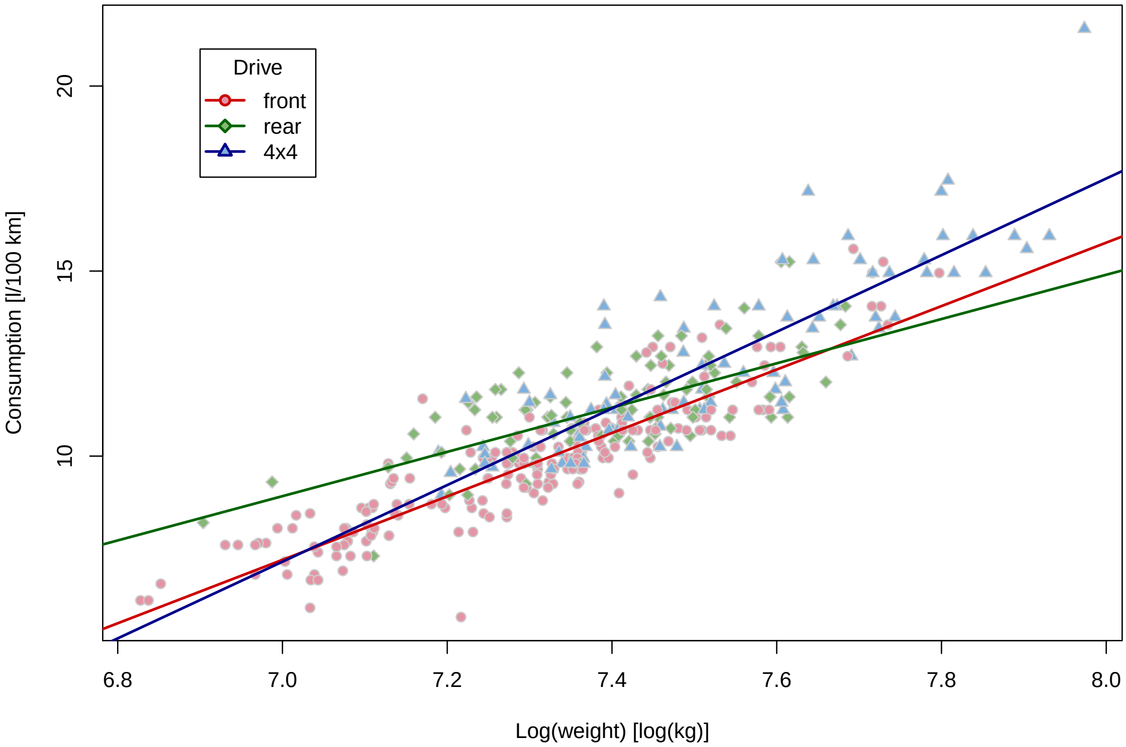
The row of the effect (term) E in the type I ANOVA table has in general the following interpretation and properties.

- 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 5.2. Similarly, interpretation of the F-tests is different for rows labeled equally B in the two tables in Example 5.2.
- 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 7.3 (Breakdown of the total sum of squares in a linear model with intercept).

Illustrations

Cars2004nh (subset, $n = 409$)

`consumption ∼ drive + log(weight) + drive:log(weight)`,
$\widehat{m}(z, w) = -52.80 + 19.84\,\mathbb{I}[z = \textbf{rear}] - 12.54\mathbb{I}[z = \textbf{4x4}] + 8.57\,\log(w) - 2.59\,\mathbb{I}[z = \textbf{rear}]\,\log(w) + 1.78\,\mathbb{I}[z = \textbf{4x4}]\,\log(w)$



Illustrations

Cars2004nh (subset, $n = 409$)

consumption ∼ drive + log(weight) + drive:log(weight)

Reference group pseudocontrasts for drive

$$
\begin{aligned}
m(z,\,w) = \beta_0 + \beta_1\,\mathbb{I}[z = \text{rear}] + \beta_2\,\mathbb{I}[z = \text{4x4}] + \beta_3\,\log(w)& \\
+ \beta_4\,\mathbb{I}[z = \text{rear}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{4x4}]\,\log(w)&
\end{aligned}
$$

```
mInter1 <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow)
anova(mInter1)
```

```
Analysis of Variance Table

Response: consumption
                Df Sum Sq Mean Sq  F value    Pr(>F)
fdrive           2 519.89  259.94  293.935 < 2.2e-16 ***
lweight          1 954.26  954.26 1079.040 < 2.2e-16 ***
fdrive:lweight   2  26.70   13.35   15.097 4.758e-07 ***
Residuals      403 356.40    0.88
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Cars2004nh (subset, $n = 409$)

consumption ∼ log(weight) + drive + drive:log(weight)

Reference group pseudocontrasts for drive

$$
\begin{aligned}
m(z,\,w) = \beta_0 + +\beta_1\,\log(w) + \beta_2\,\mathbb{I}[z = \text{rear}] + \beta_3\,\mathbb{I}[z = \text{4x4}]& \\
+ \beta_4\,\mathbb{I}[z = \text{rear}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{4x4}]\,\log(w)&
\end{aligned}
$$

```
mInter2 <- lm(consumption ~ lweight + fdrive + fdrive:lweight, data = CarsNow)
anova(mInter2)
```

```
Analysis of Variance Table

Response: consumption
                Df  Sum Sq Mean Sq  F value    Pr(>F)
lweight          1 1421.57 1421.57 1607.458 < 2.2e-16 ***
fdrive           2   52.58   26.29   29.726 9.079e-13 ***
lweight:fdrive   2   26.70   13.35   15.097 4.758e-07 ***
Residuals      403  356.40    0.88
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Type II ANOVA table

Example 5.3 (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	$\star$	$\mathsf{SS}(\mathsf{A}+\mathsf{B} \mid \mathsf{A})$	$\star$	$\star$	$\star$
A: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.

Illustrations

`Cars2004nh` (subset, $n = 409$)

`consumption` $\sim$ `drive + log(weight) + drive:log(weight)`

Reference group pseudocontrasts for `drive`

$$
\begin{aligned}
m(z,\,w) = \beta_0 + \beta_1\,\mathbb{I}[z = \text{rear}] + \beta_2\,\mathbb{I}[z = \text{4x4}] + \beta_3\,\log(w)\\
+\,\beta_4\,\mathbb{I}[z = \text{rear}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{4x4}]\,\log(w)
\end{aligned}
$$

```
mInter1 <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow)
car::Anova(mInter1, type = "II")
```

```
Anova Table (Type II tests)

Response: consumption
               Sum Sq  Df  F value    Pr(>F)
fdrive          52.58   2   29.726 9.079e-13 ***
lweight        954.26   1 1079.040 < 2.2e-16 ***
fdrive:lweight  26.70   2   15.097 4.758e-07 ***
Residuals      356.40 403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`Cars2004nh` (subset, $n = 409$)

`consumption` $\sim$ `log(weight) + drive + drive:log(weight)`

Reference group pseudocontrasts for `drive`

$$
\begin{aligned}
m(z,\,w) = \beta_0 + +\beta_1\,\log(w) + \beta_2\,\mathbb{I}[z = \text{rear}] + \beta_3\,\mathbb{I}[z = \text{4x4}]\\
+\,\beta_4\,\mathbb{I}[z = \text{rear}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{4x4}]\,\log(w)
\end{aligned}
$$

```
mInter2 <- lm(consumption ~ lweight + fdrive + fdrive:lweight, data = CarsNow)
car::Anova(mInter2, type = "II")
```

```
Anova Table (Type II tests)

Response: consumption
               Sum Sq  Df  F value    Pr(>F)
lweight        954.26   1 1079.040 < 2.2e-16 ***
fdrive          52.58   2   29.726 9.079e-13 ***
fdrive:lweight  26.70   2   15.097 4.758e-07 ***
Residuals      356.40 403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type III ANOVA table

Example 5.4 (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	$\star$	$\mathsf{SS}\big(\mathsf{A}+\mathsf{B}+\mathsf{A}{:}\mathsf{B}\,\big\vert\,\mathsf{B}+\mathsf{A}{:}\mathsf{B}\big)$	$\star$	$\star$	$\star$
B	$\star$	$\mathsf{SS}\big(\mathsf{A}+\mathsf{B}+\mathsf{A}{:}\mathsf{B}\,\big\vert\,\mathsf{A}+\mathsf{A}{:}\mathsf{B}\big)$	$\star$	$\star$	$\star$
A:B	$\star$	$\mathsf{SS}\big(\mathsf{A}+\mathsf{B}+\mathsf{A}{:}\mathsf{B}\,\big\vert\,\mathsf{A}+\mathsf{B}\big)$	$\star$	$\star$	$\star$
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 may depend on parameterizations of covariates included in the full model M_2. 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.

Illustrations

Cars2004nh (subset, $n = 409$)

consumption ~ drive + log(weight) + drive:log(weight)

Reference (first) group pseudocontrasts for drive

$$m(z,\,w) = \beta_0 + \beta_1\,\mathbb{I}[z = \text{rear}] + \beta_2\,\mathbb{I}[z = \text{4x4}] + \beta_3\,\log(w) \\ + \beta_4\,\mathbb{I}[z = \text{rear}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{4x4}]\,\log(w)$$

- β_3: slope of $\log(w)$ in group $z = \text{front}$

```
mInter <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow)
car::Anova(mInter, type = "III")
```

```
Anova Table (Type III tests)

Response: consumption
               Sum Sq  Df F value    Pr(>F)
(Intercept)    386.28   1 436.793 < 2.2e-16 ***
fdrive          26.49   2  14.979 5.310e-07 ***
lweight        542.30   1 613.216 < 2.2e-16 ***
fdrive:lweight  26.70   2  15.097 4.758e-07 ***
Residuals      356.40 403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cars2004nh (subset, $n = 409$)

consumption ~ drive + log(weight) + drive:log(weight)

Reference (last) group pseudocontrasts for drive

$$m(z,\,w) = \beta_0 + \beta_1\,\mathbb{I}[z = \text{front}] + \beta_2\,\mathbb{I}[z = \text{rear}] + \beta_3\,\log(w) \\ + \beta_4\,\mathbb{I}[z = \text{front}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{rear}]\,\log(w)$$

- β_3: slope of $\log(w)$ in group $z = \text{4x4}$

```
mInterSAS <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow,
                contrasts = list(fdrive = contr.SAS))
car::Anova(mInterSAS, type = "III")
```

```
Anova Table (Type III tests)

Response: consumption
               Sum Sq  Df F value    Pr(>F)
(Intercept)    247.68   1 280.063 < 2.2e-16 ***
fdrive          26.49   2  14.979 5.310e-07 ***
lweight        351.72   1 397.714 < 2.2e-16 ***
fdrive:lweight  26.70   2  15.097 4.758e-07 ***
Residuals      356.40 403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustrations

Cars2004nh (subset, $n = 409$)

`consumption ∼ drive + log(weight) + drive:log(weight)`

Sum contrasts for `drive`

$$m(z,\,w) = \beta_0 + \beta_1\,\mathbb{I}[z = \text{front}] + \beta_2\,\mathbb{I}[z = \text{rear}] - (\beta_1 + \beta_2)\,\mathbb{I}[z = \text{4x4}] + \beta_3\,\log(w)$$
$$+ \beta_4\,\mathbb{I}[z = \text{front}]\,\log(w) + \beta_5\,\mathbb{I}[z = \text{rear}]\,\log(w) - (\beta_4 + \beta_5)\,\mathbb{I}[z = \text{4x4}]\,\log(w)$$

- β_3: mean of the slopes of $\log(w)$ in the three `drive` groups

```
mIntersum <- lm(consumption ~ fdrive + lweight + fdrive:lweight, data = CarsNow,
                contrasts = list(fdrive = contr.sum))
car::Anova(mIntersum, type = "III")
```

```
Anova Table (Type III tests)

Response: consumption
               Sum Sq  Df F value    Pr(>F)    
(Intercept)    485.88   1 549.416 < 2.2e-16 ***
fdrive          26.49   2  14.979 5.310e-07 ***
lweight        728.22   1 823.440 < 2.2e-16 ***
fdrive:lweight  26.70   2  15.097 4.758e-07 ***
Residuals      356.40 403                      
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Chapter 6

Normal Linear Model

Until now, all proved theorems did not pose any distributional assumptions on the random vectors $\big(Y_i,\,\boldsymbol{X}_i^\top\big)^\top$, $\boldsymbol{X}_i = \big(X_{i,0},\dots,X_{i,k-1}\big)^\top$, $i = 1,\dots,n$, that represent the data (upon possible transformations of original covariates). We only assumed a certain form of the (conditional) expectation and the (conditional) covariance matrix of $\boldsymbol{Y} = \big(Y_1,\dots,Y_n\big)^\top$ given $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ (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.

6.1 Normal linear model

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

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\,\mathbf{I}_n\bigr). \tag{6.1}$$

Strictly speaking, the original data vectors $\bigl(Y_i,\,\boldsymbol{X}_i^\top\bigr)^\top$, $i = 1, \ldots, n$, do not have to be i.i.d. with respect to their joint distribution to satisfy (6.1). Remember that the joint density of the response vector and all the regressors can be decomposed as

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

Property (6.1) is related to the conditional density $f_{\boldsymbol{Y}|\mathbb{X}}$ which is then given as

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

On the other hand, the property (6.1) 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 (6.1) 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 6.1 Normal linear model with general data.

The data $\bigl(\boldsymbol{Y},\,\mathbb{X}\bigr)$*, satisfy a* normal linear model[1] *if*

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

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

Lemma 6.1 Error terms in a normal linear model.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\,\mathbf{I}_n\bigr)$*. The error terms*

$$\boldsymbol{\varepsilon} \;=\; \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} \;=\; \bigl(Y_1 - \boldsymbol{X}_1^\top\boldsymbol{\beta},\,\ldots,\,Y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\bigr)^\top \;=\; \bigl(\varepsilon_1,\,\ldots,\,\varepsilon_n\bigr)^\top$$

then satisfy

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

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

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

[1] *normální lineární model*

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).

(ii) follows from (i) and the fact that the conditional distribution $\varepsilon \,\big|\, \mathbb{X}$ 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).

❏

6.2 Properties of the least squares estimators under the normality

Theorem 6.2 Least squares estimators under the normality.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$*,* $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = r \leq k$*. Let* $\mathbb{L}_{m\times k}$ *be a real matrix with non-zero rows* $\mathbf{l}_1^\top, \dots, \mathbf{l}_m^\top$ *and* $\boldsymbol{\theta} := \mathbb{L}\boldsymbol{\beta} = \big(\mathbf{l}_1^\top\boldsymbol{\beta}, \dots, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \big(\theta_1, \dots, \theta_m\big)^\top$ *be a vector of linear combinations of regression parameters.*

If additionally $r = k$*, let* $\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ *be the least squares estimator of regression coefficients,* $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}} = \big(\mathbf{l}_1^\top\widehat{\boldsymbol{\beta}}, \dots, \mathbf{l}_m^\top\widehat{\boldsymbol{\beta}}\big)^\top = \big(\widehat{\theta}_1, \dots, \widehat{\theta}_m\big)^\top$ *and*

$$\mathbb{V} = \mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top = \big(v_{j,t}\big)_{j,t=1,\dots,m}, \qquad \mathbb{D} = \mathsf{diag}\left(\frac{1}{\sqrt{v_{1,1}}}, \dots, \frac{1}{\sqrt{v_{m,m}}}\right),$$

$$T_j = \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\mathsf{MS}_e\, v_{j,j}}}, \qquad j = 1, \dots, m, \qquad \boldsymbol{T} = \big(T_1, \dots, T_m\big)^\top = \frac{1}{\sqrt{\mathsf{MS}_e}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big).$$

The following then holds.

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

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

(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 $j = 1, \dots, m$*,* $\quad T_j \sim \mathsf{t}_{n-r}$.

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

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

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

Proof.

(i) We already know (Gauss-Markov theorem, Theorem 2.4) that $\mathbb{E}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X}\big) = \mathbb{X}\boldsymbol{\beta}$ and $\mathsf{var}\big(\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbb{H}$ (for any $\boldsymbol{\beta} \in \mathbb{R}^k$ and any $\sigma^2 > 0$). At the same time $\widehat{\boldsymbol{Y}} = \mathbb{H}\,\boldsymbol{Y}$ is a linear function of $\boldsymbol{Y}$ for which we assume $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n$. Hence, from properties of a (multivariate) normal distribution, we get $\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbb{H}\big)$.

(ii) Analogously to point (i), we already know (Lemma 2.7) that $\mathbb{E}\big(\boldsymbol{U} \,\big|\, \mathbb{X}\big) = \mathbf{0}_n$ and $\mathsf{var}\big(\boldsymbol{U} \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbb{M}$ (for any $\boldsymbol{\beta} \in \mathbb{R}^k$ and any $\sigma^2 > 0$). At the same time $\boldsymbol{U} = \mathbb{M}\,\boldsymbol{Y}$ is a linear function of $\boldsymbol{Y}$ for which we assume $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n$. Hence, from properties of a (multivariate) normal distribution, we get $\boldsymbol{U} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbf{0}_n,\, \sigma^2\,\mathbb{M}\big)$.

(iii) Based on Theorem 2.5 (Gauss-Markov theorem for linear combinations) we know that $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}}$ is BLUE of the vector parameter $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ (i.e., among the other things, $\mathbb{E}\big(\widehat{\boldsymbol{\theta}} \,\big|\, \mathbb{X}\big) = \boldsymbol{\theta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^k$) with $\mathsf{var}\big(\widehat{\boldsymbol{\theta}} \,\big|\, \mathbb{X}\big) = \sigma^2\,\mathbb{V}$, where $\mathbb{V} = \mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top$.

Since $r = k$, $\mathcal{M}\big(\mathbb{X}^\top\big) = \mathbb{R}^k$ (the property used also in the proof of Theorem 2.5). That is, for each $j = 1, \ldots, m$, $\mathbf{l}_j \in \mathcal{M}\big(\mathbb{X}^\top\big)$. In other words, $\mathcal{M}\big(\mathbb{L}^\top\big) \subset \mathcal{M}\big(\mathbb{X}^\top\big)$. That is, there exist a matrix $\mathbb{A}_{n\times m}$ such that $\mathbb{L}^\top = \mathbb{X}^\top\mathbb{A}$, i.e., $\mathbb{L} = \mathbb{A}^\top\mathbb{X}$. Then

$$\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}} = \mathbb{A}^\top\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{A}^\top\widehat{\boldsymbol{Y}}.$$

From point (i), we already know that $\widehat{\boldsymbol{Y}} \,\big|\, \mathbb{X} \sim \mathcal{N}_n$ and since $\widehat{\boldsymbol{\theta}}$ is a linear function of $\widehat{\boldsymbol{Y}}$, we get that also $\widehat{\boldsymbol{\theta}} \,\big|\, \mathbb{X} \sim \mathcal{N}_m$. In summary (the mean and the covariance matrix have been justified above), we have $\widehat{\boldsymbol{\theta}} \,\big|\, \mathbb{X} \sim \mathcal{N}_m\big(\boldsymbol{\theta},\, \sigma^2\,\mathbb{V}\big)$.

(iv) For a vector $\big(\widehat{\boldsymbol{Y}}^\top, \boldsymbol{U}^\top\big)^\top$ we can write

$$\begin{pmatrix}\widehat{\boldsymbol{Y}}\\ \boldsymbol{U}\end{pmatrix} = \begin{pmatrix}\mathbb{H}\,\boldsymbol{Y}\\ \mathbb{M}\,\boldsymbol{Y}\end{pmatrix} = \begin{pmatrix}\mathbb{H}\\ \mathbb{M}\end{pmatrix}\boldsymbol{Y}.$$

That is, the vector $\big(\widehat{\boldsymbol{Y}}^\top, \boldsymbol{U}^\top\big)^\top$ is a linear function of the random vector $\boldsymbol{Y}$ for which we assume $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n$. Hence, from properties of the normal distribution, a **joint** (conditional, given $\mathbb{X}$) distribution of the vector $\big(\widehat{\boldsymbol{Y}}^\top, \boldsymbol{U}^\top\big)^\top$ is also multivariate normal. Now, to show that $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{U}$ are (conditionally, given $\mathbb{X}$) independent, we only have to show that they are uncorrelated. This is easily obtained as follows.

$$\mathsf{cov}\big(\widehat{\boldsymbol{Y}}, \boldsymbol{U} \,\big|\, \mathbb{X}\big) = \mathsf{cov}\big(\mathbb{H}\,\boldsymbol{Y}, \mathbb{M}\,\boldsymbol{Y} \,\big|\, \mathbb{X}\big) = \mathbb{H}\,\underbrace{\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{X}\big)}_{\sigma^2\,\mathbf{I}_n}\,\mathbb{M}^\top = \sigma^2\,\mathbb{H}\,\mathbb{M} = \sigma^2\,\mathbb{Q}\,\underbrace{\mathbb{Q}^\top\,\mathbb{N}}_{\mathbf{0}}\,\mathbb{N}^\top = \sigma^2\,\mathbf{0},$$

where $\mathbb{Q}$ and $\mathbb{N}$ are matrices with orthonormal bases of the regression and residual space in their columns, respectively.

(v) $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}} = \mathbb{A}^\top\widehat{\boldsymbol{Y}}$ for some matrix $\mathbb{A}$ (see point iii), i.e., $\widehat{\boldsymbol{\theta}}$ is a measurable function of $\widehat{\boldsymbol{Y}}$.

$\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2$, i.e., SS_e is a measurable function of $\boldsymbol{U}$.

From point (iv), vectors $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{U}$ are (conditionally, given $\mathbb{X}$) independent and hence also $\widehat{\boldsymbol{\theta}}$ and SS_e are (conditionally, given $\mathbb{X}$) independent.

(vi) + (vii) Let us write the response vector $\boldsymbol{Y}$ as $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbf{0}_n, \sigma^2\mathbf{I}_n\big)$ and also unconditionally $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\big(\mathbf{0}_n, \sigma^2\mathbf{I}_n\big)$ (Lemma 6.1). Then,

$$\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta} = \mathbb{H}\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} = \underbrace{\mathbb{H}\mathbb{X}}_{\mathbb{X}}\,\boldsymbol{\beta} + \mathbb{H}\boldsymbol{\varepsilon} - \mathbb{X}\boldsymbol{\beta} = \mathbb{H}\boldsymbol{\varepsilon},$$

$$\boldsymbol{U} = \mathbb{M}\boldsymbol{Y} = \mathbb{M}\boldsymbol{\varepsilon} \qquad \text{(Lemma 2.6)}.$$

From here,

$$\big\|\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big\|^2 = \boldsymbol{\varepsilon}^\top\mathbb{H}^\top\mathbb{H}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbb{H}\boldsymbol{\varepsilon}, \qquad \mathbb{H} = \mathbb{Q}\mathbb{Q}^\top,$$

$$\big\|\boldsymbol{U}\big\|^2 = \boldsymbol{\varepsilon}^\top\mathbb{M}^\top\mathbb{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbb{M}\boldsymbol{\varepsilon}, \qquad \mathbb{M} = \mathbb{N}\mathbb{N}^\top,$$

where $\mathbb{Q}$ is an $n \times r$ matrix with the orthonormal basis of the regression space in its columns and $\mathbb{N}$ is an $n \times (n - r)$ matrix with the orthonormal basis of the residual space in its columns.

Let us first explore the residual sum of squares $\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2$. We have

$$\mathsf{SS}_e = \big\|\boldsymbol{U}\big\|^2 = \boldsymbol{\varepsilon}^\top\mathbb{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbb{N}\mathbb{N}^\top\boldsymbol{\varepsilon} = \big\|\mathbb{N}^\top\boldsymbol{\varepsilon}\big\|^2.$$

Hence,

$$\frac{1}{\sigma^2}\,\mathsf{SS}_e = \frac{1}{\sigma^2}\big\|\boldsymbol{U}\big\|^2 = \Big\|\frac{1}{\sigma}\,\mathbb{N}^\top\boldsymbol{\varepsilon}\Big\|^2.$$

As mentioned above, $\boldsymbol{\varepsilon}\,\big|\,\mathbb{X} \sim \mathcal{N}_n\big(\mathbf{0}_n,\,\sigma^2\mathbf{I}_n\big)$. Hence, from linearity and properties of the normal distribution,

$$\frac{1}{\sigma}\mathbb{N}^\top\boldsymbol{\varepsilon}\,\Big|\,\mathbb{X} \sim \mathcal{N}_{n-r}.$$

Further,

$$\mathbb{E}\Big(\frac{1}{\sigma}\mathbb{N}^\top\boldsymbol{\varepsilon}\,\Big|\,\mathbb{X}\Big) = \frac{1}{\sigma}\mathbb{N}^\top\mathbb{E}\big(\boldsymbol{\varepsilon}\,\big|\,\mathbb{X}\big) = \mathbf{0}_{n-r},$$

$$\mathsf{var}\Big(\frac{1}{\sigma}\mathbb{N}^\top\boldsymbol{\varepsilon}\,\Big|\,\mathbb{X}\Big) = \frac{1}{\sigma^2}\mathbb{N}^\top\underbrace{\mathsf{var}\big(\boldsymbol{\varepsilon}\,\big|\,\mathbb{X}\big)}_{\sigma^2\mathbf{I}_n}\mathbb{N} == \frac{\sigma^2}{\sigma^2}\mathbb{N}^\top\mathbb{N} = \mathbf{I}_{n-r}.$$

That is,

$$\frac{1}{\sigma}\mathbb{N}^\top\boldsymbol{\varepsilon}\,\Big|\,\mathbb{X} \sim \mathcal{N}_{n-r}\big(\mathbf{0}_{n-r},\,\mathbf{I}_{n-r}\big).$$

From here (sum of squares of independent normals),

$$\frac{1}{\sigma^2}\,\mathsf{SS}_e\,\Big|\,\mathbb{X} \sim \chi^2_{n-r}.$$

The above conditional distribution is the same for almost all values of the condition and hence also (unconditionally)

$$\frac{1}{\sigma^2}\,\mathsf{SS}_e \sim \chi^2_{n-r}.$$

The properties

$$\frac{1}{\sigma^2}\,\big\|\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big\|^2 \sim \chi^2_r, \qquad \frac{1}{\sigma^2}\,\big\|\widehat{\boldsymbol{Y}} - \mathbb{X}\boldsymbol{\beta}\big\|^2\Big|\,\mathbb{X} \sim \chi^2_r$$

are shown analogously.

(viii) We can write (for each $j = 0,\,\dots,\,m$)

$$T_j = \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\mathsf{MS}_e\,v_{j,j}}} = \frac{\dfrac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2\,v_{j,j}}}}{\sqrt{\dfrac{\mathsf{SS}_e}{\sigma^2\,(n-r)}}},$$

where from points (iii) and (vii)

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2\,v_{j,j}}}\,\Bigg|\,\mathbb{X} \sim \mathcal{N}(0,\,1), \qquad \frac{\mathsf{SS}_e}{\sigma^2}\,\Bigg|\,\mathbb{X} \sim \chi^2_{n-r}.$$

Moreover, from point (v), both statistics $\frac{\widehat{\theta}_j-\theta_j}{\sqrt{\sigma^2\,v_{j,j}}}$ and $\frac{\mathsf{SS}_e}{\sigma^2}$ are conditionally, given $\mathbb{X}$, independent. That is, from the definition of the Student t-distribution, $T_j\,\big|\,\mathbb{X} \sim \mathsf{t}_{n-r}$. This for almost all values of the condition $\mathbb{X}$. Hence, also unconditionally, $T_j \sim \mathsf{t}_{n-r}$.

(ix) We have

$$\boldsymbol{T} = \frac{1}{\sqrt{\mathsf{MS}_e}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) = \sqrt{\frac{(n-r)\sigma^2}{\mathsf{SS}_e}}\,\sqrt{\frac{1}{\sigma^2}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big),$$

where from points (iii) and (vii)

$$\sqrt{\frac{1}{\sigma^2}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)\,\Big|\,\mathbb{X} \sim \mathcal{N}_m\Big(\mathbf{0}_m,\,\frac{1}{\sigma^2}\,\mathbb{D}\,\sigma^2\,\mathbb{V}\,\mathbb{D}\Big) \equiv \mathcal{N}_m\Big(\mathbf{0}_m,\,\mathbb{D}\,\mathbb{V}\,\mathbb{D}\Big), \qquad \frac{\mathsf{SS}_e}{\sigma^2}\,\Bigg|\,\mathbb{X} \sim \chi^2_{n-r}.$$

Moreover, from point (v), both statistics $\sqrt{\frac{1}{\sigma^2}}\,\mathbb{D}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)$ and $\frac{\mathsf{SS}_e}{\sigma^2}$ are conditionally, given $\mathbb{X}$, independent. That is, from the definition of the multivariate t-distribution (Definition B.4), $\boldsymbol T \,\big|\, \mathbb{X} \sim \mathsf{mvt}_{m,\,n-r}\bigl(\mathbb{DVD}\bigr)$.

(x) The property that the matrix $\mathbb{V} = \mathbb{L}\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{L}^\top$ is invertible if $\mathsf{rank}\bigl(\mathbb{L}_{m\times r}\bigr) = m \le r (= k)$ was shown in the proof of Theorem 2.5 (Gauss-Markov for linear combinations). Further,

$$Q := \frac{1}{m}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)^\top\,\bigl(\mathsf{MS}_e\,\mathbb{V}\bigr)^{-1}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr) = \frac{\frac{1}{m}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)^\top\,\bigl(\sigma^2\,\mathbb{V}\bigr)^{-1}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)}{\frac{\mathsf{SS}_e}{\sigma^2\,(n-r)}}.$$

From point (iii), $\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)\,\big|\,\mathbb{X} \sim \mathcal{N}_m\bigl(\mathbf{0}_m,\,\sigma^2\,\mathbb{V}\bigr)$. Hence,

$$\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)^\top\,\bigl(\sigma^2\,\mathbb{V}\bigr)^{-1}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)\,\big|\,\mathbb{X} \sim \chi^2_m.$$

Further, from point (vii), $\frac{\mathsf{SS}_e}{\sigma^2}\,\big|\,\mathbb{X} \sim \chi^2_{n-r}$, and from point (v), both statistics $\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)^\top\,\bigl(\sigma^2\,\mathbb{V}\bigr)^{-1}\,\bigl(\widehat{\boldsymbol\theta}-\boldsymbol\theta\bigr)$ and $\frac{\mathsf{SS}_e}{\sigma^2}$ are conditionally, given $\mathbb{X}$, independent. That is, from the definition of the F-distribution, $Q\,\big|\,\mathbb{X} \sim \mathcal{F}_{m,\,n-r}$. This for almost all values of the condition $\mathbb{X}$. Hence, also unconditionally, $Q \sim \mathcal{F}_{m,\,n-r}$. ❑

Consequence of Theorem 6.2: Least squares estimator of the regression coefficients in a full-rank normal linear model.

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

$$\mathbb{V} = \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1} = \bigl(v_{j,t}\bigr)_{j,t=0,\dots,k-1},$$

$$\mathbb{D} = \mathsf{diag}\Bigl(\frac{1}{\sqrt{v_{0,0}}},\,\dots,\,\frac{1}{\sqrt{v_{k-1,k-1}}}\Bigr).$$

The following then holds.

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

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

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

(iv) $\boldsymbol T := \bigl(T_0,\dots,T_{k-1}\bigr)^\top = \dfrac{1}{\sqrt{\mathsf{MS}_e}}\,\mathbb{D}\,\bigl(\widehat{\boldsymbol\beta}-\boldsymbol\beta\bigr) \sim \mathsf{mvt}_{k,n-k}\bigl(\mathbb{DVD}\bigr)$, *conditionally given* $\mathbb{X}$.

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

Proof. Use $\mathbb{L} = \mathbf{I}_k$ in Theorem 6.2. ❑

Theorem 6.2 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.

6.2.1 Statistical inference in a full-rank normal linear model

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

Inference on a chosen regression coefficient

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

- **Standard error of $\widehat{\beta}_j$ and confidence interval for β_j**

  We have $\mathsf{var}\big(\widehat{\beta}_j \,\big|\, \mathbb{X}\big) = \sigma^2\, v_{j,j}$ (Lemma 2.2) which is unbiasedly estimated as $\mathsf{MS}_e\, v_{j,j}$ (Lemma 2.7). 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,

  $$\mathsf{S.E.}\big(\widehat{\beta}_j\big) = \sqrt{\mathsf{MS}_e\, v_{j,j}}. \tag{6.2}$$

  The standard error (6.2) is also the denominator of the t-statistic T_j from point (iii) of Consequence of Theorem 6.2. 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

  $$\begin{aligned} \beta_j^L &= \widehat{\beta}_j \;-\; \mathsf{S.E.}\big(\widehat{\beta}_j\big)\, \mathsf{t}_{n-k}\Big(1-\frac{\alpha}{2}\Big), \\ \beta_j^U &= \widehat{\beta}_j \;+\; \mathsf{S.E.}\big(\widehat{\beta}_j\big)\, \mathsf{t}_{n-k}\Big(1-\frac{\alpha}{2}\Big). \end{aligned} \tag{6.3}$$

  That is, for any $\boldsymbol{\beta}^0 = \big(\beta_0^0, \ldots, \beta_{k-1}^0\big)^\top \in \mathbb{R}^k$ and for any $j = 0, \ldots, k-1$

  $$\mathsf{P}\Big(\big(\beta_j^L,\, \beta_j^U\big) \ni \beta_j^0;\ \boldsymbol{\beta} = \boldsymbol{\beta}^0\Big) \;=\; 1-\alpha.$$

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

- **Test on a value of β_j**

  Suppose that for a given $\beta_j^0 \in \mathbb{R}$, we aim in testing H_0: $\beta_j = \beta_j^0$,
  H_1: $\beta_j \neq \beta_j^0$.

  The Wald-type test based on point (iii) of Consequence of Theorem 6.2 proceeds as follows:

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

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

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

  Analogously, also one-sided tests can be conducted.

Notation. In the following, for simplicity, expressions of the Wald-type confidence intervals like (6.3) will be briefly written as

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

[2] *směrodatná, příp. standardní chyba*

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 6.2 that the simultaneous $(1-\alpha)\,100\%$ confidence region for $\boldsymbol{\beta}$ is the set

$$\mathcal{S}(\alpha) \;=\; \Bigl\{\boldsymbol{\beta} \in \mathbb{R}^k :\; \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathsf{MS}_e^{-1}\mathbb{X}^\top\mathbb{X}\bigr)\,\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr) \;<\; k\,\mathcal{F}_{k,n-k}(1-\alpha)\Bigr\}.$$

  That is, for any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$

$$\mathsf{P}\Bigl(\mathcal{S}(\alpha) \ni \boldsymbol{\beta}^0;\; \boldsymbol{\beta} = \boldsymbol{\beta}^0\Bigr) \;=\; 1-\alpha.$$

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

  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 $\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{X}\bigr)$.

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

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

  The Wald-type test based on point (v) of Consequence of Theorem 6.2 proceeds as follows:

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

  Reject H$_0$ if $Q_0 \geq \mathcal{F}_{k,n-k}(1-\alpha)$.

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

[3] *simultánní konfidenční oblast*

Inference on a chosen linear combination

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

- **Standard error of $\widehat{\theta}$ and confidence interval for θ**

  We have $\mathsf{var}\big(\widehat{\theta}\,\big|\,\mathbb{X}\big) = \sigma^2\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}$ (Theorem 2.5) which is unbiasedly estimated as $\mathsf{MS}_e\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}$ (Lemma 2.7). Hence the standard error of $\widehat{\theta}$ is

  $$\mathsf{S.E.}\big(\widehat{\theta}\big) = \sqrt{\mathsf{MS}_e\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}}. \tag{6.4}$$

  The standard error (6.4) is also the denominator of the appropriate t-statistic from point (viii) of Theorem 6.2. Hence the lower and the upper bounds of the Wald-type $(1-\alpha)\,100\%$ confidence interval for θ based on this t-statistic are

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

  That is, for any $\theta^0 \in \mathbb{R}$,

  $$\mathsf{P}\Big(\big(\theta^L,\,\theta^U\big) \ni \theta^0;\; \theta = \theta^0\Big) \;=\; 1-\alpha.$$

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

- **Test on a value of θ**

  Suppose that for a given $\theta^0 \in \mathbb{R}$, we aim in testing H_0: $\theta = \theta^0$,
  H_1: $\theta \neq \theta^0$.

  The Wald-type test based on point (viii) of Theorem 6.2 proceeds as follows:

  Test statistic: $$T_0 = \frac{\widehat{\theta}-\theta^0}{\mathsf{S.E.}\big(\widehat{\theta}\big)} = \frac{\widehat{\theta}-\theta^0}{\sqrt{\mathsf{MS}_e\,\mathbf{l}^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}}}.$$

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

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

  Analogously, also one-sided tests can be conducted.

Simultaneous inference on a set of linear combinations

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

- **Simultaneous confidence region for $\boldsymbol{\theta}$**

  It follows from point (x) of Theorem 6.2 that the simultaneous $(1-\alpha)\,100\%$ confidence region for $\boldsymbol{\theta}$ is the set

  $$\mathcal{S}(\alpha) \;=\; \Bigl\{\boldsymbol{\theta} \in \mathbb{R}^m : \;(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \Bigl\{\mathsf{MS}_e\, \mathbb{L}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{L}^\top\Bigr\}^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \;<\; m\,\mathcal{F}_{m,n-k}(1-\alpha)\Bigr\}.$$

  That is, for any $\boldsymbol{\theta}^0 \in \mathbb{R}^m$

  $$\mathsf{P}\Bigl(\mathcal{S}(\alpha) \ni \boldsymbol{\theta}^0;\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \;=\; 1-\alpha.$$

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

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

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

  The Wald-type test based on point (x) of Theorem 6.2 proceeds as follows:

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

  Reject H$_0$ if $Q_0 \geq \mathcal{F}_{m,n-k}(1-\alpha)$.

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

Note. If we 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 16 that under certain conditions, all those methods of statistical inference remain *asymptotically* valid even if normality does not hold.

6.3 Confidence interval for the model based mean, prediction interval

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

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

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 6.1).

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 $\boldsymbol{\varepsilon} = \bigl(\varepsilon_1, \ldots, \varepsilon_n\bigr)^\top$. 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 $\mu_{new} := \mathbb{E}\bigl(Y_{new} \,\big|\, \boldsymbol{X}_{new} = \boldsymbol{x}_{new}\bigr) = \boldsymbol{x}_{new}^\top\boldsymbol{\beta}$.

(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 6.3 Confidence interval for the model based mean, prediction interval.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\bigr)$, $\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k$ (full-rank model), $\widehat{\boldsymbol{\beta}} = \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ is the LSE of the regression parameters $\boldsymbol{\beta}$. Let $\boldsymbol{x}_{new} \in \mathcal{X}$, $\boldsymbol{x}_{new} \neq \mathbf{0}_k$. 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) The quantity $\widehat{\mu}_{new} := \boldsymbol{x}_{new}^\top\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\mu_{new} := \boldsymbol{x}_{new}^\top\boldsymbol{\beta}$. The standard error of $\widehat{\mu}_{new}$ is

$$\mathsf{S.E.}\bigl(\widehat{\mu}_{new}\bigr) = \sqrt{\mathsf{MS}_e\, \boldsymbol{x}_{new}^\top\, \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_{new}}$$

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

$$\bigl(\mu_{new}^L,\, \mu_{new}^U\bigr) \;\equiv\; \widehat{\mu}_{new} \;\pm\; \mathsf{S.E.}\bigl(\widehat{\mu}_{new}\bigr)\, t_{n-k}\Bigl(1 - \frac{\alpha}{2}\Bigr). \tag{6.5}$$

(ii) A (random) interval with the bounds

$$\bigl(Y_{new}^L,\, Y_{new}^U\bigr) \;\equiv\; \widehat{\mu}_{new} \;\pm\; \mathsf{S.E.P.}\bigl(\boldsymbol{x}_{new}\bigr)\, t_{n-k}\Bigl(1 - \frac{\alpha}{2}\Bigr), \tag{6.6}$$

where

$$\mathsf{S.E.P.}\bigl(\boldsymbol{x}_{new}\bigr) = \sqrt{\mathsf{MS}_e\Bigl\{1 + \boldsymbol{x}_{new}^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_{new}\Bigr\}}, \tag{6.7}$$

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

Notes.

- Statement (i) of Theorem 6.3 says that for any $\boldsymbol{x}_{new} \in \mathcal{X}$ and any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$ providing $\mu^0_{new} = \boldsymbol{x}^\top_{new}\boldsymbol{\beta}^0$ the following holds:
$$\mathsf{P}\Big(\big(\mu^L_{new},\, \mu^U_{new}\big) \ni \mu^0_{new};\, \boldsymbol{\beta} = \boldsymbol{\beta}^0\Big) \;=\; 1-\alpha.$$
- Statement (ii) of Theorem 6.3 says that for any $\boldsymbol{x}_{new} \in \mathcal{X}$, any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$ and any $\sigma_0^2 \in (0,\,\infty)$ leading to $Y^0_{new} = \boldsymbol{x}^\top_{new}\boldsymbol{\beta}^0 + \varepsilon^0_{new}$, where ε^0_{new} is sampled from $\mathcal{N}\big(0,\,\sigma_0^2\big)$ independently of $\boldsymbol{\varepsilon}$, the following holds:
$$\mathsf{P}\Big(\big(Y^L_{new},\, Y^U_{new}\big) \ni Y^0_{new};\, \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\Big) \;=\; 1-\alpha.$$

Proof.

(i) The BLUE property of the estimator $\widehat{\mu}_{new}$ follows from the Gauss-Markov theorem for linear combinations (Theorem 2.5), the form of the confidence interval for μ_{new} follows from point (viii) of Theorem 6.2, see also Section 6.2.1.

(ii) Conditionally, given $X_1,\,\dots,\,X_n,\,X_{new}$:

$Y_{new} \sim \mathcal{N}(\mu_{new},\,\sigma^2)$ (since $\varepsilon_{new} \sim \mathcal{N}(0,\,\sigma^2)$),

$\widehat{\mu}_{new} \sim \mathcal{N}(\mu_{new},\,\sigma^2\,v)$, $\quad v = \boldsymbol{x}^\top_{new}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{x}_{new}$ (point iii of Theorem 6.2),

$\boldsymbol{Y}$ and Y_{new} are independent (since $\boldsymbol{\varepsilon}$ and ε_{new} are independent).

At the same time, $\widehat{\mu}_{new}$ is a (linear) function of $\boldsymbol{Y}$ and hence also $\widehat{\mu}_{new}$ and Y_{new} are independent (conditionally, given $X_1,\,\dots,\,X_n,\,X_{new}$).

It now follows from above and properties of the normal distribution that
$$\big(Y_{new} - \widehat{\mu}_{new}\big)\,\big|\,X_1,\,\dots,\,X_n,\,X_{new} \sim \mathcal{N}\big(0,\,\sigma^2\,(1+v)\big),$$
or written in an alternative way
$$\frac{Y_{new} - \widehat{\mu}_{new}}{\sqrt{\sigma^2\,(1+v)}}\,\Bigg|\,X_1,\,\dots,\,X_n,\,X_{new} \sim \mathcal{N}\big(0,\,1\big).$$

Further, SS_e and $\widehat{\mu}_{new}$ are conditionally independent (point v of Theorem 6.2). At the same time, SS_e is a measurable function of $\boldsymbol{Y}$ and hence, SS_e and Y_{new} are again conditionally independent. It then follows that also $(Y_{new} - \widehat{\mu}_{new})$ and SS_e are conditionally independent.

Finally, $\frac{\mathsf{SS}_e}{\sigma^2} \sim \chi^2_{n-r}$ (both conditionally and unconditionally, see proof of point vii of Theorem 6.2).

Hence, conditionally, given $X_1,\,\dots,\,X_n,\,X_{new}$,
$$\frac{\dfrac{Y_{new} - \widehat{\mu}_{new}}{\sqrt{\sigma^2\,(1+v)}}}{\dfrac{\mathsf{SS}_e}{\sigma^2\,(n-k)}} \;=\; \frac{Y_{new} - \widehat{\mu}_{new}}{\sqrt{\mathsf{MS}_e\,(1+v)}} \;\sim\; \mathsf{t}_{n-k}.$$

This holds for almost all values of $X_1 = x_1,\,\dots,\,X_n = x_n,\,X_{new} = x_{new}$ and hence for any $\boldsymbol{x}_{new} \in \mathcal{X}$, any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$ and any $\sigma_0^2 \in (0,\,\infty)$ leading to $Y^0_{new} = \boldsymbol{x}^\top_{new}\boldsymbol{\beta}^0 + \varepsilon^0_{new}$
$$\mathsf{P}\left(\left|\frac{Y^0_{new} - \mu_{new}}{\sqrt{\mathsf{MS}_e\,(1+v)}}\right| < \mathsf{t}_{n-k}\big(1-\frac{\alpha}{2}\big),\;\; \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\right) \;=\; 1-\alpha.$$

That is, if we define
$$\text{S.E.P.}(\boldsymbol{x}_{new}) \;:=\; \sqrt{\mathsf{MS}_e\,(1+v)} \;=\; \sqrt{\mathsf{MS}_e\,\Big(1+\boldsymbol{x}^\top_{new}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{x}_{new}\Big)},$$
we get
$$\mathsf{P}\Bigg(\Big(\widehat{\mu}_{new} \pm \text{S.E.P.}(\boldsymbol{x}_{new})\,\mathsf{t}_{n-k}\big(1-\frac{\alpha}{2}\big)\Big) \ni Y^0_{new},\;\; \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\Bigg) \;=\; 1-\alpha.$$

❑

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

- The interval with the bounds (6.5) is called the *confidence interval for the model based mean.*
- The interval with the bounds (6.6) is called the *prediction interval.*
- The quantity (6.7) is called the *standard error of prediction.*

Terminology *(Fitted regression function).*

The function

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

which, by Theorem 6.3, 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. 6.5) is called the *confidence band around the regression function*;[5]

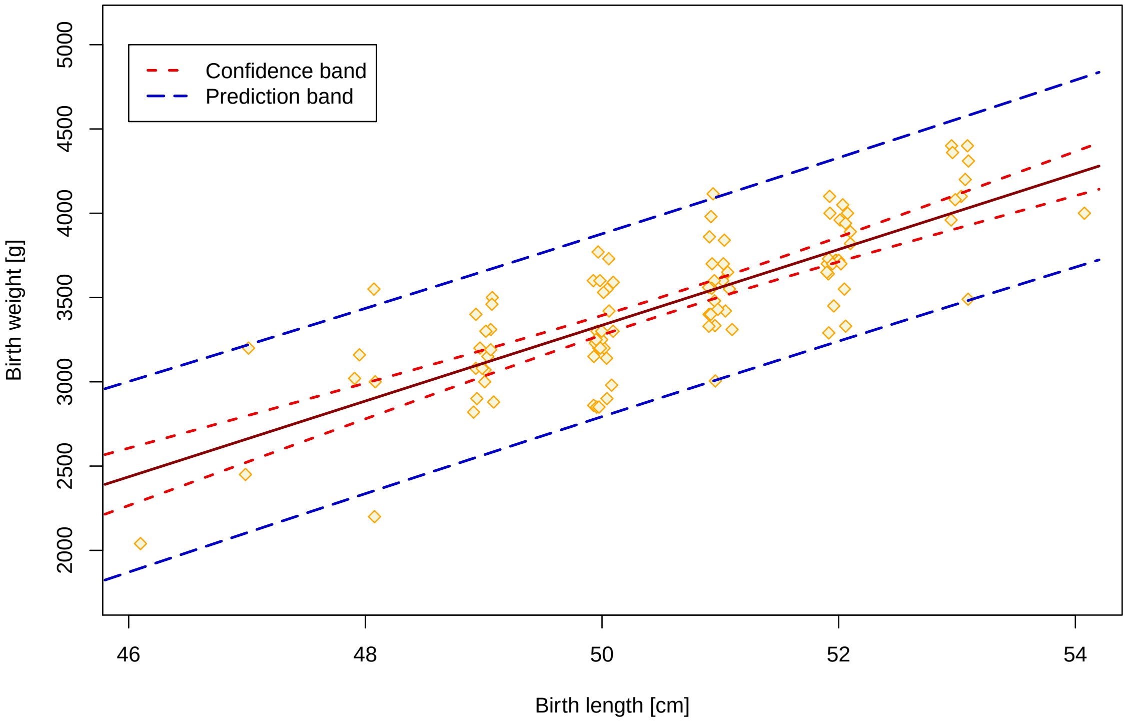
(ii) The band based on the prediction intervals (Eq. 6.6) is called the *prediction band.*[6]

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

Illustrations

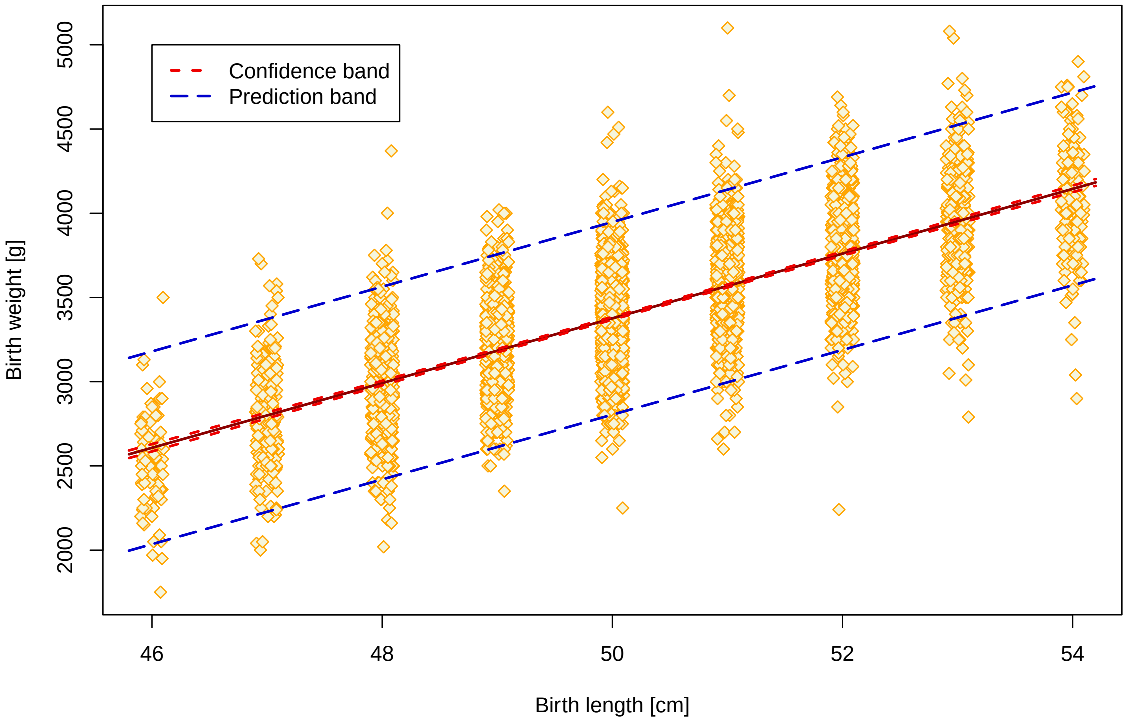
`Kojeni` ($n = 99$)

`bweight` $\sim$ `blength`



`Hosi0` ($n = 4838$)

`bweight` $\sim$ `blength`



6.4 Distribution of the linear hypotheses test statistics under the alternative

This section of the text contains materials which will not be examined.

Section 6.2 provided classical tests of the linear hypotheses (hypotheses on the values of linear combinations of regression coefficients). To allow for power or sample size calculations, we additionally need distribution of the test statistics under the alternatives.

Theorem 6.4 Distribution of the linear hypothesis test statistics under the alternative.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$. Let $\mathbf{l} \neq \mathbf{0}_k$ and let $\widehat{\theta} = \mathbf{l}^\top\widehat{\boldsymbol{\beta}}$ be the LSE of the parameter $\theta = \mathbf{l}^\top\boldsymbol{\beta}$. Let θ^0, $\theta^1 \in \mathbb{R}$, $\theta^0 \neq \theta^1$ and let

$$T_0 = \frac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\, \mathbf{l}^\top \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}}}.$$

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

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

Note. The statistic T_0 is the test statistic to test the null hypothesis H_0: $\theta = \theta^0$ using point (viii) of Theorem 6.2.

Theorem 6.5 Distribution of the linear hypotheses test statistics under the alternative.

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

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

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

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

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 6.2.

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]

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

Chapter 7

Coefficient of Determination

In this chapter, we develop a quantity called *coefficient of determination* which is often considered as a basic measure of a model quality. Nevertheless, as will be shown, it only measures a prediction quality (in certain sense) of the model. For derivations used in this chapter, we do not need to know the transformation which links original covariates and regressors considered in a particular model. That is, we will now assume that data are represented by n random vectors $\bigl(Y_i,\, \boldsymbol{X}_i^\top\bigr)^\top$, $i = 1, \ldots, n$, and a linear model with the response vector $\boldsymbol{Y} = \bigl(Y_1, \ldots, Y_n\bigr)^\top$ and the model matrix $\mathbb{X}$ with vectors $\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top$ in rows is used to model the conditional expectations $\mathbb{E}\bigl(Y_i \bigm| \boldsymbol{X}_i\bigr)$, $i = 1, \ldots, n$.

7.1 Intercept only model

Notation *(Response sample mean).*

The sample mean over the response vector $\boldsymbol{Y} = \bigl(Y_1, \dots, Y_n\bigr)^\top$ 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 7.1 Regression and total sums of squares in a linear model.
Consider a linear model $\boldsymbol{Y} \bigm| \mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\bigr)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k$. *The following expressions define the following quantities:*

(i) Regression sum of squares[1] *and corresponding degrees of freedom:*

$$\mathsf{SS}_R = \bigl\|\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}_n\bigr\|^2 = \sum_{i=1}^{n}\bigl(\widehat{Y}_i - \overline{Y}\bigr)^2, \quad \nu_R = r - 1,$$

(ii) Total sum of squares[2] *and corresponding degrees of freedom:*

$$\mathsf{SS}_T = \bigl\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\bigr\|^2 = \sum_{i=1}^{n}\bigl(Y_i - \overline{Y}\bigr)^2, \quad \nu_T = n - 1.$$

Lemma 7.1 Model with intercept only.
Let $\boldsymbol{Y} \sim \bigl(\mathbf{1}_n\gamma, \zeta^2\mathbf{I}_n\bigr)$. *Then*

(i) $\widehat{\boldsymbol{Y}} = \overline{Y}\mathbf{1}_n = \bigl(\overline{Y}, \dots, \overline{Y}\bigr)^\top$.

(ii) $\mathsf{SS}_e = \mathsf{SS}_T$.

Proof. This is a full-rank model with $\mathbb{X} = \mathbf{1}_n$. Further,

$$\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1} = \bigl(\mathbf{1}_n^\top\mathbf{1}_n\bigr)^{-1} = \frac{1}{n}, \qquad \mathbb{X}^\top\boldsymbol{Y} = \mathbf{1}_n^\top\boldsymbol{Y} = \sum_{i=1}^{n} Y_i.$$

Hence $\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$ and $\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\gamma} = \mathbf{1}_n\overline{Y} = \overline{Y}\mathbf{1}_n$.

❑

[1] *regresní součet čtverců* [2] *celkový součet čtverců*

7.2 Models with intercept

Lemma 7.2 Identity in a linear model with intercept.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *where* $\mathbf{1}_n \in \mathcal{M}(\mathbb{X})$. *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 $\mathbf{1}_n \in \mathcal{M}(\mathbb{X})$.

❑

Lemma 7.3 Breakdown of the total sum of squares in a linear model with intercept.

Let $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ *where* $\mathbf{1}_n \in \mathcal{M}(\mathbb{X})$. *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 7.1) $\widehat{\boldsymbol{Y}} = \overline{Y}\mathbf{1}_n$. Then $\mathsf{SS}_T = \mathsf{SS}_e$, $\mathsf{SS}_R = 0$.

The identity $\mathsf{SS}_T = \mathsf{SS}_e + \mathsf{SS}_R$ for general rank $r \geq 1$ can 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

$$\sum_{i=1}^n Y_i\widehat{Y}_i = \boldsymbol{Y}^\top\widehat{\boldsymbol{Y}} = \boldsymbol{Y}^\top\mathbb{H}\boldsymbol{Y}, \qquad \sum_{i=1}^n \widehat{Y}_i^2 = \widehat{\boldsymbol{Y}}^\top\widehat{\boldsymbol{Y}} = \boldsymbol{Y}^\top\mathbb{H}\mathbb{H}\boldsymbol{Y} = \boldsymbol{Y}^\top\mathbb{H}\boldsymbol{Y}.$$

❑

Using materials from the chapter on submodels (Chapter 8) we even do not need the algebra used above. In the following, let $r = \mathsf{rank}\bigl(\mathbb{X}\bigr) > 1$. Then, model $\boldsymbol{Y} \,|\, \mathbb{X} \sim \bigl(\mathbf{1}_n\beta^0,\, \sigma^2\mathbf{I}_n\bigr)$ is a submodel of the model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$ and by Lemma 7.1, $\mathsf{SS}_T = \mathsf{SS}_e^0$. Further, from definition of SS_R, it equals to $\mathsf{SS}_R = \bigl\|\boldsymbol{D}\bigr\|^2$, where $\boldsymbol{D} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0$. By point (iv) of Theorem 8.1 (on a submodel), $\bigl\|\boldsymbol{D}\bigr\|^2 = \mathsf{SS}_e^0 - \mathsf{SS}_e$. In other words,

$$\mathsf{SS}_R = \mathsf{SS}_T - \mathsf{SS}_e.$$

❑

7.3 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 $\bigl(Y_i,\,\boldsymbol{X}_i^\top\bigr)^\top$, $i = 1,\dots,n$, are a random sample from some joint distribution of a generic random vector $\bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$, $\boldsymbol{X} = \bigl(X_0,\,\dots,\,X_{k-1}\bigr)^\top$ and the conditional distribution $Y \,|\, \boldsymbol{X}$ can be described by a linear model, i.e.,

$$\mathbb{E}\bigl(Y \,\big|\, \boldsymbol{X}\bigr) = \boldsymbol{X}^\top\boldsymbol{\beta}, \qquad \mathsf{var}\bigl(Y \,\big|\, \boldsymbol{X}\bigr) = \sigma^2 \tag{7.1}$$

for some $\boldsymbol{\beta} = \bigl(\beta_0,\,\dots,\,\beta_{k-1}\bigr)^\top \in \mathbb{R}^k$ and some $\sigma^2 > 0$, which leads to the linear model

$$\boldsymbol{Y} \,|\, \mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\mathbf{I}_n\bigr), \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},\ \mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}$$

for the data. As usually, we assume $\mathsf{rank}\bigl(\mathbb{X}\bigr) = r \le k < n$ (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.,

$$\mathbb{E}\bigl(Y\bigr) = \gamma, \qquad \mathsf{var}\bigl(Y\bigr) = \zeta^2. \tag{7.2}$$

This corresponds to the only intercept linear model

$$\boldsymbol{Y} \sim \bigl(\mathbf{1}_n\gamma,\,\zeta^2\mathbf{I}_n\bigr)$$

for the data with a model matrix $\mathbf{1}_n$ of rank 1.

Suppose now that all model parameters ($\boldsymbol{\beta}$, γ, σ^2, ζ^2) related to the distribution of the random vector $\bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$ 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 $\bigl(Y_{new},\,\boldsymbol{X}_{new}^\top\bigr)^\top$ which is distributed as the generic vector $\bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$. Nevertheless, for simplicity of notation, we will not use the subscript $_{new}$ and will simply work with the random vector $\bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$ whose distribution satisfies (7.1) and (7.2).

Suppose further that the random vector $\bigl(Y,\,\boldsymbol{X}^\top\bigr)^\top$ is defined on a probability space $\bigl(\Omega,\,\mathcal{A},\,\mathsf{P}\bigr)$ 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 $\mathsf{L}_2(\boldsymbol{X}) = \mathsf{L}_2\bigl(\Omega,\,\sigma(\boldsymbol{X}),\,\mathsf{P}|_{\sigma(\boldsymbol{X})}\bigr)$. Further, let $\sigma(\emptyset) = \bigl\{\emptyset,\ \Omega\bigr\}$ be a trivial σ-algebra on Ω, $\mathsf{P}|_{\sigma(\emptyset)}$ the related restricted probability measure and $\mathsf{L}_2(\emptyset) = \mathsf{L}_2\bigl(\Omega,\,\sigma(\emptyset),\,\mathsf{P}|_{\sigma(\emptyset)}\bigr)$.

A problem of prediction of a value of the random variable $Y \in \mathsf{L}_2\bigl(\Omega,\,\mathcal{A},\,\mathsf{P}\bigr)$ classically corresponds to looking for $\widehat{Y}$ which in a certain sense minimizes the mean squared error of prediction[3] (MSEP)

$$\mathsf{MSEP}\bigl(\widehat{Y}\bigr) = \mathbb{E}\bigl(\widehat{Y} - Y\bigr)^2.$$

We now distinguish 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)

$$\widehat{Y} = \operatorname*{argmin}_{\widetilde{Y}\in\mathsf{L}_2(\emptyset)} \mathbb{E}\bigl(\widetilde{Y} - Y\bigr)^2 = \operatorname*{argmin}_{\widetilde{Y}\in\mathbb{R}} \mathbb{E}\bigl(\widetilde{Y} - Y\bigr)^2 = \mathbb{E}\bigl(Y\bigr) = \gamma := \widehat{Y}^M.$$

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

$$\mathsf{MSEP}\bigl(\widehat{Y}^M\bigr) = \mathbb{E}\bigl(\gamma - Y\bigr)^2 = \mathsf{var}\bigl(Y\bigr) = \zeta^2.$$

[3] *střední čtvercová chyba predikce*

(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)

$$\widehat{Y} = \underset{\widetilde{Y} \in \mathsf{L}_2(\boldsymbol{X})}{\mathsf{argmin}}\; \mathbb{E}\bigl(\widetilde{Y} - Y\bigr)^2 = \mathbb{E}\bigl(Y \bigm| \boldsymbol{X}\bigr) = \boldsymbol{X}^\top\boldsymbol{\beta} := \widehat{Y}^C,$$

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

$$\begin{aligned}\mathsf{MSEP}\bigl(\widehat{Y}^C\bigr) = \mathbb{E}\bigl(\boldsymbol{X}^\top\boldsymbol{\beta} - Y\bigr)^2 = \mathbb{E}\Bigl[\mathbb{E}\Bigl\{\bigl(\boldsymbol{X}^\top\boldsymbol{\beta} - Y\bigr)^2 \Bigm| \boldsymbol{X}\Bigr\}\Bigr]& \\ = \mathbb{E}\Bigl\{\mathsf{var}\bigl(Y \bigm| \boldsymbol{X}\bigr)\Bigr\} = \mathbb{E}\bigl(\sigma^2\bigr) = \sigma^2.&\end{aligned}$$

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.

To compare the marginal and the conditional prediction, we introduce the ratio of the two MSEP's:

$$\frac{\mathsf{MSEP}\bigl(\widehat{Y}^C\bigr)}{\mathsf{MSEP}\bigl(\widehat{Y}^M\bigr)} = \frac{\sigma^2}{\zeta^2}.$$

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.

7.4 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 $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1,\, \dots,\, \widehat{Y}_n\big)^\top$ 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

$$\frac{\frac{1}{n-r}\,\mathsf{SS}_e}{\frac{1}{n-1}\mathsf{SS}_T} = \frac{n-1}{n-r}\,\frac{\mathsf{SS}_e}{\mathsf{SS}_T}. \tag{7.3}$$

Alternatively, if $Y_i \overset{\text{i.i.d.}}{\sim} Y$, $i = 1, \dots, n$, $Y \sim \mathcal{N}(\gamma,\, \zeta^2)$, that is, if $Y_1, \dots, Y_n$ 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[4] (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

$$\frac{\frac{1}{n}\,\mathsf{SS}_e}{\frac{1}{n}\mathsf{SS}_T} = \frac{\mathsf{SS}_e}{\mathsf{SS}_T}. \tag{7.4}$$

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,

$$\underbrace{\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)^2}_{\mathsf{SS}_T} = \underbrace{\sum_{i=1}^{n}\big(Y_i - \widehat{Y}_i\big)^2}_{\mathsf{SS}_e} + \underbrace{\sum_{i=1}^{n}\big(\widehat{Y}_i - \overline{Y}\big)^2}_{\mathsf{SS}_R},$$

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 (7.3) and (7.4) then motivate the following definition.

[4] *maximálně věrohodný odhad*

Definition 7.2 Coefficients of determination.

Consider a linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}(\mathbb{X}) = r$ *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[5] *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[6] *of the linear model.*

Notes.

- By Theorem 7.3, $\mathsf{SS}_T = \mathsf{SS}_e + \mathsf{SS}_R$ 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 $\mathbb{E}\big(Y \,\big|\, \boldsymbol{X}\big)$. Even a model with a low value of R^2 (R^2_{adj}) might be useful with respect to modelling the conditional mean $\mathbb{E}\big(Y \,\big|\, \boldsymbol{X}\big)$. The model is perhaps only useless for prediction purposes.

[5] *koeficient determinace* [6] *upravený koeficient determinace*

Chapter 8

Submodels

In this chapter, we will again consider the original response-covariate data being represented by n random vectors $\bigl(Y_i,\,\boldsymbol{Z}_i^\top\bigr)^\top$, $\boldsymbol{Z}_i = \bigl(Z_{i,1},\,\dots,\,Z_{i,p}\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1,\dots,n$. The main aim is still to find a suitable model to express the (conditional) response expectation $\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z}\bigr)$, 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

$$\mathbb{X}^0 = \begin{pmatrix} \boldsymbol{X}_1^{0\,\top} \\ \vdots \\ \boldsymbol{X}_n^{0\,\top} \end{pmatrix}, \quad \begin{array}{rcl} \boldsymbol{X}_1^0 & = & \boldsymbol{t}_0(\boldsymbol{Z}_1), \\ & \vdots & \\ \boldsymbol{X}_n^0 & = & \boldsymbol{t}_0(\boldsymbol{Z}_n), \end{array} \qquad \mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^{\top} \\ \vdots \\ \boldsymbol{X}_n^{\top} \end{pmatrix}, \quad \begin{array}{rcl} \boldsymbol{X}_1 & = & \boldsymbol{t}(\boldsymbol{Z}_1), \\ & \vdots & \\ \boldsymbol{X}_n & = & \boldsymbol{t}(\boldsymbol{Z}_n). \end{array} \tag{8.1}$$

Briefly, we will write

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

Let (almost surely),

$$\mathsf{rank}(\mathbb{X}^0) = r_0, \qquad \mathsf{rank}(\mathbb{X}) = r, \tag{8.2}$$

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:

$$\text{Model } \mathsf{M}_0 : \boldsymbol{Y}\,|\,\mathbb{Z} \sim \bigl(\mathbb{X}^0\boldsymbol{\beta}^0,\,\sigma^2\,\mathbf{I}_n\bigr),$$

$$\text{Model } \mathsf{M} : \;\; \boldsymbol{Y}\,|\,\mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\,\mathbf{I}_n\bigr),$$

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.

8.1 Submodel

Definition 8.1 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 $\mathbb{E}(\boldsymbol{Y} \,|\, \mathbb{Z}) = \mathbb{X}^0\boldsymbol{\beta}^0$.

8.1.1 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

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

of the n-dimensional Euclidean space as

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

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

Further,

[1] *podmodel* [2] *vnořený model*

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

$$\mathcal{M}\bigl(\mathbb{X}^0\bigr)^{\perp} = \mathcal{M}\bigl(\mathbb{N}^0\bigr) = \mathcal{M}\Bigl(\bigl(\mathbb{Q}^1,\ \mathbb{N}\bigr)\Bigr).$$

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

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

We have

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

8.1.2 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).
- $\boldsymbol{U}^0 = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0 = \mathbb{M}^0\boldsymbol{Y} = \bigl(\mathbb{Q}^1{\mathbb{Q}^1}^\top + \mathbb{N}\mathbb{N}^\top\bigr)\boldsymbol{Y}$: residuals of the submodel.
- $\mathsf{SS}_e^0 = \|\boldsymbol{U}^0\|^2$: 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 $\mathcal{M}\bigl(\mathbb{Q}^1\bigr)$, i.e.,

$$\boldsymbol{D} = \mathbb{Q}^1\,{\mathbb{Q}^1}^\top\boldsymbol{Y} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0 = \boldsymbol{U}^0 - \boldsymbol{U}. \tag{8.3}$$

Theorem 8.1 On a submodel.

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

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

(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*

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

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

$$F_0 = \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{r - r_0}}{\dfrac{\mathsf{SS}_e}{n - r}} = \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{\nu_e^0 - \nu_e}}{\dfrac{\mathsf{SS}_e}{\nu_e}} \sim \mathcal{F}_{r - r_0,\,n - r} = \mathcal{F}_{\nu_e^0 - \nu_e,\,\nu_e}. \tag{8.4}$$

Proof. **Proof/calculations are available in the handnotes.** ❑

8.1.3 Series of submodels

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

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

where, analogously to (8.1), 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 (8.2), 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

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

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

Theorem 8.2 On submodels.

Consider three normal *linear models* M : $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\bigr)$, M_1 : $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \mathcal{N}_n\bigl(\mathbb{X}^1\boldsymbol{\beta}^1,\, \sigma^2\,\mathbf{I}_n\bigr)$, M_0 : $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \mathcal{N}_n\bigl(\mathbb{X}^0\boldsymbol{\beta}^0,\, \sigma^2\,\mathbf{I}_n\bigr)$ *such that* $\mathsf{M}_0 \subset \mathsf{M}_1 \subset \mathsf{M}$. *Let the (smallest) submodel* M_0 *hold, i.e., let* $\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z}\bigr) \in \mathcal{M}\bigl(\mathbb{X}^0\bigr)$. *Then*

$$F_{0,1} = \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e^1}{r_1 - r_0}}{\dfrac{\mathsf{SS}_e}{n - r}} = \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e^1}{\nu_e^0 - \nu_e^1}}{\dfrac{\mathsf{SS}_e}{\nu_e}} \sim \mathcal{F}_{r_1 - r_0,\, n-r} = \mathcal{F}_{\nu_e^0 - \nu_e^1,\, \nu_e}. \tag{8.5}$$

Proof. **Proof/calculations are available in the handnotes.** ❑

Note. Both F-statistics (8.4) and (8.5) 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 (8.4) or (8.5), 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 (8.4) and (8.5) contain certain $\mathsf{SS}\big(B \,\big|\, A\big)$ in their numerators.
- Point (iv) of Theorem 8.1 gives

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

8.1.4 Statistical test to compare nested models

Theorems 8.1 and 8.2 provide a way to compare two nested models by the mean of a statistical test.

F-test on a submodel based on Theorem 8.1

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

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

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 (8.4) from Theorem 8.1 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 8.2

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

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

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

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 (8.5) from Theorem 8.2 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 $F_{0,1} = f_{0,1}$: $p = 1 - \mathsf{CDF}_{\mathcal{F},\, r_1 - r_0, n - r}\big(f_{0,1}\big)$.

8.2 Omitting some regressors

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

where the submodel matrix $\mathbb{X}^0_{n\times k_0}$ is obtained by omitting selected columns from the model matrix $\mathbb{X}_{n\times k}$. In other words, some regressors are omitted from the original regressor vectors $\boldsymbol{X}^0,\dots,\boldsymbol{X}^{k-1}$ to get the submodel and the matrix $\mathbb{X}^0$. In the following, we will consider only the full-rank models, that is, $\mathsf{rank}(\mathbb{X}^0_{n\times k_0}) = k_0$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$. Without the loss of generality, let

$$\mathbb{X} \;=\; \bigl(\mathbb{X}^0,\, \mathbb{X}^1\bigr), \qquad 0 < \mathsf{rank}\bigl(\mathbb{X}^0_{n\times k_0}\bigr) = k_0 < k = \mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) < n, \mathsf{rank}\bigl(\mathbb{X}^1_{n\times k_1}\bigr) = k_1 = k - k_0.$$

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.

Lemma 8.3 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 $\mathcal{M}\bigl(\mathbb{X}^1\bigr) \perp \mathcal{M}\bigl(\mathbb{X}^0\bigr)$ then

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

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

(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 $\bigl(\mathbb{R}^n,\, \mathcal{B}_n\bigr)$ then

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

Proof. Let

$$\mathbb{M}^0 \;:=\; \mathbf{I}_n \;-\; \mathbb{X}^0\bigl(\mathbb{X}^{0\top}\mathbb{X}^0\bigr)^{-1}\mathbb{X}^{0\top}$$

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

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

Hence

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

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 $\mathcal{M}\bigl(\mathbb{X}^0\bigr)^{\perp}$, 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

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

Let

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

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 $\mathcal{M}\bigl(\mathbb{X}^0\bigr)$, i.e.,

$$\mathcal{M}\bigl(\mathbb{X}^0\bigr) \;=\; \mathcal{M}\bigl(\mathbb{Q}^0\bigr).$$

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

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

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

Then
$$\boldsymbol{D} \;=\; \mathbb{H}^1\,\boldsymbol{Y} \;=\; \bigl(\mathbb{M}^0\,\mathbb{X}^1\bigr)\,\bigl(\mathbb{X}^{1\top}\mathbb{M}^0\mathbb{X}^1\bigr)^{-1}\,\mathbb{X}^{1\top}\,\underbrace{\mathbb{M}^0\boldsymbol{Y}}_{\boldsymbol{U}^0}. \tag{8.6}$$

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

(i) If $\mathcal{M}\bigl(\mathbb{X}^1\bigr) \perp \mathcal{M}\bigl(\mathbb{X}^0\bigr)$, 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 (8.6), while realizing $\mathbb{M}^0 = \mathbb{M}^0\,\mathbb{M}^0$, we get
$$\boldsymbol{D} \;=\; \mathbb{X}^1\,\bigl(\mathbb{X}^{1\top}\mathbb{X}^1\bigr)^{-1}\mathbb{X}^{1\top}\boldsymbol{Y}.$$

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

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

❑

Note. If we take the residual sum of squares as a measure of a quality of the model, point (ii) of Theorem 8.3 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).

8.3 Linear constraints

Suppose that a full-rank linear model $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$ 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

$$\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big) := \big\{\boldsymbol{v} :\, \boldsymbol{v} = \mathbb{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^k,\ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big\}, \tag{8.7}$$

where $\mathbb{L}_{m\times k}$ is a given real matrix with m linearly independent rows, $m < k$ 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 $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$.

Definition 8.2 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*:* $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$*, if the response expectation* $\mathbb{E}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big)$ *under the model* M_0 *is assumed to lie in a space* $\mathcal{M}\big(\mathbb{X};\, \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0\big)$*, where* $\mathbb{L}_{m\times k}$ *is a real matrix with* m *linearly independent rows,* $m < k$ *and* $\boldsymbol{\theta}^0 \in \mathbb{R}^m$ *is a given vector.*

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

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

Definition 8.3 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* $\mathsf{SS}(\boldsymbol{\beta}) = \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\big\|^2$ *over* $\boldsymbol{\beta} \in \mathbb{R}^k$ *subject to* $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$*. For the submodel* $\mathsf{M}_0 : \boldsymbol{Y}\,|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big),\ \mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$*, 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: $\mathsf{SS}_e^0 := \big\|\boldsymbol{U}^0\big\|^2$.

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

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

Note. The fitted values could also be defined as

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

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

Theorem 8.4 On a submodel given by linear constraints.

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

[3] *podmodel zadaný lineárními omezeními*

(i) *There is a unique minimizer* $\boldsymbol{b}^0$ *to* $\mathsf{SS}(\boldsymbol{\beta}) = \left\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\right\|^2$ *subject to* $\mathbb{L}\boldsymbol{\beta} = \boldsymbol{\theta}^0$ *and is given as*

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

where $\widehat{\boldsymbol{\beta}} = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ *is the (unconstrained) least squares estimator of the vector* $\boldsymbol{\beta}$.

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

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

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

$$\left\|\boldsymbol{D}\right\|^2 = \mathsf{SS}_e^0 - \mathsf{SS}_e = \big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big). \tag{8.8}$$

Proof. First mention that under our assumptions, the matrix $\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top$ is invertible. This follows from Theorem 2.5 (Gauss–Markov for linear combinations).

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

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

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

$$\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta} = \mathbb{X}^\top\boldsymbol{Y} - \mathbb{L}^\top\boldsymbol{\lambda}. \tag{8.9}$$

Note that the linear system (8.9) is consistent for any $\boldsymbol{\lambda} \in \mathbb{R}^m$ and any $\boldsymbol{Y} \in \mathbb{R}^n$. This follows from the fact that $\mathcal{M}\big(\mathbb{L}^\top\big) \subset \mathbb{R}^k = \mathcal{M}\big(\mathbb{X}^\top\big)$ Hence the right-hand-side of the system (8.9) 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 (8.9) lies in $\mathcal{M}\big(\mathbb{X}^\top\mathbb{X}\big)$, for any $\boldsymbol{\beta} \in \mathbb{R}^k$. On top of that, $\mathcal{M}\big(\mathbb{X}^\top\big) = \mathbb{R}^k = \mathcal{M}\big(\mathbb{X}^\top\mathbb{X}\big)$. This proves that there always exist a solution to the linear system (8.9).

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,

$$\begin{aligned}\boldsymbol{b}^0(\boldsymbol{\lambda}) &= \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y} - \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\boldsymbol{\lambda}\\ &= \widehat{\boldsymbol{\beta}} - \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\boldsymbol{\lambda}.\end{aligned}$$

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

$$\begin{aligned}\mathbb{L}\boldsymbol{b}^0(\boldsymbol{\lambda}) &= \boldsymbol{\theta}^0\\ \mathbb{L}\widehat{\boldsymbol{\beta}} - \mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{L}^\top\boldsymbol{\lambda} &= \boldsymbol{\theta}^0\\ \underbrace{\mathbb{L}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{L}^\top}_{\text{invertible as we already know}}\boldsymbol{\lambda} &= \mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0.\end{aligned}$$

That is,

$$\boldsymbol{\lambda} = \Big\{\mathbb{L}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\,\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big).$$

Finally,

$$\boldsymbol{b}^0 = \widehat{\boldsymbol{\beta}} - \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big),$$

$$\widehat{\boldsymbol{Y}}^0 = \mathbb{X}\boldsymbol{b}^0 = \widehat{\boldsymbol{Y}} - \mathbb{X}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big).$$

It then follows that

$$\boldsymbol{D} = \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^0 = \mathbb{X}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big).$$

Hence,

$$\begin{aligned}\big\|\boldsymbol{D}\big\|^2 &= \big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\mathbb{L}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)\\ &= \big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big).\end{aligned}$$

It remains to be shown that $\big\|\boldsymbol{D}\big\|^2 = \mathsf{SS}_e^0 - \mathsf{SS}_e$. We have

$$\begin{aligned}\mathsf{SS}_e^0 = \big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^0\big\|^2 &= \Big\|\underbrace{\boldsymbol{Y} - \widehat{\boldsymbol{Y}}}_{\boldsymbol{U}\in\mathcal{M}(\mathbb{X})^\perp} + \underbrace{\mathbb{X}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)}_{\boldsymbol{D}\in\mathcal{M}(\mathbb{X})}\Big\|^2\\ &= \big\|\boldsymbol{U}\big\|^2 + \big\|\boldsymbol{D}\big\|^2 = \mathsf{SS}_e + \big\|\boldsymbol{D}\big\|^2.\end{aligned}$$

❑

8.3.1 F-statistic to verify a set of linear constraints

Let us take the expression (8.8) 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 (8.4):

$$\begin{aligned}F_0 = \frac{\dfrac{\mathsf{SS}_e^0 - \mathsf{SS}_e}{k - r_0}}{\dfrac{\mathsf{SS}_e}{n-k}} &= \frac{\dfrac{\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)^\top\Big\{\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)}{m}}{\dfrac{\mathsf{SS}_e}{n-k}}\\ &= \frac{1}{m}\,\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)^\top\Big\{\mathsf{MS}_e\,\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}^0\big)\\ &= \frac{1}{m}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big)^\top\Big\{\mathsf{MS}_e\,\mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big), \qquad (8.10)\end{aligned}$$

where $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}}$ is the LSE of the vector parameter $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ in the linear model $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$ without constraints. Note now that (8.10) is exactly equal to the Wald-type statistic Q_0 (see page 175) that we used in Section 6.2 to test the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ on a vector parameter $\boldsymbol{\theta}$ in a *normal* linear model $\boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$. If normality can be assumed, point (x) of Theorem 6.2 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 $\mathcal{F}_{m,n-k}$. This shows that the Wald-type test on the estimable vector parameter in a normal linear model based on Theorem 6.2 is equivalent to the submodel F-test based on Theorem 8.1.

8.3.2 t-statistic to verify a linear constraint

Consider $\mathbb{L} = \mathbf{l}^\top$, $\mathbf{l} \in \mathbb{R}^k$, $\mathbf{l} \neq \mathbf{0}_k$ and a normal linear model $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$. Take $\theta^0 \in \mathbb{R}$ and consider the submodel given by $m = 1$ linear constraint $\mathbf{l}^\top\boldsymbol{\beta} = \theta^0$. Let $\widehat{\theta} = \mathbf{l}^\top\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the least squares estimator of the regression coefficients $\boldsymbol{\beta}$ in the model without constraints. The statistic (8.10) then takes the form

$$F_0 = \frac{1}{m}\,\bigl(\widehat{\theta} - \theta^0\bigr)\,\Bigl\{\mathsf{MS}_e\,\mathbf{l}^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbf{l}\Bigr\}^{-1}\,\bigl(\widehat{\theta} - \theta^0\bigr) = \left(\frac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\,\mathbf{l}^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbf{l}}}\right)^2 = T_0^2,$$

where

$$T_0 = \frac{\widehat{\theta} - \theta^0}{\sqrt{\mathsf{MS}_e\,\mathbf{l}^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbf{l}}}$$

is the Wald-type test statistic introduced in Section 6.2 (on page 174) to test the null hypothesis H_0: $\theta = \theta^0$ in a *normal* linear model $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$. Point (viii) of Theorem 6.2 provided that under the null hypothesis H_0: $\theta = \theta^0$, the statistic T_0 follows the Student t-distribution t_{n-k} which is indeed in agreement with the fact that $T_0^2 = F_0$ follows the F-distribution $\mathcal{F}_{1,n-k}$.

8.4 Overall F-test

Lemma 8.5 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* $\mathbf{1}_n \in \mathcal{M}\big(\mathbb{X}\big)$*. 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*

$$F_0 = \frac{R^2}{1-R^2}\cdot\frac{n-r}{r-1}. \tag{8.11}$$

Proof.

- $R^2 = 1 - \frac{\mathsf{SS}_e}{\mathsf{SS}_T}$ and according to Lemma 7.1: $\mathsf{SS}_T = \mathsf{SS}_e^0$.
- 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 (8.11) 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 9

Checking Model Assumptions

In Chapter 3, 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 3, we assume that data are represented by n random vectors $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $\boldsymbol{Z}_i = \bigl(Z_{i,1},\, \dots,\, Z_{i,p}\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$ $i = 1, \dots, n$. Further, we assume that possibly two sets of regressors have been derived from the covariates:

(i) $\boldsymbol{X}_i$, $i = 1, \dots, n$, 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

$$\mathbb{X}_{n\times k} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \Bigl(\boldsymbol{X}^0,\, \dots,\, \boldsymbol{X}^{k-1}\Bigr).$$

For most practical problems, $\boldsymbol{X}^0 = \bigl(1,\, \dots,\, 1\bigr)^\top$ (almost surely).

(ii) $\boldsymbol{V}_i$, $i = 1, \dots, n$, 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

$$\mathbb{V}_{n\times l} = \begin{pmatrix} \boldsymbol{V}_1^\top \\ \vdots \\ \boldsymbol{V}_n^\top \end{pmatrix} = \Bigl(\boldsymbol{V}^1,\, \dots,\, \boldsymbol{V}^l\Bigr).$$

Primarily, we will assume that the model matrix $\mathbb{X}$ is sufficient to be able to assume that $\mathbb{E}\bigl(\boldsymbol{Y} \,\big|\, \mathbb{Z}\bigr) = \mathbb{E}\bigl(\boldsymbol{Y} \,\big|\, \mathbb{X}\bigr) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} = \bigl(\beta_0,\, \dots,\, \beta_{k-1}\bigr)^\top \in \mathbb{R}^k$. That is, we will arrive from assuming

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

or even from assuming normality, i.e.,

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

The task is now to verify appropriateness of those assumptions that, in principle, consist of four subassumptions outlined in Chapter 3, which can all be written while using the error terms $\boldsymbol{\varepsilon} = \bigl(\varepsilon_1,\, \dots,\, \varepsilon_n\bigr)^\top = \bigl(Y_1 - \boldsymbol{X}_1^\top\boldsymbol{\beta},\, \dots,\, Y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\bigr)^\top = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}$:

(A1) Correct regression function $\equiv$ (Conditionally) errors with zero mean $\equiv \mathbb{E}\bigl(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\bigr) = 0$, $i = 1, \dots, n$.

(A2) (Conditional) homoscedasticity of errors $\equiv \mathsf{var}\bigl(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\bigr) = \sigma^2 = \text{const}$, $i = 1, \dots, n$.

(A3) (Conditionally) uncorrelated/independent errors $\equiv \mathsf{cov}\bigl(\varepsilon_i,\, \varepsilon_j \,\big|\, \mathbb{Z}\bigr) = 0$, $i \neq j$.

(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 $\equiv \mathsf{var}\big(\varepsilon_i\big) = \sigma^2 =$ 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 $\equiv \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}$.

9.1 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 11. 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 $\mathcal{M}\big(\mathbb{X}\big)$):
$$\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top = \big(h_{i,t}\big)_{i,t=1,\dots,n};$$
- Fitted values $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \big(\widehat{Y}_1, \dots, \widehat{Y}_n\big)^\top$;
- Projection matrix into the residual space $\mathcal{M}\big(\mathbb{X}\big)^\perp$:
$$\mathbb{M} = \mathbf{I}_n - \mathbb{H} = \big(m_{i,t}\big)_{i,t=1,\dots,n};$$
- Residuals: $\boldsymbol{U} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \mathbb{M}\boldsymbol{Y} = \big(U_1, \dots, U_n\big)^\top$;
- 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 $\mathcal{M}\big(\mathbb{G}\big)$):
$$\mathbb{H}_g = \mathbb{G}\big(\mathbb{G}^\top\mathbb{G}\big)^{-}\mathbb{G}^\top = \big(h_{g,i,t}\big)_{i,t=1,\dots,n};$$
- Fitted values $\widehat{\boldsymbol{Y}}_g = \mathbb{H}_g\boldsymbol{Y} = \big(\widehat{Y}_{g,1}, \dots, \widehat{Y}_{g,n}\big)^\top$;
- Projection matrix into the residual space $\mathcal{M}\big(\mathbb{G}\big)^\perp$:
$$\mathbb{M}_g = \mathbf{I}_n - \mathbb{H}_g = \big(m_{g,i,t}\big)_{i,t=1,\dots,n};$$
- Residuals: $\boldsymbol{U}_g = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_g = \mathbb{M}_g\boldsymbol{Y} = \big(U_{g,1}, \dots, U_{g,n}\big)^\top$;
- Residual sum of squares: $\mathsf{SS}_{e,g} = \big\|\boldsymbol{U}_g\big\|^2$.

Lemma 9.1 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.*

$$\begin{aligned}\widehat{\boldsymbol{Y}}_g &= \widehat{\boldsymbol{Y}} + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^{-}\mathbb{V}^\top\boldsymbol{U}\\ &= \mathbb{X}\boldsymbol{b}_g + \mathbb{V}\boldsymbol{c}_g, \qquad \textit{for some } \boldsymbol{b}_g \in \mathbb{R}^k,\ \boldsymbol{c}_g \in \mathbb{R}^l.\end{aligned}$$

Vectors $\boldsymbol{b}_g$ *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 "$\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top$":

$$\mathbb{H}_g = \begin{pmatrix}\mathbb{X}, & \mathbb{M}\mathbb{V}\end{pmatrix}\begin{pmatrix}\mathbb{X}^\top\mathbb{X} & \underbrace{\mathbb{X}^\top\mathbb{M}\mathbb{V}}_{\mathbf{0}}\\ \underbrace{\mathbb{V}^\top\mathbb{M}\mathbb{X}}_{\mathbf{0}} & \mathbb{V}^\top\mathbb{M}\mathbb{V}\end{pmatrix}^{-}\begin{pmatrix}\mathbb{X}^\top\\ \mathbb{V}^\top\mathbb{M}\end{pmatrix}$$

$$= \begin{pmatrix}\mathbb{X}, & \mathbb{M}\mathbb{V}\end{pmatrix}\begin{pmatrix}\big(\mathbb{X}^\top\mathbb{X}\big)^{-} & \mathbf{0}\\ \mathbf{0} & \big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^{-}\end{pmatrix}\begin{pmatrix}\mathbb{X}^\top\\ \mathbb{V}^\top\mathbb{M}\end{pmatrix}$$

$$= \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^{-}\mathbb{V}^\top\mathbb{M}.$$

- So that,

$$\widehat{\boldsymbol{Y}}_g = \mathbb{H}_g\boldsymbol{Y} = \underbrace{\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top\boldsymbol{Y}}_{\widehat{\boldsymbol{Y}}} + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^{-}\mathbb{V}^\top\underbrace{\mathbb{M}\boldsymbol{Y}}_{\boldsymbol{U}}$$

$$= \widehat{\boldsymbol{Y}} + \mathbb{M}\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}\mathbb{V}\big)^{-}\mathbb{V}^\top\boldsymbol{U} \qquad 🚲$$

- The fitted values $\widehat{\boldsymbol{Y}}_g$ must lie in the corresponding regression space $\mathcal{M}\big(\mathbb{X},\,\mathbb{V}\big)$, that is, it must be possible to write the vector of fitted values as

$$\widehat{\boldsymbol{Y}}_g = \mathbb{X}\boldsymbol{b}_g + \mathbb{V}\boldsymbol{c}_g$$

  for some $\boldsymbol{b}_g \in \mathbb{R}^k$, $\boldsymbol{c}_g \in \mathbb{R}^l$. At the same time, the vector $\big(\boldsymbol{b}_g^\top,\,\boldsymbol{c}_g^\top\big)^\top$ must minimize the sum of squares SS of model M_g. As was shown in proof of Lemma 2.1, the vector $\big(\boldsymbol{b}_g^\top,\,\boldsymbol{c}_g^\top\big)^\top$ minimizes the sum of squares if and only if it solves corresponding normal equations.
- We rewrite 🚲 to see what $\boldsymbol{b}_g$ and $\boldsymbol{c}_g$ could be.

- Remember (Lemma 2.1) that $\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{b}$ for any $\boldsymbol{b} = \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-}\mathbb{X}^\top\boldsymbol{Y}$ (any solution of normal equations in model M). Take now 🚲 and further calculate:

$$\begin{aligned}
\widehat{\boldsymbol{Y}}_g &= \underbrace{\mathbb{X}\boldsymbol{b}}_{\widehat{\boldsymbol{Y}}} + \underbrace{\bigl\{\mathbf{I}_n - \mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-}\mathbb{X}^\top\bigr\}}_{\mathbb{M}}\mathbb{V}\bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U}\\
&= \mathbb{X}\boldsymbol{b} + \mathbb{V}\bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U} - \mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-}\mathbb{X}^\top\mathbb{V}\bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U}\\
&= \mathbb{X}\underbrace{\bigl\{\boldsymbol{b} - \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-}\mathbb{X}^\top\mathbb{V}\bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U}\bigr\}}_{\boldsymbol{b}_g} + \mathbb{V}\underbrace{\bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U}}_{\boldsymbol{c}_g}.
\end{aligned}$$

- That is, $\boldsymbol{c}_g = \bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U},$

  $\boldsymbol{b}_g = \boldsymbol{b} - \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-}\mathbb{X}^\top\mathbb{V}\boldsymbol{c}_g.$

- Finally

$$\mathsf{SS}_e - \mathsf{SS}_{e,g} = \bigl\|\widehat{\boldsymbol{Y}}_g - \widehat{\boldsymbol{Y}}\bigr\|^2 = \bigl\|\mathbb{M}\mathbb{V}\bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-}\mathbb{V}^\top\boldsymbol{U}\bigr\|^2 = \bigl\|\mathbb{M}\mathbb{V}\boldsymbol{c}_g\bigr\|^2.$$

❑

Note. If all model matrices are of a full column rank, i.e., if

$$\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k, \quad \mathsf{rank}\bigl(\mathbb{V}_{n\times l}\bigr) = l, \quad \mathsf{rank}\bigl(\mathbb{G}_{n\times(k+l)}\bigr) = k + l$$

then the least squares estimators $\widehat{\boldsymbol{\beta}}$ and $\bigl(\widehat{\boldsymbol{\beta}}_g^\top, \widehat{\boldsymbol{\gamma}}_g^\top\bigr)^\top$ of the vector of regression coefficients in models M and M$_g$ are mutually in the following relationship:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\boldsymbol{Y}, & \widehat{\boldsymbol{\gamma}}_g &= \bigl(\mathbb{V}^\top\mathbb{M}\mathbb{V}\bigr)^{-1}\mathbb{V}^\top\boldsymbol{U},\\
& & \widehat{\boldsymbol{\beta}}_g &= \widehat{\boldsymbol{\beta}} - \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_g.
\end{aligned}$$

9.2 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,

$$\mathbb{E}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}\big), \qquad \mathbb{E}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big) = \mathbb{X}\boldsymbol{\beta} \quad \text{for some } \boldsymbol{\beta} \in \mathbb{R}^k,$$

$$\mathbb{E}\big(\boldsymbol{\varepsilon}\,\big|\,\mathbb{Z}\big) = \mathbf{0}_n \qquad \big(\Longrightarrow \quad \mathbb{E}\big(\boldsymbol{\varepsilon}\big) = \mathbf{0}_n\big).$$

As (also) explained in Section 3.1, 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 full rank k with intercept, that is,

$$\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k < n, \qquad \mathbb{X} = \Big(\boldsymbol{X}^0,\, \ldots,\, \boldsymbol{X}^{k-1}\Big),\ \boldsymbol{X}^0 = \mathbf{1}_n.$$

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).

9.2.1 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

$$\mathsf{M}^{(-j)}:\quad \boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \big(\mathbb{X}^{(-j)}\boldsymbol{\beta}^{(-j)},\, \sigma^2\mathbf{I}_n\big).$$

Note. We are assuming a full rank model, i.e., $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$. This implies that also the matrix $\mathbb{X}^{(-j)}$ is of a full column rank, i.e., $\mathsf{rank}\big(\mathbb{X}^{(-j)}\big) = k-1$. This further implies, among the other things that

(i) $\boldsymbol{X}^j \notin \mathcal{M}\big(\mathbb{X}^{(-j)}\big)$;

(ii) $\boldsymbol{X}^j \neq \mathbf{0}_n$;

(iii) $\boldsymbol{X}^j$ is not a multiple of a vector $\mathbf{1}_n$.

All quantities related to the model $\mathsf{M}^{(-j)}$ will be indicated by a superscript $(-j)$. In particular,

$$\mathbb{M}^{(-j)} = \mathbf{I}_n - \mathbb{X}^{(-j)} \left(\mathbb{X}^{(-j)\top}\mathbb{X}^{(-j)}\right)^{-1} \mathbb{X}^{(-j)\top}$$

is a projection matrix into the residual space $\mathcal{M}\big(\mathbb{X}^{(-j)}\big)^{\perp}$;

$$\boldsymbol{U}^{(-j)} = \mathbb{M}^{(-j)}\boldsymbol{Y}$$

is a vector of residuals of the model $\mathsf{M}^{(-j)}$.

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). At the same time, both models are of a full rank given our assumptions. Let

$$\widehat{\boldsymbol{\beta}} = \big(\widehat{\beta}_0, \ldots, \widehat{\beta}_{j-1}, \widehat{\beta}_j, \widehat{\beta}_{j+1}, \ldots, \widehat{\beta}_{k-1}\big)^\top$$

be the least squares estimator of the regression coefficients in model M. Lemma 9.1 (Model with added regressors) provides

$$\widehat{\beta}_j = \big(\boldsymbol{X}^{j\top}\mathbb{M}^{(-j)}\boldsymbol{X}^j\big)^{-1}\boldsymbol{X}^{j\top}\boldsymbol{U}^{(-j)}. \tag{9.1}$$

Further, since a matrix $\mathbb{M}^{(-j)}$ is idempotent and symmetric, we have

$$\boldsymbol{X}^{j\top}\mathbb{M}^{(-j)}\boldsymbol{X}^j = \big\|\mathbb{M}^{(-j)}\boldsymbol{X}^j\big\|^2.$$

At the same time, $\mathbb{M}^{(-j)}\boldsymbol{X}^j \neq \mathbf{0}_n$ since $\boldsymbol{X}^j \notin \mathcal{M}\big(\mathbb{X}^{(-j)}\big)$, $\boldsymbol{X}^j \neq \mathbf{0}_n$. Hence, $\boldsymbol{X}^{j\top}\mathbb{M}^{(-j)}\boldsymbol{X}^j > 0$ and we can indeed calculate its inverse used in expression (9.1). That is, the jth element of the LSE of the vector $\boldsymbol{\beta}$ in model M (= BLUE of β_j) is given as

$$\widehat{\beta}_j = \big(\boldsymbol{X}^{j\top}\mathbb{M}^{(-j)}\boldsymbol{X}^j\big)^{-1}\boldsymbol{X}^{j\top}\boldsymbol{U}^{(-j)} = \frac{\boldsymbol{X}^{j\top}\boldsymbol{U}^{(-j)}}{\boldsymbol{X}^{j\top}\mathbb{M}^{(-j)}\boldsymbol{X}^j}.$$

Consequently, we define a vector of jth partial residuals of model M as follows.

Definition 9.1 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

$$\boldsymbol{U}^{part,j} = \boldsymbol{U} + \widehat{\beta}_j\boldsymbol{X}^j = \boldsymbol{Y} - \big(\mathbb{X}\widehat{\boldsymbol{\beta}} - \widehat{\beta}_j\boldsymbol{X}^j\big) = \boldsymbol{Y} - \big(\widehat{\boldsymbol{Y}} - \widehat{\beta}_j\boldsymbol{X}^j\big).$$

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.

[1] *vektor $j$tých parciálních reziduí*

Lemma 9.2 Property of partial residuals.

Let $\boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k$, $\boldsymbol{X}^0 = \mathbf{1}_n$, $\boldsymbol{\beta} = \big(\beta_0,\, \dots,\, \beta_{k-1}\big)^\top$. Let $\widehat{\beta}_j$ be the LSE of β_j, $j \in \big\{1,\, \dots, k-1\big\}$. 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 $\Big\|\boldsymbol{U}^{part,j} - \gamma_{j,0}\,\mathbf{1}_n - \gamma_{j,1}\,\boldsymbol{X}^j\Big\|^2 = \Big\|\boldsymbol{U} - \big\{\gamma_{j,0}\,\mathbf{1}_n + (\gamma_{j,1} - \widehat{\beta}_j)\boldsymbol{X}^j\big\}\Big\|^2 = \text{🚲}$.
- 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

$$\text{🚲} = \big\|\boldsymbol{U}\big\|^2 + \Big\|\gamma_{j,0}\,\mathbf{1}_n + \big(\gamma_{j,1} - \widehat{\beta}_j\big)\,\boldsymbol{X}^j\Big\|^2 \geq \big\|\boldsymbol{U}\big\|^2$$

with equality if and only if $\gamma_{j,0} = 0 \quad \& \quad \gamma_{j,1} = \widehat{\beta}_j$.

❑

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 9.2 Shifted partial residuals.

A vector of jth response-mean *partial residuals of model* M *is a vector*

$$\boldsymbol{U}^{part,j,Y} = \boldsymbol{U}^{part,j} + \left(\overline{Y} - \widehat{\beta}_j\overline{X}^j\right)\mathbf{1}_n.$$

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 $\boldsymbol{Y}$;
- a *partial* effect of the jth regressor on the response $\boldsymbol{Y}$ 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 $\left(\boldsymbol{X}^j, \boldsymbol{U}^{part,j}\right)$ 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 $\left(\boldsymbol{X}^j, \boldsymbol{U}^{part,j}\right)$ 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 $\left(\boldsymbol{X}^j, \boldsymbol{U}^{part,j}\right)$ equipped by a line with zero intercept and slope equal to $\widehat{\beta}_j$.

Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**

`consumption` $\sim$ `log(weight)` $+$ `engine size` $+$ `horsepower`

```
m <- lm(consumption ~ lweight + engine.size + horsepower, data = CarsNow)
summary(m)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1174 -0.6923 -0.1127  0.5473  5.2275

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.353265   2.948614 -14.364  < 2e-16 ***
lweight       6.935604   0.428971  16.168  < 2e-16 ***
engine.size   0.352687   0.096730   3.646 0.000301 ***
horsepower    0.003983   0.001085   3.672 0.000273 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9706 on 405 degrees of freedom
Multiple R-squared:  0.7946,	Adjusted R-squared:  0.793
F-statistic: 522.1 on 3 and 405 DF,  p-value: < 2.2e-16
```

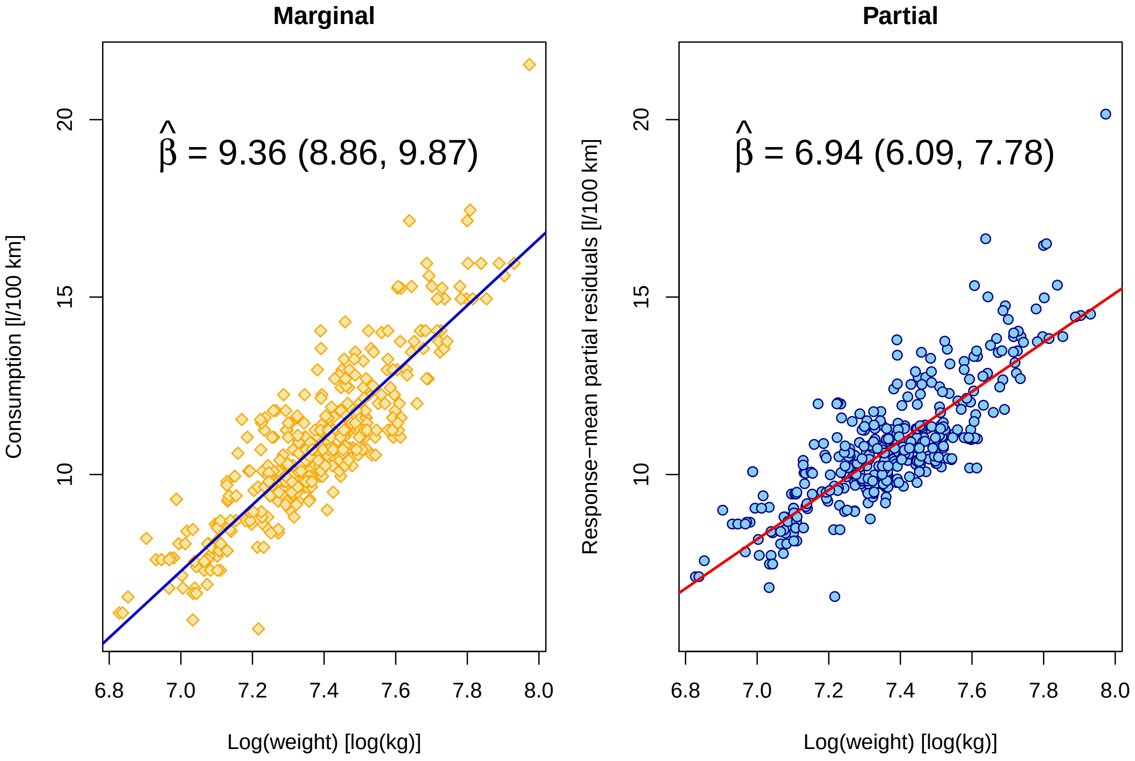
Consumption: $\overline{Y} = 10.75$, Log(weight): $\overline{X}^1 = 7.37$,

Engine size: $\overline{X}^2 = 3.18$,

Horsepower: $\overline{X}^3 = 215.8$.

`Cars2004nh` **(subset,** $n = 409$**)**

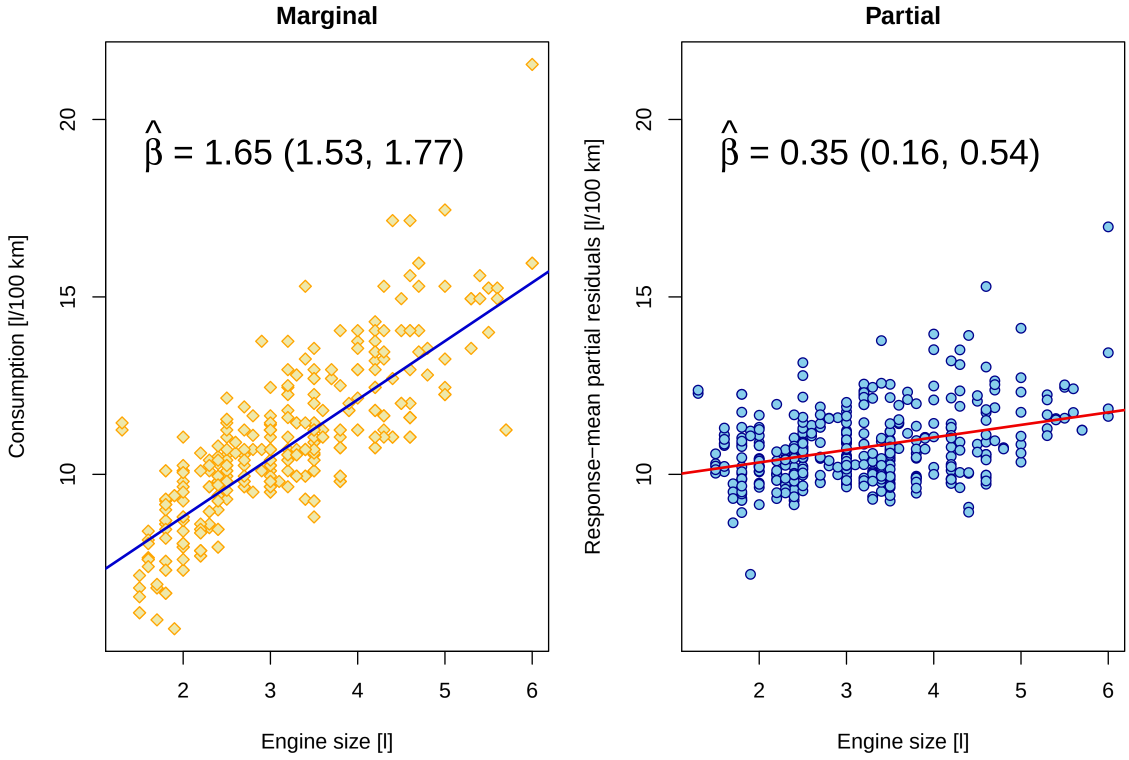
`consumption` $\sim$ `log(weight)` $+$ `engine size` $+$ `horsepower`



Illustrations

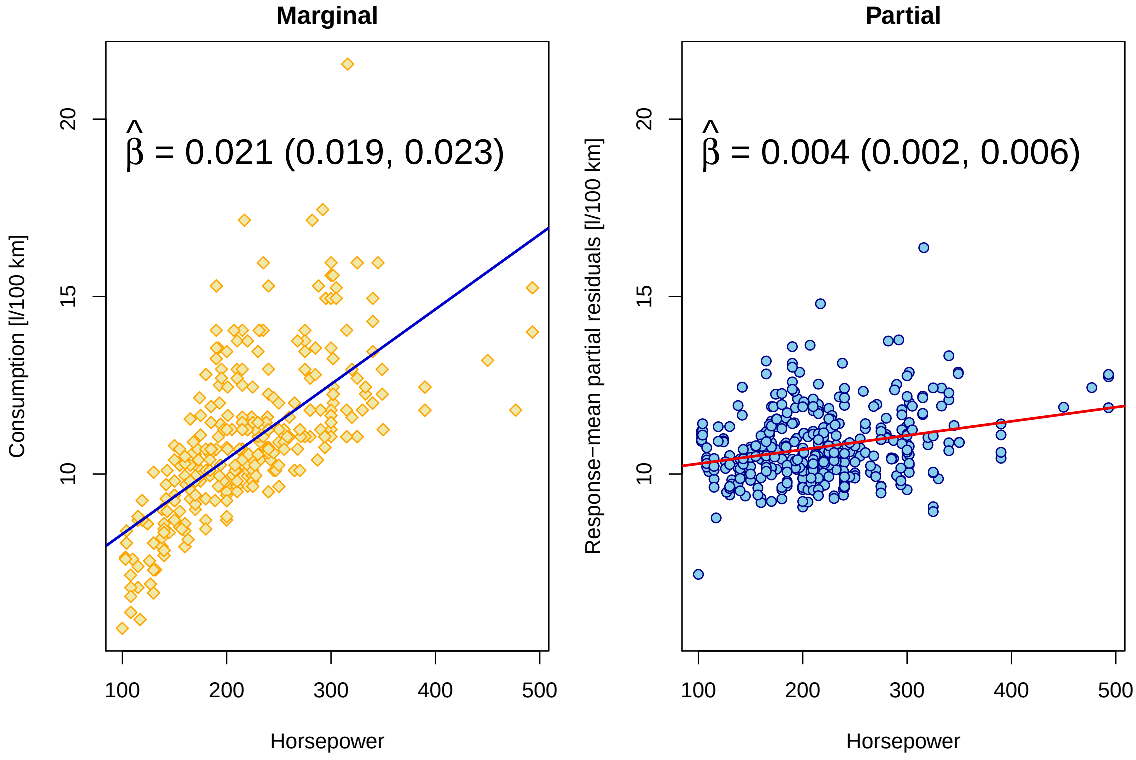
Cars2004nh (**subset,** $n = 409$)

consumption $\sim$ log(weight) + engine size + horsepower



Cars2004nh (**subset,** $n = 409$)

consumption $\sim$ log(weight) + engine size + horsepower



Illustrations

Policie ($n = 50$)

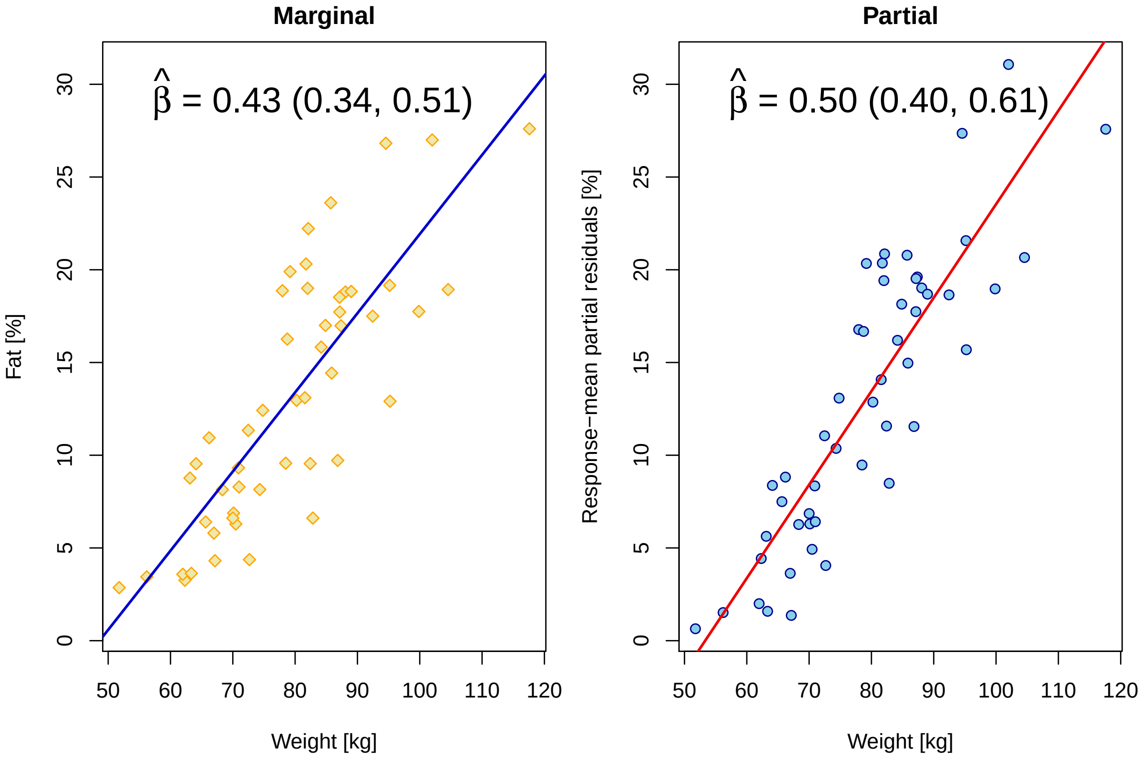
fat $\sim$ weight $+$ height

```
summary(mHeWe <- lm(fat ~ weight + height, data = Policie))
```

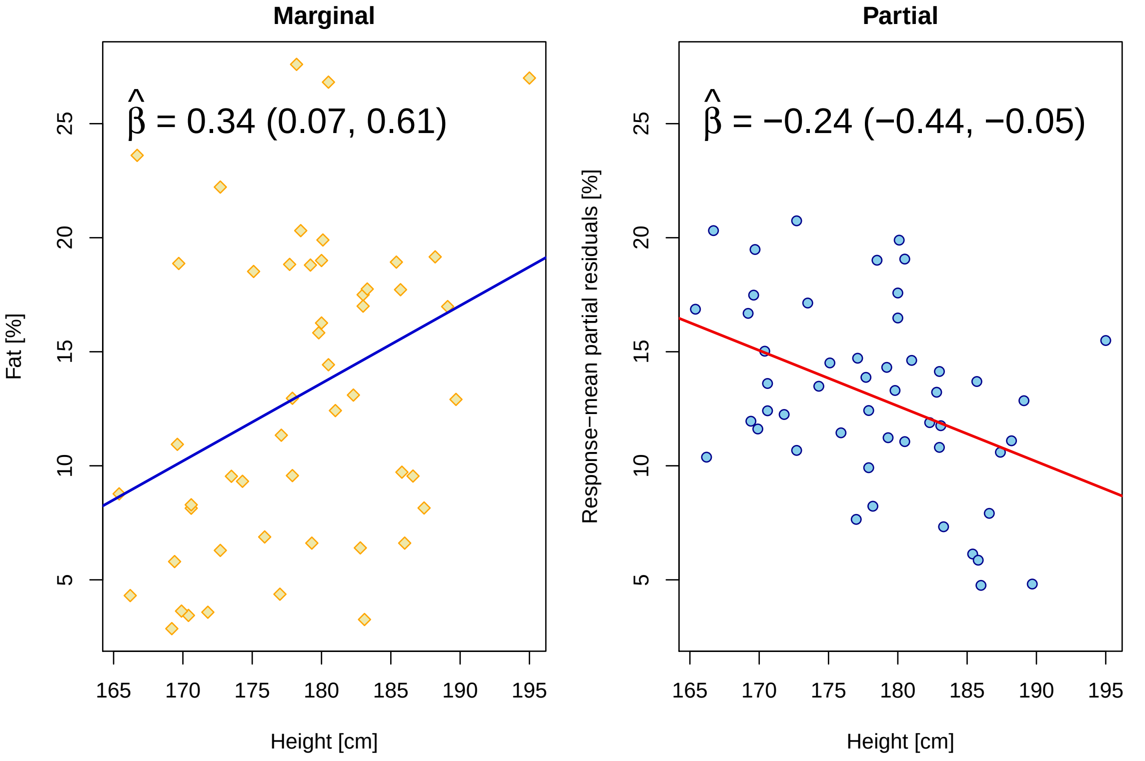
```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.4011 -2.9482 -0.0211  2.3072  7.2968 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.55309   15.24621   1.086   0.2831    
weight       0.50418    0.05095   9.896 4.49e-13 ***
height      -0.24362    0.09728  -2.504   0.0158 *  
---

Residual standard error: 3.731 on 47 degrees of freedom
Multiple R-squared:  0.714,	Adjusted R-squared:  0.7018 
F-statistic: 58.66 on 2 and 47 DF,  p-value: 1.681e-13
```

Illustrations

`Policie` ($n = 50$)
`fat` $\sim$ `weight` $+$ `height`



`Policie` ($n = 50$)
`fat` $\sim$ `weight` $+$ `height`



9.2.2 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 $\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z}\bigr)$ 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

$$\mathbb{X} = \Bigl(\mathbf{1}_n,\, \mathbb{X}^0,\, \boldsymbol{X}^j\Bigr).$$

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

$$\boldsymbol{X}^j \in \mathcal{M}\bigl(\mathbb{V}\bigr), \qquad \mathsf{rank}\bigl(\mathbb{V}\bigr) \geq 2.$$

That is, the model matrices of the submodel and the (larger) model are

$$\begin{array}{ll} \text{Submodel } \mathsf{M}: & \Bigl(\mathbf{1}_n,\, \mathbb{X}^0,\, \boldsymbol{X}^j\Bigr) = \mathbb{X}; \\ \text{(Larger) model } \mathsf{M}_g: & \Bigl(\mathbf{1}_n,\, \mathbb{X}^0,\, \mathbb{V}\Bigr). \end{array}$$

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

$$\Bigl(\mathbb{X}^0,\, \mathbb{V}\Bigr)$$

that does not explicitely include the intercept term which is included implicitely.

Illustrations

Cars2004nh **(subset,** $n = 409$**)**

`consumption` ∼ `log(weight) + engine.size + horsepower`

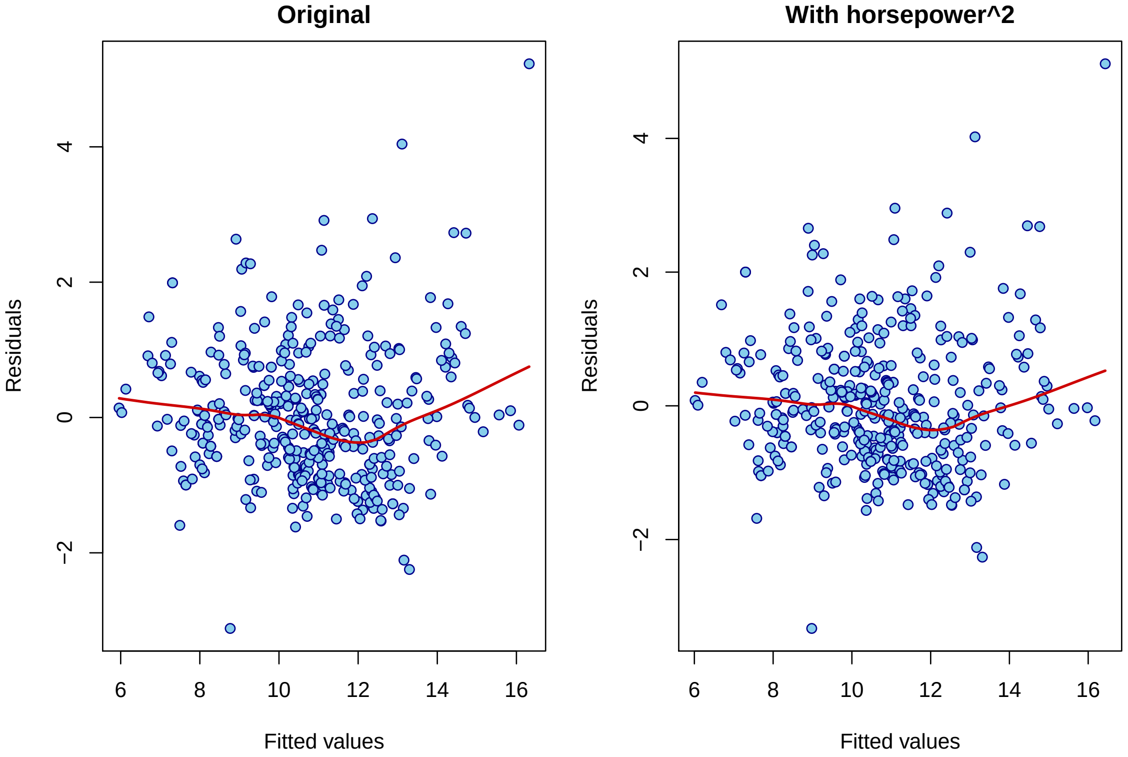
Quadratic term added for `horsepower`

```
mh2 <- lm(consumption ~ lweight + engine.size + horsepower + I(horsepower^2),
          data = CarsNow)
summary(mh2)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3298 -0.6501 -0.1307  0.5178  5.1163

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.386e+01  3.065e+00 -14.308  < 2e-16 ***
lweight          7.249e+00  4.641e-01  15.621  < 2e-16 ***
engine.size      3.482e-01  9.652e-02   3.607 0.000348 ***
horsepower      -2.578e-03  3.914e-03  -0.659 0.510515
I(horsepower^2)  1.221e-05  7.001e-06   1.744 0.081873 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9682 on 404 degrees of freedom
Multiple R-squared:  0.7961,        Adjusted R-squared:  0.7941
F-statistic: 394.3 on 4 and 404 DF,  p-value: < 2.2e-16
```

Cars2004nh **(subset,** $n = 409$**)**

`consumption` ∼ `log(weight)` + `engine size` + `horsepower`



Illustrations

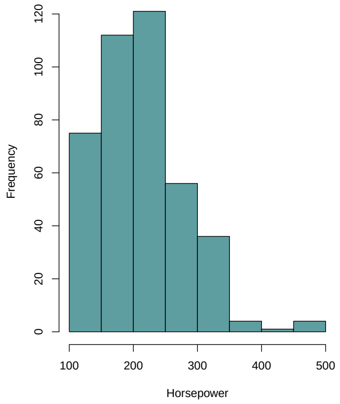
Cars2004nh (**subset,** $n = 409$)

`consumption` $\sim$ `log(weight) + engine.size + horsepower`

Cubic spline parameterization of `horsepower` (knots: 100, 200, 300, 500)

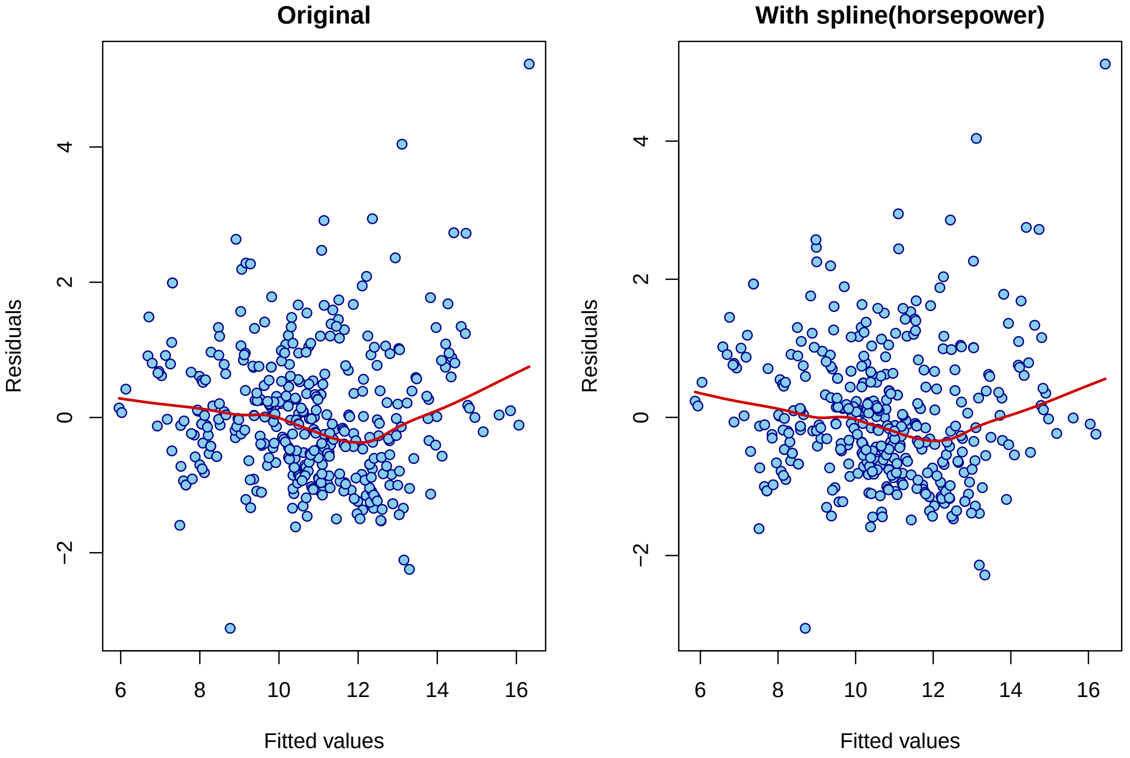
```
library("splines")
knots <- c(100, 200, 300, 500)
inner <- knots[-c(1, length(knots))]
bound <- knots[c(1, length(knots))]
hB <- bs(CarsNow[, "horsepower"], knots = inner, Boundary.knots = bound, degree = 3,
         intercept = TRUE)
mhB <- lm(consumption ~ -1 + lweight + engine.size + hB, data = CarsNow)
summary(mhB)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.0533 -0.6471 -0.1273  0.5095  5.1164

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
lweight       7.19154    0.48080  14.958  < 2e-16 ***
engine.size   0.36108    0.09911   3.643 0.000304 ***
hB1         -43.88205    3.25963 -13.462  < 2e-16 ***
hB2         -43.40426    3.32369 -13.059  < 2e-16 ***
hB3         -43.58750    3.39894 -12.824  < 2e-16 ***
hB4         -43.18531    3.38594 -12.754  < 2e-16 ***
hB5         -41.93832    3.43966 -12.193  < 2e-16 ***
hB6         -41.83870    3.37295 -12.404  < 2e-16 ***
...
```



Cars2004nh (**subset,** $n = 409$)

`consumption` $\sim$ `log(weight)` + `engine size` + `horsepower`



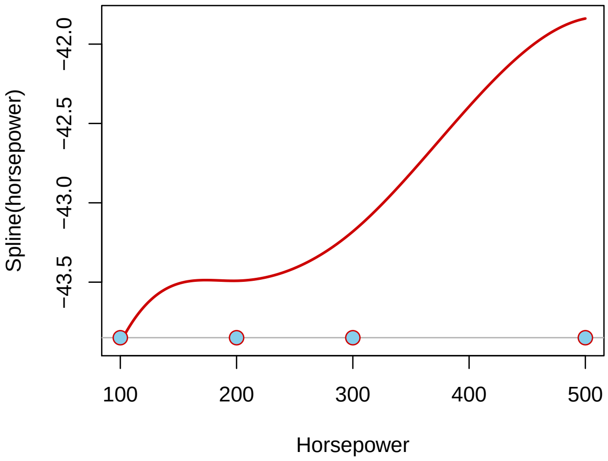
Illustrations

`Cars2004nh` **(subset,** $n = 409$**)**
`consumption` $\sim$ `log(weight) + engine.size + horsepower`

Cubic spline parameterization of `horsepower` (knots: 100, 200, 300, 500)

```
m <- lm(consumption ~ lweight +
           engine.size +
           horsepower,
           data = CarsNow)
anova(m, mhB)
```

```
Analysis of Variance Table

Model 1: consumption ~ lweight +
             engine.size + horsepower
Model 2: consumption ~ -1 + lweight +
             engine.size + hB

  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    405 381.56
2    401 377.08  4    4.4797 1.191 0.3142
```



Categorization of the jth regressor

Let $-\infty < x_j^{low} < x_j^{upp} < \infty$ be chosen such that interval $\bigl(x_j^{low},\, x_j^{upp}\bigr)$ covers the values $X_{1,j},\,\dots,\,X_{n,j}$ of the jth regressor. That is,

$$x_j^{low} < \min_i X_{i,j}, \qquad \max_i X_{i,j} < x_j^{upp}.$$

Let $\mathcal{I}_1,\,\dots,\,\mathcal{I}_H$ be $H > 1$ subintervals of $\bigl(x_j^{low},\, x_j^{upp}\bigr]$ based on a grid

$$x_j^{low} < \lambda_1 < \cdots < \lambda_{H-1} < x_j^{upp}.$$

Let $x_h \in \mathcal{I}_h$, $h = 1,\dots,H$, be chosen representative values for each of the subintervals $\mathcal{I}_1,\dots,\mathcal{I}_H$ (e.g., their midpoints) and let

$$\boldsymbol{X}^{j,cut} = \bigl(X_1^{j,cut},\,\dots,\,X_n^{j,cut}\bigr)^\top$$

be obtained by categorization of the jth regressor using the division $\mathcal{I}_1,\,\dots,\,\mathcal{I}_H$ and representatives $x_1,\,\dots,\,x_H$, i.e., ($i = 1,\dots,n$):

$$X_i^{j,cut} = x_h \quad \equiv \quad X_i^j \in \mathcal{I}_h, \qquad h = 1,\dots,H.$$

In this way, we obtained a categorical ordinal regressor $\boldsymbol{X}^{j,cut}$ whose values $x_1,\,\dots,\,x_H$, 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 $\mathcal{I}_1,\,\dots,\,\mathcal{I}_H$ and the representatives $x_1,\,\dots,\,x_H$) 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

$$\boldsymbol{X}^{j,cut} \in \mathcal{M}\bigl(\mathbb{V}\bigr),$$

and test for linearity of the jth regressor is obtained by considering the following model matrices in the submodel and the (larger) model:

$$\text{Submodel M:} \qquad \Bigl(\mathbf{1}_n,\, \mathbb{X}^0,\, \boldsymbol{X}^{j,cut}\Bigr);$$

$$\text{(Larger) model M}_g\text{:} \qquad \Bigl(\mathbf{1}_n,\, \mathbb{X}^0,\, \mathbb{V}\Bigr).$$

Additional insight concerning the correct inclusion of the jth regressor can be obtained by using the orthonormal polynomial contrasts (Section 4.4.3) in place of the $\mathbb{V}$ matrix.

Illustrations

Cars2004nh (subset, $n = 409$)

consumption ∼ log(weight) + engine.size + horsepower

Categorized horsepower (100–150, 150–200, 250–300, 300–500)

```
BREAKS <- c(0, 150, 200, 250, 300, 500)
CarsNow <- transform(CarsNow, horseord = cut(horsepower, breaks = BREAKS))
levels(CarsNow[, "horseord"])[1] <- "[100, 150]"
table(CarsNow[, "horseord"])
```

```
[100, 150]  (150,200]  (200,250]  (250,300]  (300,500]
        75        112        121         56         45
```

horsepower categories represented by **midpoints**

```
MIDS <- c(125, 175, 225, 275, 400)
CarsNow <- transform(CarsNow, horsemid = as.numeric(horseord))
CarsNow[, "horsemid"] <- MIDS[CarsNow[, "horsemid"]]
table(CarsNow[, "horsemid"])
```

```
125 175 225 275 400
 75 112 121  56  45
```

Cars2004nh (subset, $n = 409$)

consumption ∼ log(weight) + engine.size + horsepower

Larger model (horsepower as categorical, reference group pseudocontrasts)

```
mhord <- lm(consumption ~ lweight + engine.size + horseord, data = CarsNow)
summary(mhord)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -43.4282     3.1974 -13.582  < 2e-16 ***
lweight             7.1578     0.4676  15.307  < 2e-16 ***
engine.size         0.3312     0.0981   3.376 0.000806 ***
horseord(150,200]   0.3928     0.1637   2.400 0.016852 *
horseord(200,250]   0.2206     0.1832   1.204 0.229119
horseord(250,300]   0.5249     0.2338   2.245 0.025332 *
horseord(300,500]   1.0871     0.2626   4.140 4.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9628 on 402 degrees of freedom
Multiple R-squared:  0.7994,        Adjusted R-squared:  0.7964
F-statistic:   267 on 6 and 402 DF,  p-value: < 2.2e-16
```

Illustrations

Cars2004nh (subset, $n = 409$)

consumption $\sim$ log(weight) + engine.size + horsepower

Submodel (horsepower intervals represented by midpoints)

```
mhmid <- lm(consumption ~ lweight + engine.size + horsemid, data = CarsNow)
summary(mhmid)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -43.121394   2.944142 -14.647  < 2e-16 ***
lweight          7.057884   0.427803  16.498  < 2e-16 ***
engine.size      0.338626   0.096994   3.491 0.000534 ***
horsemid         0.003519   0.009049   3.889 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9687 on 405 degrees of freedom
Multiple R-squared:  0.7954,	Adjusted R-squared:  0.7938
F-statistic: 524.7 on 3 and 405 DF,  p-value: < 2.2e-16
```

F-test on a submodel

```
anova(mhmid, mhord)
```

```
Model 1: consumption ~ lweight + engine.size + horsemid
Model 2: consumption ~ lweight + engine.size + horseord
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    405 380.07
2    402 372.61  3    7.4566 2.6816 0.04653 *
```

Illustrations

Cars2004nh (subset, $n = 409$)
consumption ∼ log(weight) + engine.size + horsepower

Approximate submodel (original horsepower values)

```
m <- lm(consumption ~ lweight + engine.size + horsepower, data = CarsNow)
summary(m)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.353265   2.948614 -14.364  < 2e-16 ***
lweight       6.935604   0.428971  16.168  < 2e-16 ***
engine.size   0.352687   0.096730   3.646 0.000301 ***
horsepower    0.003983   0.001085   3.672 0.000273 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9706 on 405 degrees of freedom
Multiple R-squared:  0.7946,        Adjusted R-squared:  0.793
F-statistic: 522.1 on 3 and 405 DF,  p-value: < 2.2e-16
```

Approximate F-test on a submodel

```
anova(m, mhord)
```

```
Model 1: consumption ~ lweight + engine.size + horsepower
Model 2: consumption ~ lweight + engine.size + horseord
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1    405 381.56
2    402 372.61  3    8.9427 3.216 0.02285 *
```

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.

9.3 Homoscedasticity

We are again 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):

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

The assumption (A2) of homoscedasticity is, in particular,

$$\mathsf{var}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z}\big) = \sigma^2\,\mathbf{I}_n, \qquad \mathsf{var}\big(\boldsymbol{\varepsilon}\,\big|\,\mathbb{Z}\big) = \sigma^2\,\mathbf{I}_n, \qquad \big(\Longrightarrow \;\mathsf{var}\big(\boldsymbol{\varepsilon}\big) = \sigma^2\mathbf{I}_n\big),$$

where σ^2 is unknown but most importantly constant.

9.3.1 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.

9.3.2 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.,

$$\mathsf{M}:\quad \boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big), \quad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k,$$

and an alternative model is a generalization of a general normal linear model

$$\mathsf{M}_{hetero}:\quad \boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\big),$$

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

$$\begin{aligned} \mathrm{H}_0: &\quad \boldsymbol{\lambda} = \mathbf{0}, \\ \mathrm{H}_1: &\quad \boldsymbol{\lambda} \neq \mathbf{0}. \end{aligned} \tag{9.2}$$

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 (9.2). 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 ($\boldsymbol{x} = \boldsymbol{t}_X(\boldsymbol{z})$ 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 $i = 1, \ldots, n$,

$$\mathsf{var}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr) = \mathsf{var}\bigl(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\bigr) = \sigma^2 \exp\bigl(\lambda\, \boldsymbol{X}_i^\top \boldsymbol{\beta}\bigr) = \sigma^2 \exp\Bigl(\lambda\, \mathbb{E}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr)\Bigr), \tag{9.3}$$

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 (9.3) 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.

Linear dependence on the regressors

Let $\boldsymbol{t}_W : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ be a given transformation, $\boldsymbol{w} := \boldsymbol{t}_W(\boldsymbol{z})$, $\boldsymbol{W}_i = t_W(\boldsymbol{Z}_i)$, $i = 1, \ldots, n$. 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 $i = 1, \ldots, n$,

$$\mathsf{var}\bigl(Y_i \,\big|\, \boldsymbol{Z}_i\bigr) \;=\; \mathsf{var}\bigl(\varepsilon_i \,\big|\, \boldsymbol{Z}_i\bigr) \;=\; \sigma^2 \exp\bigl(\boldsymbol{\lambda}^\top \boldsymbol{W}_i\bigr).$$

[2] *vnořený*

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.

9.3.3 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`.

9.4 Normality

In this section, we are assuming a *normal* linear model

$$\mathsf{M}: \ \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbf{I}_n\bigr), \ \mathsf{rank}(\mathbb{X}_{n\times k}) = r \leq k,$$

where the error terms $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} = \bigl(\varepsilon_1, \ldots, \varepsilon_n\bigr)^\top$ satisfy (Lemma 6.1):

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma^2), \quad i = 1, \ldots, n. \tag{9.4}$$

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 $\mathcal{M}(\mathbb{X})$):

$$\mathbb{H} = \mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-}\mathbb{X}^\top = \bigl(h_{i,t}\bigr)_{i,t=1,\ldots,n};$$

(ii) Projection matrix into the residual space $\mathcal{M}(\mathbb{X})^\perp$:

$$\mathbb{M} = \mathbf{I}_n - \mathbb{H} = \bigl(m_{i,t}\bigr)_{i,t=1,\ldots,n};$$

(iii) Residuals: $\boldsymbol{U} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \mathbb{M}\boldsymbol{Y} = \bigl(U_1, \ldots, U_n\bigr)^\top$;

(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: $\boldsymbol{U}^{std} = \bigl(U_1^{std}, \ldots, U_n^{std}\bigr)^\top$, 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 (9.4) holds then Lemma 3.1 and Theorem 6.2 provide:

(i) For (raw) residuals:

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

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):

$$\mathbb{E}\bigl(U_i^{std} \,\big|\, \mathbb{Z}\bigr) = 0, \qquad \mathsf{var}\bigl(U_i^{std} \,\big|\, \mathbb{Z}\bigr) = 1, \qquad i = 1, \ldots, n.$$

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.

9.4.1 Tests of normality

There exist formal tests of the null hypothesis on a normality of the error terms:

$$\mathrm{H}_0: \ \text{distribution of } \varepsilon_1, \ldots, \varepsilon_n \text{ is normal}, \tag{9.5}$$

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 (9.5), 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`.

9.5 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,

$$\mathsf{cov}\big(\varepsilon_i,\, \varepsilon_l \,\big|\, \mathbb{X}\big) = 0,\ i \neq l \qquad \big(\Longrightarrow \quad \mathsf{cov}\big(\varepsilon_i,\, \varepsilon_l\big) = 0,\ i \neq l\big). \tag{9.6}$$

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.

9.5.1 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

$$\begin{aligned} \mathsf{M}: \quad & Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, & & i = 1, \ldots, n, \\ & \mathbb{E}\big(\varepsilon_i \,\big|\, \mathbb{X}\big) = 0, \quad \mathsf{var}\big(\varepsilon_i \,\big|\, \mathbb{X}\big) = \sigma^2, & & i = 1, \ldots, n, \\ & \mathsf{cor}\big(\varepsilon_i,\, \varepsilon_l \,\big|\, \mathbb{X}\big) = 0, & & i \neq l. \end{aligned} \tag{9.7}$$

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 (9.7) leads to a more general model

$$\begin{aligned} \mathsf{M}_{AR}: \quad & Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, & & i = 1, \ldots, n, \\ & \varepsilon_1 = \eta_1, \quad \varepsilon_i = \varrho\, \varepsilon_{i-1} + \eta_i, & & i = 2, \ldots, n, \\ & \mathbb{E}\big(\eta_i \,\big|\, \mathbb{X}\big) = 0, \quad \mathsf{var}\big(\eta_i \,\big|\, \mathbb{X}\big) = \sigma^2, & & i = 1, \ldots, n, \\ & \mathsf{cor}\big(\eta_i,\, \eta_l \,\big|\, \mathbb{X}\big) = 0, & & i \neq l, \end{aligned} \tag{9.8}$$

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

$$\begin{aligned} \mathrm{H}_0\colon &\quad \varrho = 0, \\ \mathrm{H}_1\colon &\quad \varrho \neq 0 \end{aligned}$$

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

$$DW = \frac{\displaystyle\sum_{i=2}^{n}(U_i - U_{i-1})^2}{\displaystyle\sum_{i=1}^{n} U_i^2}.$$

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 Lemma 2.7).

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 $\mathrm{H}_0: \ \varrho = 0$ 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)*.

9.6 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

$$\begin{aligned} \boldsymbol{Y}^\star \,\big|\, \mathbb{X} &\sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big), \\ \boldsymbol{Y}^\star &= \big(t(Y_1),\,\ldots,\,t(Y_n)\big)^\top, \end{aligned} \tag{9.9}$$

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.

9.6.1 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 (9.9) 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 $\widehat{Y}_{new}^\star$ and a $(1-\alpha)\,100\%$ prediction interval $\big(\widehat{Y}_{new}^{\star,L},\;\widehat{Y}_{new}^{\star,U}\big)$ for $Y_{new}^\star = t(Y_{new})$ based on the model (9.9).
- Trivially, an interval
$$\big(\widehat{Y}_{new}^{L},\;\widehat{Y}_{new}^{U}\big) = \Big(t^{-1}\big(\widehat{Y}_{new}^{\star,L}\big),\;t^{-1}\big(\widehat{Y}_{new}^{\star,U}\big)\Big) \tag{9.10}$$
covers a value of Y_{new} with a probability of $1-\alpha$.
- A value $\widehat{Y}_{new} = t^{-1}\big(\widehat{Y}_{new}^\star\big)$ lies inside the prediction interval (9.10) and can be considered as a point prediction of Y_{new}. Only note that the prediction interval $\big(\widehat{Y}_{new}^{L},\;\widehat{Y}_{new}^{U}\big)$ is not necessarily centered around a value of $\widehat{Y}_{new}$.

9.6.2 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:

$$\begin{aligned} \log(Y_i) = \boldsymbol{X}_i^\top\boldsymbol{\beta} + \varepsilon_i, &\qquad i = 1,\ldots,n, \\ \varepsilon_i \,\big|\, \mathbb{X} &\overset{\text{indep.}}{\sim} \mathcal{N}\big(0,\,\sigma^2\big), \end{aligned} \tag{9.11}$$

which also implies $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\,\sigma^2\big)$. We then have

$$\begin{aligned} Y_i = \exp\big(\boldsymbol{X}_i^\top\boldsymbol{\beta}\big)\,\eta_i, &\qquad i = 1,\ldots,n, \\ \eta_i \,\big|\, \mathbb{X} &\overset{\text{indep.}}{\sim} \mathcal{LN}\big(0,\,\sigma^2\big), \end{aligned}$$

which also implies $\eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{LN}\bigl(0,\,\sigma^2\bigr)$, 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 (9.11) 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 $i = 1,\ldots,n$),

$$\begin{aligned} M &:= \mathbb{E}\bigl(\eta_i\bigr) = \mathbb{E}\bigl(\eta_i \bigm| \mathbb{X}\bigr) = \exp\Bigl(\frac{\sigma^2}{2}\Bigr) \qquad\qquad > 1 \ \ (\text{with } \sigma^2 > 0),\\ V &:= \mathsf{var}\bigl(\eta_i\bigr) = \mathsf{var}\bigl(\eta_i \bigm| \mathbb{X}\bigr) = \bigl\{\exp(\sigma^2) - 1\bigr\}\exp(\sigma^2). \end{aligned}$$

Hence, for $\boldsymbol{x} \in \mathcal{X}$:

$$\begin{aligned} \mathbb{E}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr) &= M\,\exp\bigl(\boldsymbol{x}^\top\boldsymbol{\beta}\bigr),\\ \mathsf{var}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr) &= V\,\exp\bigl(2\,\boldsymbol{x}^\top\boldsymbol{\beta}\bigr) = V\cdot\biggl(\frac{\mathbb{E}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr)}{M}\biggr)^2. \end{aligned} \tag{9.12}$$

A log-normal model (9.11) 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 $\mathsf{var}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr)$ increases with a conditional expectation $\mathbb{E}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr)$. This feature is captured by the log-normal model as shown by (9.12). Indeed, under the log-normal model, $\mathsf{var}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr)$ increases with $\mathbb{E}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr)$. It is then said that the logarithmic transformation *stabilizes* the variance.

Interpretation of regression coefficients

With a log-normal model (9.11), the (non-intercept) regression coefficients have the following interpretation. Let for $j \in \{1,\,\ldots,\,k-1\}$,

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

and suppose that $\boldsymbol{\beta} = \bigl(\beta_0,\,\ldots,\,\beta_{k-1}\bigr)^\top$ We then have

$$\frac{\mathbb{E}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}^{j(+1)}\bigr)}{\mathbb{E}\bigl(Y \bigm| \boldsymbol{X} = \boldsymbol{x}\bigr)} = \frac{M\,\exp\bigl({\boldsymbol{x}^{j(+1)}}^\top\boldsymbol{\beta}\bigr)}{M\,\exp\bigl(\boldsymbol{x}^\top\boldsymbol{\beta}\bigr)} = \exp(\beta_j).$$

Notes.

- If a linear model with only a single categorical covariate and 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. In particular, let $Z \in \bigl\{1,\,\ldots,\,G\bigr\}$ be a categorical covariate which group means of the log-response are parameterized as

$$\mathbb{E}\bigl(\log(Y) \bigm| Z = g\bigr) = \beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta}^Z, \quad g = 1,\,\ldots,\,G,$$

where $\boldsymbol{c}_1^\top,\,\ldots,\,\boldsymbol{c}_G^\top$ are rows of the (pseudo)contrast matrix. If normality of the log-transformed response is assumed, we get (as above):

$$\begin{aligned} \frac{\mathbb{E}\bigl(Y \bigm| Z = g\bigr)}{\mathbb{E}\bigl(Y \bigm| Z = h\bigr)} &= \frac{M\,\exp\bigl(\beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta}^Z\bigr)}{M\,\exp\bigl(\beta_0 + \boldsymbol{c}_h^\top\boldsymbol{\beta}^Z\bigr)} = \exp\Bigl\{\bigl(\boldsymbol{c}_g^\top - \boldsymbol{c}_h^\top\bigr)\boldsymbol{\beta}^Z\Bigr\}\\ &= \exp\Bigl\{\mathbb{E}\bigl(\log(Y) \bigm| Z = g\bigr) - \mathbb{E}\bigl(\log(Y) \bigm| Z = h\bigr)\Bigr\}, \qquad g \neq h. \end{aligned}$$

- If a linear model with logarithmically transformed response if fitted, estimated regression coefficients, estimates of linear combinations 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 10

Consequences of a Problematic Regression Space

As in Chapter 9, we assume that data are represented by n random vectors $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $\boldsymbol{Z}_i = \bigl(Z_{i,1},\, \dots,\, Z_{i,p}\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$ $i = 1, \dots, n$. As usual, let $\boldsymbol{Y} = \bigl(Y_1, \dots, Y_n\bigr)^\top$ and let $\mathbb{Z}_{n\times p}$ denote a matrix with covariate vectors $\boldsymbol{Z}_1,\, \dots,\, \boldsymbol{Z}_n$ in its rows. Finally, let $\boldsymbol{X}_i$, $i = 1, \dots, n$, 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

$$\mathbb{X}_{n\times k} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \Bigl(\boldsymbol{X}^0,\, \dots,\, \boldsymbol{X}^{k-1}\Bigr).$$

It will be assumed that $\boldsymbol{X}^0 = \bigl(1,\, \dots,\, 1\bigr)^\top$ (almost surely) leading to the model matrix

$$\mathbb{X}_{n\times k} = \Bigl(\mathbf{1}_n,\, \boldsymbol{X}^1,\, \dots,\, \boldsymbol{X}^{k-1}\Bigr),$$

with explicitely included intercept term.

Primarily, we will assume that the model matrix $\mathbb{X}$ is sufficient to be able to assume that $\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}\bigr) = \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{X}\bigr) = \mathbb{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} = \bigl(\beta_0,\, \dots,\, \beta_{k-1}\bigr)^\top \in \mathbb{R}^k$. That is, we will arrive from assuming

$$\boldsymbol{Y} \bigm| \mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr).$$

It will finally be assumed in the whole chapter that the model matrix $\mathbb{X}$ is of full rank, i.e.,

$$\mathsf{rank}\bigl(\mathbb{X}\bigr) = k < n.$$

10.1 Multicollinearity

A principal assumption of any regression model is correct specification of the regression function. While assuming a linear model $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, this means that $\mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) \in \mathcal{M}\big(\mathbb{X}\big)$. 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.

10.1.1 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

$$\begin{aligned}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1} &= \sum_{j=0}^{k-1} \frac{1}{d_j^2}\, \boldsymbol{v}_j\, \boldsymbol{v}_j^\top = \mathbb{V}\,\mathbb{D}^{-2}\,\mathbb{V}^\top,\\ \mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\} &= \sum_{j=0}^{k-1} \frac{1}{d_j^2}.\end{aligned} \tag{10.1}$$

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*

10.1.2 Multicollinearity and its impact on precision of the LSE

It is seen from (10.1) 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 $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1,\,\dots,\,\widehat{Y}_n\big)^\top = \mathbb{H}\boldsymbol{Y}$, 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 $\widehat{\boldsymbol{\beta}} = \big(\widehat{\beta}_0,\,\dots,\,\widehat{\beta}_{k-1}\big)^\top = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ 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.

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 10.1 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 $\mathbb{Z}$.

$\underline{\mathbb{E}\Big(\|\widehat{\boldsymbol{Y}}\|^2 - \|\mathbb{X}\boldsymbol{\beta}\|^2 \;\Big|\; \mathbb{Z}\Big)}$

- 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:

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

- So that, $$\begin{aligned}\mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 - \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 &= \sigma^2\,k,\\ \mathbb{E}\big\|\widehat{\boldsymbol{Y}}\big\|^2 &= \big\|\mathbb{X}\boldsymbol{\beta}\big\|^2 + \sigma^2\,k.\end{aligned}$$

$\underline{\mathbb{E}\Big(\|\widehat{\boldsymbol{\beta}}\|^2 - \|\boldsymbol{\beta}\|^2 \;\Big|\; \mathbb{Z}\Big)}$

- 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, $$\begin{aligned}\mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2 - \big\|\boldsymbol{\beta}\big\|^2 &= \sigma^2\,\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\},\\ \mathbb{E}\big\|\widehat{\boldsymbol{\beta}}\big\|^2 &= \big\|\boldsymbol{\beta}\big\|^2 + \underbrace{\sigma^2\,\mathsf{tr}\Big\{\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big\}}_{\sum_{j=0}^{k-1}\mathsf{var}(\widehat{\beta}_j)}.\end{aligned}$$

❑

10.1.3 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

$$\boldsymbol{Y} = \big(Y_1,\,\dots,\,Y_n\big)^\top,$$

$$\mathbb{X} = \big(\mathbf{1}_n,\,\boldsymbol{X}^1,\,\dots,\,\boldsymbol{X}^{k-1}\big), \qquad \boldsymbol{X}^j = \big(X_{1,j},\,\dots,\,X_{n,j}\big)^\top, \quad j = 1,\dots,k-1,$$

the following (partly standard) notation, will be used:

Response sample mean: $\overline{Y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i;$

Square root of the total sum of squares: $T_Y = \sqrt{\displaystyle\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)^2} = \big\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\big\|;$

Fitted values: $\widehat{\boldsymbol{Y}} = \big(\widehat{Y}_1,\,\dots,\,\widehat{Y}_n\big)^\top;$

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 $j = 1,\dots,k-1$, consider a linear model M_j, where the vector $\boldsymbol{X}^j$ acts as a response and the model matrix is

$$\mathbb{X}^{(-j)} = \big(\mathbf{1}_n,\,\boldsymbol{X}^1,\,\dots,\,\boldsymbol{X}^{j-1},\,\boldsymbol{X}^{j+1},\,\dots,\,\boldsymbol{X}^{k-1}\big).$$

The following notation will be used:

Column sample mean: $\overline{X}^j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_{i,j};$

Square root of the total sum of squares from model M_j:

$$T_j = \sqrt{\sum_{i=1}^{n}\big(X_{i,j} - \overline{X}^j\big)^2} = \big\|\boldsymbol{X}^j - \overline{X}^j\mathbf{1}_n\big\|;$$

Fitted values from model M_j: $\widehat{\boldsymbol{X}}^j = \big(\widehat{X}_{1,j},\,\dots,\,\widehat{X}_{n,j}\big)^\top;$

Coefficient of determination from model M_j:

$$R_j^2 = 1 - \frac{\big\|\boldsymbol{X}^j - \widehat{\boldsymbol{X}}^j\big\|^2}{\big\|\boldsymbol{X}^j - \overline{X}^j\mathbf{1}_n\big\|^2} = 1 - \frac{\big\|\boldsymbol{X}^j - \widehat{\boldsymbol{X}}^j\big\|^2}{T_j^2}.$$

Notes.

(i) If data (response random variables and non-intercept covariates) $\big(Y_i,\,X_{i,1},\,\dots,\,X_{i,k-1}\big)^\top$, $i = 1,\dots,n$ are a random sample from a distribution of a generic random vector $\big(Y,\,X_1,\,\dots,\,X_{k-1}\big)^\top$ then

- The coefficient of determination R^2 is also a squared value of a sample coefficient of multiple correlation between Y and $\boldsymbol{X} := \big(X_1,\,\dots,\,X_{k-1}\big)^\top$.

- 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 $\boldsymbol{X}_{(-j)} := \bigl(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_{k-1}\bigr)^\top$.

(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 $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$,

$$\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\bigr) = \mathsf{MS}_e\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}.$$

The following Theorem shows that diagonal elements of the matrix $\mathsf{MS}_e\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}$, i.e., values $\widehat{\mathsf{var}}\bigl(\widehat{\beta}_j \,\big|\, \mathbb{Z}\bigr)$ can also be calculated, for $j = 1, \ldots, k-1$, using above defined quantities T_Y, T_j, R^2, R_j^2.

Theorem 10.2 Estimated variances of the LSE of the regression coefficients.

For a given dataset for which a linear model $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$, $\boldsymbol{X} = \bigl(\mathbf{1}_n, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^{k-1}\bigr)$ *is applied, diagonal elements of the matrix* $\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\bigr) = \mathsf{MS}_e\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}$, *can also be calculated, for* $j = 1, \ldots, k-1$, *as*

$$\widehat{\mathsf{var}}\bigl(\widehat{\beta}_j \,\big|\, \mathbb{Z}\bigr) = \left(\frac{T_Y}{T_j}\right)^2 \cdot \frac{1-R^2}{n-k} \cdot \frac{1}{1-R_j^2}.$$

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

$$\boldsymbol{Y}^\star = \frac{1}{T_Y}\bigl(\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\bigr),$$

$$\boldsymbol{X}^{j,\star} = \frac{1}{T_j}\bigl(\boldsymbol{X} - \overline{X}^j\mathbf{1}_n\bigr), \qquad j = 1, \ldots, k-1$$

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 10.1.3. 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 $\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}} \,\big|\, \mathbb{Z}\bigr) = \mathsf{MS}_e\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}$, 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

$$\mathbb{X}^\star = \bigl(\boldsymbol{X}^{1,\star}, \ldots, \boldsymbol{X}^{k-1,\star}\bigr)$$

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.,

$$\bigl(\boldsymbol{Y}^\star\bigr)^\top\mathbf{1}_n = 0, \qquad \bigl(\mathbb{X}^\star\bigr)^\top\mathbf{1}_n = \mathbf{0}_{k-1}.$$

Let us now consider a linear model based on *standardized variables* (as if T_Y, $T_1, \ldots, T_{k-1}$ were prespecified constants). Let $\bigl(\beta_0^\star, \beta_1^\star, \ldots, \beta_{k-1}^\star\bigr)^\top$ be the regression coefficients in a model

$$\mathsf{M}^\star\colon\ \boldsymbol{Y}^\star \,\big|\, \mathbb{Z} \sim \Bigl(\bigl(\mathbf{1}_n, \mathbb{X}^\star\bigr) \begin{pmatrix} \beta_0^\star \\ \boldsymbol{\beta}^\star \end{pmatrix}, (\sigma^\star)^2 \mathbf{I}_n \Bigr),$$

with the model matrix $\mathbb{X}_{st} = \bigl(\mathbf{1}_n, \mathbb{X}^\star\bigr)$. Let $\boldsymbol{\beta}^\star = \bigl(\beta_1^\star, \ldots, \beta_{k-1}^\star\bigr)^\top$ be the subvector of the regression coefficients related to the non-intercept columns of the model matrix.

As usually, let $\boldsymbol{\beta} = \bigl(\beta_0, \beta_1, \ldots, \beta_{k-1}\bigr)^\top$ be the regression coefficients in the original model

$$\mathsf{M}\colon\ \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta}, \sigma^2\,\mathbf{I}_n\bigr).$$

Model M can be written as

$$\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \sum_{j=1}^{k-1} \boldsymbol{X}^j \beta_j + \boldsymbol{\varepsilon}, \tag{10.2}$$

where $\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z} \sim \bigl(\mathbf{0}_n\, \sigma^2\, \mathbf{I}_n\bigr)$.

That is, data satisfying model M also satisfy

$$\boldsymbol{Y} - \overline{Y}\mathbf{1}_n = (\beta_0 - \overline{Y})\mathbf{1}_n + \sum_{j=1}^{k-1}(\boldsymbol{X}^j - \overline{X}^j\mathbf{1}_n)\beta_j + \sum_{j=1}^{k-1}\overline{X}^j\beta_j\mathbf{1}_n + \boldsymbol{\varepsilon},$$

$$\underbrace{\frac{1}{T_Y}\,(\boldsymbol{Y} - \overline{Y}\mathbf{1}_n)}_{\boldsymbol{Y}^\star} = \underbrace{\frac{\bigl(\beta_0 - \overline{Y} + \sum_{j=1}^{k-1}\overline{X}^j\beta_j\bigr)}{T_Y}}_{\beta_0^\star}\,\mathbf{1}_n + \sum_{j=1}^{k-1}\underbrace{\frac{1}{T_j}\,(\boldsymbol{X}^j - \overline{X}^j\mathbf{1}_n)}_{\boldsymbol{X}^{j,\star}}\underbrace{\frac{T_j}{T_Y}\beta_j}_{\beta_j^\star} + \underbrace{\frac{1}{T_Y}\boldsymbol{\varepsilon}}_{\boldsymbol{\varepsilon}^\star}.$$

In other words, if data satisfy model M then the standardized data satisfy the model $\mathsf{M}^\star$ with the error terms $\boldsymbol{\varepsilon}^\star = \boldsymbol{Y}^\star - \beta_0^\star\mathbf{1}_n - \mathbb{X}^\star\boldsymbol{\beta}^\star$ having $\boldsymbol{\varepsilon}^\star \,\big|\, \mathbb{Z} \sim \bigl(\mathbf{0}_n\,(\sigma^\star)^2\,\mathbf{I}_n\bigr)$ 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,

$$\widehat{\boldsymbol{Y}}^{\star} = \frac{1}{T_Y}\bigl(\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}_n\bigr), \qquad \widehat{Y}_i^{\star} = \frac{1}{T_Y}(\widehat{Y}_i - \overline{Y}), \quad i = 1, \dots, n,$$

$$\widehat{\boldsymbol{Y}} = T_Y\widehat{\boldsymbol{Y}}^{\star} + \overline{Y}\mathbf{1}_n, \qquad \widehat{Y}_i = T_Y\,\widehat{Y}_i^{\star} + \overline{Y}, \quad i = 1, \dots, n.$$

The residual sum of squares in model $\mathsf{M}^\star$ is then:

$$\begin{aligned}
\mathsf{SS}_e^\star &= \bigl\|\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^{\star}\bigr\|^2 \\
&= \sum_{i=1}^{n}(Y_i^\star - \widehat{Y}_i^\star)^2 = \frac{1}{T_Y^2}\sum_{i=1}^{n}(T_Y\,Y_i^\star - T_Y\,\widehat{Y}_i^\star)^2 \\
&= \frac{1}{T_Y^2}\sum_{i=1}^{n}\bigl\{T_Y\,Y_i^\star + \overline{Y} - (T_Y\,\widehat{Y}_i^\star + \overline{Y})\bigr\}^2 \\
&= \frac{1}{T_Y^2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{T_Y^2}\,\mathsf{SS}_e,
\end{aligned}$$

where SS_e is the residual sum of squares in the original model M.

Moreover, note that $T_Y^2 = \bigl\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\bigr\|^2$ is also the total sum of squares SS_T for the original response vector $\boldsymbol{Y}$. That is,

$$\mathsf{SS}_e^\star = \frac{\mathsf{SS}_e}{\mathsf{SS}_T} = 1 - R^2, \tag{10.3}$$

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 $\bigl(\beta_0^\star,\, \boldsymbol{\beta}^{\star\top}\bigr)^\top$ in model $\mathsf{M}^\star$ which are given as

$$\begin{pmatrix}\widehat{\beta}_0^\star \\ \widehat{\boldsymbol{\beta}}^\star\end{pmatrix} = \Bigl(\mathbb{X}_{st}^\top\mathbb{X}_{st}\Bigr)^{-1}\mathbb{X}_{st}^\top\boldsymbol{Y}^\star.$$

First,

$$\mathbb{X}_{st}^\top\mathbb{X}_{st} \;=\; \bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr)^\top\bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr) \;=\; \begin{pmatrix} n & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \bigl(\mathbb{X}^\star\bigr)^\top\mathbb{X}^\star \end{pmatrix} \;=\; \begin{pmatrix} n & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X} \end{pmatrix},$$

where $\mathbb{R}_{X,X} := \bigl(\mathbb{X}^\star\bigr)^\top\mathbb{X}^\star = \bigl(r_{X,X}^{j,l}\bigr)_{j,l=1,\dots,k-1}$, has elements

$$\begin{aligned}
r_{X,X}^{j,l} &= \frac{\sum_{i=1}^{n}(X_{i,j} - \overline{X}^j)(X_{i,l} - \overline{X}^l)}{T_j\,T_l} \\
&= \frac{\sum_{i=1}^{n}(X_{i,j} - \overline{X}^j)(X_{i,l} - \overline{X}^l)}{\sqrt{\sum_{i=1}^{n}(X_{i,j} - \overline{X}^j)^2}\sqrt{\sum_{i=1}^{n}(X_{i,l} - \overline{X}^l)^2}}, \quad j,l = 1, \dots, k-1.
\end{aligned}$$

That is, $\mathbb{R}_{X,X} = \bigl(\mathbb{X}^\star\bigr)^\top\mathbb{X}^\star$ 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,

$$\Bigl(\mathbb{X}_{st}^\top\mathbb{X}_{st}\Bigr)^{-1} \;=\; \begin{pmatrix} n & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X} \end{pmatrix} \;=\; \begin{pmatrix} \dfrac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}^{-1} \end{pmatrix}.$$

Second,

$$\mathbb{X}_{st}^\top \boldsymbol{Y}^\star \;=\; \bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr)^\top \boldsymbol{Y}^\star \;=\; \begin{pmatrix} \sum_{i=1}^n Y_i^\star \\ \bigl(\mathbb{X}^\star\bigr)^\top \boldsymbol{Y}^\star \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ \boldsymbol{r}_{X,Y} \end{pmatrix},$$

where $\boldsymbol{r}_{X,Y} := \bigl(\mathbb{X}^\star\bigr)^\top \boldsymbol{Y}^\star = \bigl(r_{X,Y}^j\bigr)_{j=1,\dots,k-1}$, has elements

$$\begin{aligned} r_{X,Y}^j &= \frac{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)(Y_i - \overline{Y})}{T_j\, T_Y} \\ &= \frac{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_{i,j} - \overline{X}^j)^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}, \qquad j = 1, \dots, k-1, \end{aligned}$$

That is, $\boldsymbol{r}_{X,Y} = \bigl(\mathbb{X}^\star\bigr)^\top \boldsymbol{Y}^\star$ is a vector of sample correlation coefficients between the regressors from the model matrix $\mathbb{X}$ and the response $\boldsymbol{Y}$.

Hence,

$$\begin{pmatrix} \widehat{\beta}_0^\star \\ \widehat{\boldsymbol{\beta}}^\star \end{pmatrix} = \begin{pmatrix} \dfrac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ \boldsymbol{r}_{X,Y} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbb{R}_{X,X}^{-1}\, \boldsymbol{r}_{X,Y} \end{pmatrix},$$

$$\mathsf{var}\left\{ \begin{pmatrix} \widehat{\beta}_0^\star \\ \widehat{\boldsymbol{\beta}}^\star \end{pmatrix} \,\middle|\, \mathbb{Z} \right\} = (\sigma^\star)^2 \left(\mathbb{X}_{st}^\top \mathbb{X}_{st}\right)^{-1} = (\sigma^\star)^2 \begin{pmatrix} \dfrac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{X,X}^{-1} \end{pmatrix}.$$

That is, we have

$$\begin{aligned} \widehat{\beta}_0^\star &= 0, & \mathsf{var}\bigl(\widehat{\beta}_0^\star \,\big|\, \mathbb{Z}\bigr) &= \frac{(\sigma^\star)^2}{n}, \\ \widehat{\boldsymbol{\beta}}^\star &= \mathbb{R}_{X,X}^{-1}\, \boldsymbol{r}_{X,Y} & \mathsf{var}\bigl(\widehat{\boldsymbol{\beta}}^\star \,\big|\, \mathbb{Z}\bigr) &= (\sigma^\star)^2\, \mathbb{R}_{X,X}^{-1}. \end{aligned} \tag{10.4}$$

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

$$\begin{aligned} \mathbb{H}_{st} = \bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr) \left(\mathbb{X}_{st}^\top \mathbb{X}_{st}\right)^{-1} \bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr)^\top &= \bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr) \begin{pmatrix} \dfrac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \mathbb{R}_{x,x}^{-1} \end{pmatrix} \bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr)^\top \\ &= \frac{1}{n}\,\mathbf{1}_n\,{\mathbf{1}_n}^\top + \underbrace{\mathbb{X}^\star \mathbb{R}_{X,X}^{-1} \bigl(\mathbb{X}^\star\bigr)^\top}_{=:\, \mathbb{H}^\star}. \end{aligned}$$

Observe that

- $\mathbb{H}_{st}$ is the projection matrix into $\mathcal{M}\Bigl(\bigl(\mathbf{1}_n,\, \mathbb{X}^\star\bigr)\Bigr)$.
- $\mathbb{H}^\star = \mathbb{X}^\star \mathbb{R}_{X,X}^{-1} \bigl(\mathbb{X}^\star\bigr)^\top$ is the projection matrix into $\mathcal{M}\bigl(\mathbb{X}^\star\bigr)$.

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)

$$\bigl(\widehat{\boldsymbol{Y}}^\star\bigr)^\top \boldsymbol{Y}^\star \;=\; \bigl(\boldsymbol{Y}^\star\bigr)^\top \mathbb{H}^\star \boldsymbol{Y}^\star \;=\; \bigl(\boldsymbol{Y}^\star\bigr)^\top \mathbb{H}^\star \mathbb{H}^\star \boldsymbol{Y}^\star \;=\; \bigl(\widehat{\boldsymbol{Y}}^\star\bigr)^\top \widehat{\boldsymbol{Y}}^\star.$$

Consequently,

$$\mathsf{SS}_e^\star = \big\|\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\big\|^2 = \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star - \big(\boldsymbol{Y}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star - \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \boldsymbol{Y}^\star + \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star$$
$$= \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star - \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star. \qquad (10.5)$$

Let $d_{X,X}^{j,j}$, $j = 1, \dots, k-1$ be diagonal elements of the matrix $\mathbb{R}_{X,X}^{-1}$. That is, from (10.4):

$$\mathsf{var}\big(\widehat{\beta}_j^\star \,\big|\, \mathbb{Z}\big) = (\sigma^\star)^2\, d_{X,X}^{j,j}, \quad j = 1, \dots, k-1.$$

To derive the value of $d_{X,X}^{j,j}$, $j = 1, \dots, k-1$, 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:

$$\mathbb{R}_{(Y,X),(Y,X)}^{-1} = \begin{pmatrix} \big(1 - \boldsymbol{r}_{X,Y}^\top \mathbb{R}_{X,X}^{-1} \boldsymbol{r}_{X,Y}\big)^{-1} & \text{✿} \\ \text{✿} & \text{✿} \end{pmatrix}.$$

Further (while also using Eqs. 10.3 and 10.5),

$$\begin{aligned} 1 &- \boldsymbol{r}_{X,Y}^\top \mathbb{R}_{X,X}^{-1} \boldsymbol{r}_{X,Y} \\ &= \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star - \big(\boldsymbol{Y}^\star\big)^\top \underbrace{\mathbb{X}^\star\big\{(\mathbb{X}^\star)^\top \mathbb{X}^\star\big\}^{-1}(\mathbb{X}^\star)^\top}_{\mathbb{H}^\star} \boldsymbol{Y}^\star \\ &= \big(\boldsymbol{Y}^\star\big)^\top \boldsymbol{Y}^\star - \big(\widehat{\boldsymbol{Y}}^\star\big)^\top \widehat{\boldsymbol{Y}}^\star = \mathsf{SS}_e^\star = 1 - R^2, \end{aligned}$$

where R^2 is coefficient of determination from the linear model $\mathsf{M}\colon\ \boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\big)$.

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 $\boldsymbol{X}^1, \dots, \boldsymbol{X}^{k-1}$, i.e., the model matrix

$$\mathbb{X} = \big(\mathbf{1}_n,\, \boldsymbol{X}^1,\, \dots,\, \boldsymbol{X}^{k-1}\big).$$

Now, consider for given $j = 1, \dots, k-1$ 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

$$\mathbb{X}^{(-j)} = \big(\mathbf{1}_n,\, \boldsymbol{X}^1,\, \dots,\, \boldsymbol{X}^{j-1},\, \boldsymbol{X}^{j+1},\, \dots,\, \boldsymbol{X}^{k-1}\big).$$

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,

$$\mathsf{var}\big(\widehat{\beta}_j^\star \,\big|\, \mathbb{Z}\big) = \frac{(\sigma^\star)^2}{1 - R_j^2}, \qquad j = 1, \dots, k-1.$$

The jth diagonal element ($j = 1, \dots, k-1$) of the matrix $\mathsf{var}\bigl(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{Z}\bigr)$ can now be expressed as

$$\mathsf{var}\bigl(\widehat{\beta}_j\,\big|\,\mathbb{Z}\bigr) \;=\; \mathsf{var}\biggl(\frac{T_Y}{T_j}\widehat{\beta}^\star_j\,\bigg|\,\mathbb{Z}\biggr) \;=\; \biggl(\frac{T_Y}{T_j}\biggr)^2 \mathsf{var}\bigl(\widehat{\beta}^\star_j\,\big|\,\mathbb{Z}\bigr) \;=\; \biggl(\frac{T_Y}{T_j}\biggr)^2 \frac{(\sigma^\star)^2}{1-R_j^2}.$$

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

$$\widehat{\mathsf{var}}\bigl(\widehat{\beta}_j\,\big|\,\mathbb{Z}\bigr) \;=\; \biggl(\frac{T_Y}{T_j}\biggr)^2 \frac{1-R^2}{n-k}\,\frac{1}{1-R_j^2} \qquad j = 1, \dots, k-1.$$

❑

Definition 10.1 Variance inflation factor and tolerance.

For given $j = 1, \dots, k-1$, the variance inflation factor[2] *and the* tolerance[3] *of the jth regressor of the linear model $\boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$, $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$ are values VIF_j and Toler_j, respectively, defined as*

$$\mathit{VIF}_j = \frac{1}{1-R_j^2}, \qquad \mathit{Toler}_j = 1 - R_j^2 = \frac{1}{\mathit{VIF}_j}.$$

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 10.2, the VIF of the jth regressor ($j = 1, \dots, k-1$) 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

  $$\widehat{\beta}_j \;\pm\; \mathsf{t}_{n-k}\Bigl(1-\frac{\alpha}{2}\Bigr)\,\sqrt{\widehat{\mathsf{var}}\bigl(\widehat{\beta}_j\,\big|\,\mathbb{Z}\bigr)}.$$

  Using the statement of Theorem 10.2, the lower and the upper bounds of the confidence interval for β_j can also be written as

  $$\widehat{\beta}_j \;\pm\; \mathsf{t}_{n-k}\Bigl(1-\frac{\alpha}{2}\Bigr)\,\frac{T_Y}{T_j}\,\sqrt{\frac{1-R^2}{n-k}}\,\sqrt{\mathrm{VIF}_j}.$$

  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,

  $$\mathrm{VIF}_j = \biggl(\frac{\mathrm{Vol}_j}{\mathrm{Vol}_{0,j}}\biggr)^2, \tag{10.6}$$

  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.

[2] *varianční inflační faktor* [3] *tolerance*

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

- $\mathcal{J} \subset \{1,\ldots,k-1\}$, $\bigl|\mathcal{J}\bigr| = m$;
- $\boldsymbol{\beta}_{[\mathcal{J}]}$ 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

$$\Bigl\{\boldsymbol{\beta}_{[\mathcal{J}]} \in \mathbb{R}^m: \; \bigl(\boldsymbol{\beta}_{[\mathcal{J}]} - \widehat{\boldsymbol{\beta}}_{[\mathcal{J}]}\bigr)^\top \bigl(\mathsf{MS}_e\,\mathbb{V}_{[\mathcal{J}]}\bigr)^{-1} \bigl(\boldsymbol{\beta}_{[\mathcal{J}]} - \widehat{\boldsymbol{\beta}}_{[\mathcal{J}]}\bigr) \; < \; m\,\mathcal{F}_{m,n-k}(1-\alpha)\Bigr\},$$

$$\mathbb{V}_{[\mathcal{J}]} = (\mathcal{J}-\mathcal{J}) \text{ block of the matrix } \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}. \tag{10.7}$$

Let $\mathrm{Vol}_{\mathcal{J}}$: volume of the confidence ellipsoid (10.7);

$\mathrm{Vol}_{0,\mathcal{J}}$: volume of the confidence ellipsoid (10.7) 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 (10.6) 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 $j = 1,\ldots,k-1$, 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 $\mathrm{gVIF}_{\mathcal{J}}$ for index sets $\mathcal{J}$, $\bigl|\mathcal{J}\bigr|$ of different cardinality m, quantities

  $$\mathrm{gVIF}_{\mathcal{J}}^{\frac{1}{2m}} = \left(\frac{\mathrm{Vol}_{\mathcal{J}}}{\mathrm{Vol}_{0,\mathcal{J}}}\right)^{\frac{1}{m}} \tag{10.8}$$

  should be compared which all relate to volume units in 1D.
- Generalized VIF's (and standard VIF's if $m = 1$) together with (10.8) are calculated by the R function `vif` from the package `car`.

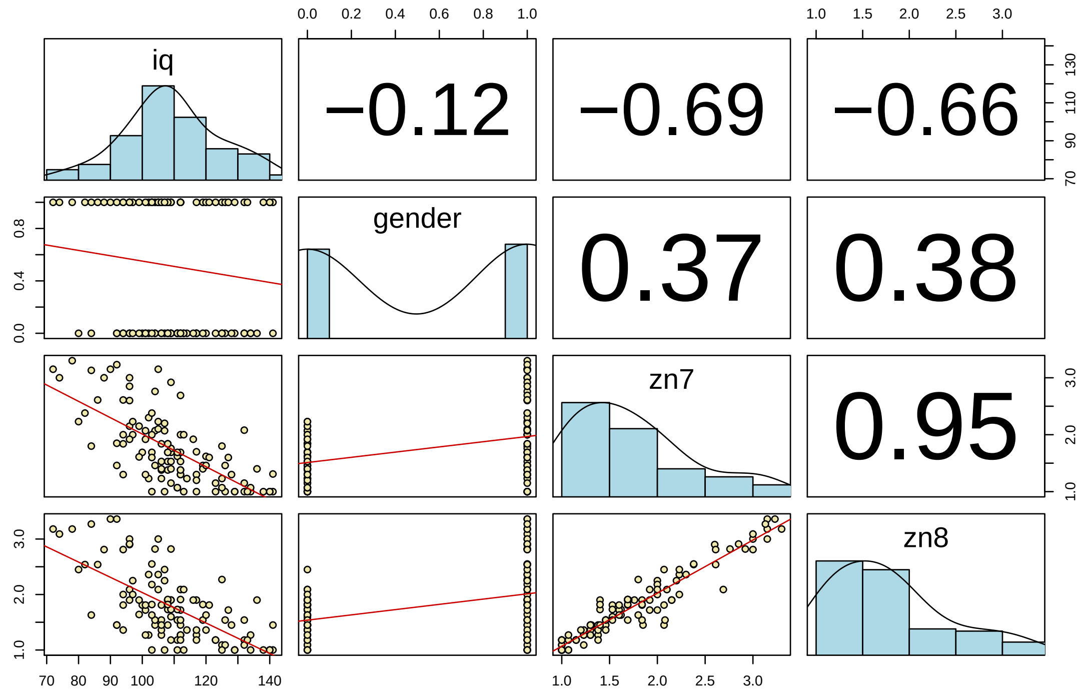
End of skipped part

10.1.4 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.

Illustrations

IQ ($n = 111$)
iq ∼ gender + zn7 + zn8



IQ ($n = 111$)
iq ∼ gender + zn7 + zn8

```
summary(m1 <- lm(iq ~ gender + zn7 + zn8, data = IQ))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-22.1677  -7.5243  -0.4338   7.1780  26.4095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  138.222      3.119  44.314  < 2e-16 ***
gender         4.563      2.221   2.055  0.04232 *
zn7          -16.767      5.536  -3.029  0.00308 **
zn8           -1.149      5.557  -0.207  0.83658
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.81 on 107 degrees of freedom
Multiple R-squared:  0.4943,	Adjusted R-squared:  0.4801
F-statistic: 34.87 on 3 and 107 DF,  p-value: 8.472e-16
```

```
library("car")
vif(m1)
```

```
  gender      zn7      zn8
 1.16923 11.26866 11.40240
```

Illustrations

IQ ($n = 111$)
iq ~ gender + zn7

```
(sm27 <- summary(m27 <- lm(iq ~ gender + zn7, data = IQ)))
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-21.9606  -7.4290  -0.1927   7.0047  26.5244 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  138.093      3.043  45.376   <2e-16 ***
gender         4.513      2.198   2.054   0.0424 *  
zn7          -17.852      1.765 -10.116   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.77 on 108 degrees of freedom
Multiple R-squared:  0.4941,	Adjusted R-squared:  0.4848 
F-statistic: 52.74 on 2 and 108 DF,  p-value: < 2.2e-16
```

```
vif(m27)
```

```
 gender     zn7 
1.15531 1.15531 
```

IQ ($n = 111$)
iq ~ gender + zn8

```
(sm28 <- summary(m28 <- lm(iq ~ gender + zn8, data = IQ)))
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-25.5378  -7.9585  -0.0763   7.1273  31.0778 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  137.402      3.223  42.634  < 2e-16 ***
gender         4.474      2.303   1.943   0.0547 .  
zn8          -17.095      1.846  -9.263 2.21e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.22 on 108 degrees of freedom
Multiple R-squared:  0.451,	Adjusted R-squared:  0.4408 
F-statistic: 44.36 on 2 and 108 DF,  p-value: 8.673e-15
```

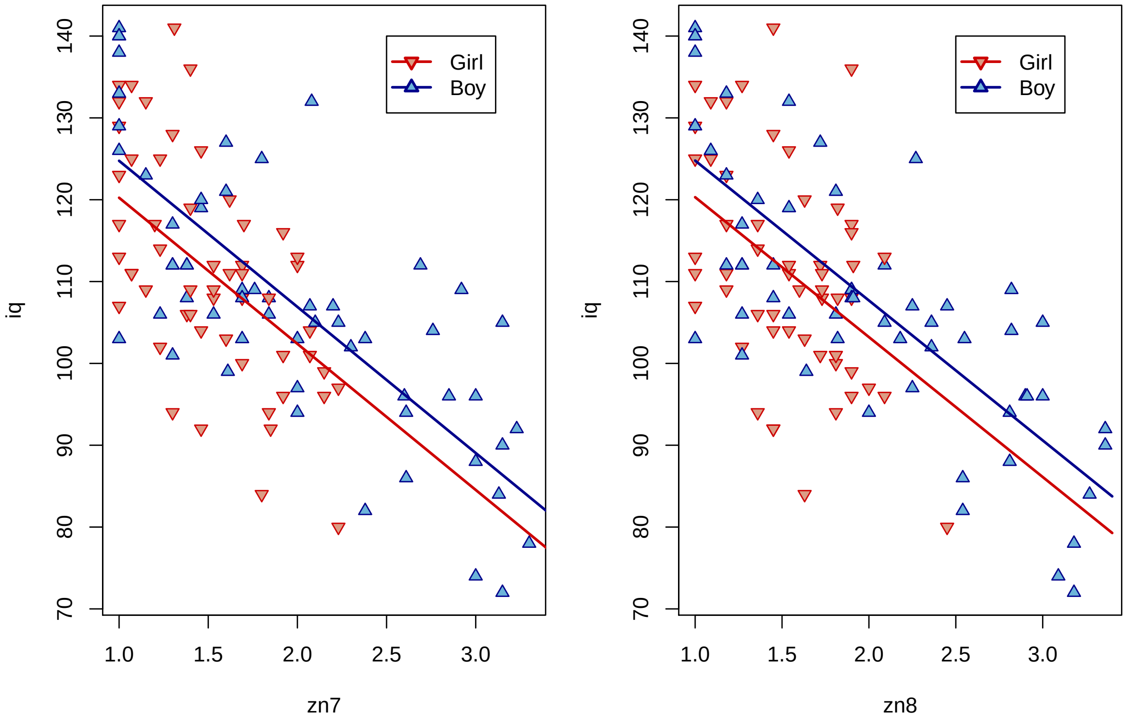
```
vif(m28)
```

```
  gender      zn8 
1.169022 1.169022 
```

Illustrations

IQ ($n = 111$)

iq $\sim$ gender + znX



10.2 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 10.1, 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.

10.2.1 Omitted and irrelevant regressors

We will assume that possibly two sets of regressors are available:

(i) $\boldsymbol{X}_i$, $i = 1, \dots, n$, 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

$$\mathbb{X}_{n\times k} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} = \Bigl(\boldsymbol{X}^0, \dots, \boldsymbol{X}^{k-1}\Bigr).$$

It will still be assumed that $\boldsymbol{X}^0 = \bigl(1, \dots, 1\bigr)^\top$ (almost surely) leading to the model matrix

$$\mathbb{X}_{n\times k} = \Bigl(\mathbf{1}_n, \boldsymbol{X}^1, \dots, \boldsymbol{X}^{k-1}\Bigr),$$

with explicitely included intercept term.

(ii) $\boldsymbol{V}_i$, $i = 1, \dots, n$, 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

$$\mathbb{V}_{n\times l} = \begin{pmatrix} \boldsymbol{V}_1^\top \\ \vdots \\ \boldsymbol{V}_n^\top \end{pmatrix} = \Bigl(\boldsymbol{V}^1, \dots, \boldsymbol{V}^l\Bigr).$$

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

$$\begin{aligned} &\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k, \quad \mathsf{rank}\bigl(\mathbb{V}_{n\times l}\bigr) = l, \\ &\text{for} \quad \mathbb{G}_{n\times(k+l)} := \bigl(\mathbb{X}, \mathbb{V}\bigr), \quad \mathsf{rank}\bigl(\mathbb{G}\bigr) = k + l < n. \end{aligned}$$

The matrices $\mathbb{X}$ and $\mathbb{G}$ give rise to two nested linear models:

Model M_X $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\bigr)$;

Model M_{XV} $\boldsymbol{Y} \,\big|\, \mathbb{Z} \sim \bigl(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n\bigr)$.

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 $\boldsymbol{\gamma} \neq \mathbf{0}_l$) 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 $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_X = \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\,\mathbb{X}^\top\boldsymbol{Y} = \bigl(\widehat{\beta}_{X,0},\,\ldots,\,\widehat{\beta}_{X,k-1}\bigr)^\top;$$

- Projection matrices into the regression space $\mathcal{M}\bigl(\mathbb{X}\bigr)$ and into the residual space $\mathcal{M}\bigl(\mathbb{X}\bigr)^\perp$:

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

- Fitted values (LSE of a vector $\mathbb{X}\boldsymbol{\beta}$):

$$\widehat{\boldsymbol{Y}}_X = \mathbb{H}_X\boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}_X = \bigl(\widehat{Y}_{X,1},\,\ldots,\,\widehat{Y}_{X,n}\bigr)^\top;$$

- Residuals

$$\boldsymbol{U}_X = \mathbb{M}_X\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_X = \bigl(U_{X,1},\,\ldots,\,U_{X,n}\bigr)^\top;$$

- Residual sum of squares and residual mean square:

$$\mathsf{SS}_{e,X} = \bigl\|\boldsymbol{U}_X\bigr\|^2, \qquad \mathsf{MS}_{e,X} = \frac{\mathsf{SS}_{e,X}}{n-k}.$$

(ii) Quantities derived while assuming model M_{XV}:

- Least squares estimator of $\bigl(\boldsymbol{\beta}^\top,\,\boldsymbol{\gamma}^\top\bigr)^\top$:

$$\bigl(\widehat{\boldsymbol{\beta}}_{XV}^\top,\,\widehat{\boldsymbol{\gamma}}_{XV}^\top\bigr)^\top = \bigl(\mathbb{G}^\top\mathbb{G}\bigr)^{-1}\,\mathbb{G}^\top\boldsymbol{Y},$$

$$\widehat{\boldsymbol{\beta}}_{XV} = \bigl(\widehat{\beta}_{XV,0},\,\ldots,\,\widehat{\beta}_{XV,k-1}\bigr)^\top, \qquad \widehat{\boldsymbol{\gamma}}_{XV} = \bigl(\widehat{\gamma}_{XV,1},\,\ldots,\,\widehat{\gamma}_{XV,l}\bigr)^\top;$$

- Projection matrices into the regression space $\mathcal{M}\bigl(\mathbb{G}\bigr)$ and into the residual space $\mathcal{M}\bigl(\mathbb{G}\bigr)^\perp$:

$$\mathbb{H}_{XV} = \mathbb{G}\bigl(\mathbb{G}^\top\mathbb{G}\bigr)^{-1}\mathbb{G}^\top, \qquad \mathbb{M}_{XV} = \mathbf{I}_n - \mathbb{H}_{XV};$$

- Fitted values (LSE of a vector $\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma}$):

$$\widehat{\boldsymbol{Y}}_{XV} = \mathbb{H}_{XV}\boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}}_{XV} + \mathbb{V}\boldsymbol{\gamma}_{XV} = \bigl(\widehat{Y}_{XV,1},\,\ldots,\,\widehat{Y}_{XV,n}\bigr)^\top;$$

- Residuals

$$\boldsymbol{U}_{XV} = \mathbb{M}_{XV}\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_{XV} = \bigl(U_{XV,1},\,\ldots,\,U_{XV,n}\bigr)^\top;$$

- Residual sum of squares and residual mean square:

$$\mathsf{SS}_{e,XV} = \left\|\boldsymbol{U}_{XV}\right\|^2, \qquad \mathsf{MS}_{e,XV} = \frac{\mathsf{SS}_{e,XV}}{n-k-l}.$$

Consequence of Lemma 9.1: 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:*

$$\begin{aligned}
\widehat{\boldsymbol{Y}}_{XV} - \widehat{\boldsymbol{Y}}_X &= \mathbb{M}_X\mathbb{V}\big(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\big)^{-1}\mathbb{V}^\top\boldsymbol{U}_X,\\
&= \mathbb{X}\big(\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X\big) + \mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV},\\
\widehat{\boldsymbol{\gamma}}_{XV} &= \big(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\big)^{-1}\mathbb{V}^\top\boldsymbol{U}_X,\\
\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X &= -\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV},\\
\mathsf{SS}_{e,X} - \mathsf{SS}_{e,XV} &= \left\|\mathbb{M}_X\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}\right\|^2,
\end{aligned}$$

$$\mathbb{H}_{XV} = \mathbb{H}_X + \mathbb{M}_X\mathbb{V}\left(\mathbb{V}^\top\mathbb{M}_X\mathbb{V}\right)^{-1}\mathbb{V}^\top\mathbb{M}_X.$$

Proof. Direct use of Lemma 9.1 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 9.1. 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 $\mathbb{X}^\top\mathbb{M}_X\mathbb{V} = \mathbf{0}_{k\times l}$.

❑

Lemma 10.3 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:*

$$\begin{aligned} \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) - \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) &\geq 0, \\ \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) - \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X} \,\big|\, \mathbb{Z}\big) &\geq 0. \end{aligned}$$

Proof.

$\mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) - \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) \geq 0$

We have,
$$\begin{aligned} \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) &= \mathsf{var}\big(\mathbb{H}_X \boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \mathbb{H}_X(\sigma^2 \mathbf{I}_n)\mathbb{H}_X \\ &= \sigma^2\, \mathbb{H}_X \qquad \text{(even if } \mathsf{M}_X \text{ is not correct).} \\ \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) &= \mathsf{var}\big(\mathbb{H}_{XV} \boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \sigma^2 \mathbb{H}_{XV} \\ &= \sigma^2 \big\{\mathbb{H}_X + \mathbb{M}_X \mathbb{V}(\mathbb{V}^\top \mathbb{M}_X \mathbb{V})^{-1}\mathbb{V}^\top \mathbb{M}_X\big\} \\ &= \mathsf{var}\big(\widehat{\boldsymbol{Y}}_{X} \,\big|\, \mathbb{Z}\big) + \underbrace{\sigma^2\, \mathbb{M}_X \mathbb{V}(\mathbb{V}^\top \mathbb{M}_X \mathbb{V})^{-1}\mathbb{V}^\top \mathbb{M}_X.}_{\text{positive semidefinite matrix}} \end{aligned}$$

$\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) - \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X} \,\big|\, \mathbb{Z}\big) \geq 0$

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):

$$\mathsf{var}\left\{\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{XV} \\ \widehat{\boldsymbol{\gamma}}_{XV}\end{pmatrix} \,\middle|\, \mathbb{Z}\right\} = \sigma^2 \begin{pmatrix} \mathbb{X}^\top\mathbb{X} & \mathbb{X}^\top\mathbb{V} \\ \mathbb{V}^\top\mathbb{X} & \mathbb{V}^\top\mathbb{V}\end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} \big\{\mathbb{X}^\top\mathbb{X} - \mathbb{X}^\top\mathbb{V}(\mathbb{V}^\top\mathbb{V})^{-1}\mathbb{V}^\top\mathbb{X}\big\}^{-1} & \text{⌘} \\ \text{⌘} & \text{⌘}\end{pmatrix}.$$

Further,

$$\begin{aligned} \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X} \,\big|\, \mathbb{Z}\big) &= \mathsf{var}\Big(\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y} \,\Big|\, \mathbb{Z}\Big) = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top(\sigma^2\mathbf{I}_n)\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1} \\ &= \sigma^2\, \big(\mathbb{X}^\top\mathbb{X}\big)^{-1} \qquad \text{(even if } \mathsf{M}_X \text{ is not correct).} \end{aligned}$$

$$\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) = \sigma^2 \Big\{\mathbb{X}^\top\mathbb{X} - \mathbb{X}^\top\mathbb{V}\big(\mathbb{V}^\top\mathbb{V}\big)^{-1}\mathbb{V}^\top\mathbb{X}\Big\}^{-1}.$$

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}.

10.2.2 Prediction quality of the fitted model

To evaluate a prediction quality of the *fitted* model, we will assume that data $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $\boldsymbol{Z}_i = \bigl(Z_{i,1},\, \dots,\, Z_{i,p}\bigr)^\top \in \mathcal{Z} \subseteq \mathbb{R}^p$, $i = 1, \dots, n$, are a random sample from a distribution of a generic random vector $\bigl(Y,\, \boldsymbol{Z}^\top\bigr)^\top$, $\boldsymbol{Z} = \bigl(Z_1,\, \dots,\, Z_p\bigr)^\top$. Let the conditional distribution $Y \,|\, \boldsymbol{Z}$ of Y given the covariates $\boldsymbol{Z}$ satisfies

$$\mathbb{E}\bigl(Y \bigm| \boldsymbol{Z}\bigr) = m(\boldsymbol{Z}), \qquad \mathsf{var}\bigl(Y \bigm| \boldsymbol{Z}\bigr) = \sigma^2, \tag{10.9}$$

for some (regression) function m and some $\sigma^2 > 0$.

Replicated response

Let $\boldsymbol{z}_1,\, \dots,\, \boldsymbol{z}_n$ be the values of the covariate vectors $\boldsymbol{Z}_1,\, \dots,\, \boldsymbol{Z}_n$ in the original data that are available to estimate the parameters of the model (10.9). Further, let $\bigl(Y_{n+i},\, \boldsymbol{Z}_{n+i}^\top\bigr)^\top$, $i = 1, \dots, n$, be independent random vectors (new or future data) being distributed as a generic random vector $\bigl(Y,\, \boldsymbol{Z}\bigr)$ and being independent of the original data $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $i = 1, \dots, n$. Suppose that our aim is to *predict* values of Y_{n+i}, $i = 1, \dots, n$, under the condition that the new covariate values are equal to the old ones. That is, we want to predict, for $i = 1, \dots, n$, values of Y_{n+i} given $\boldsymbol{Z}_{n+i} = \boldsymbol{z}_i$.

Terminology *(Replicated response).*

A random vector

$$\boldsymbol{Y}_{new} = \bigl(Y_{n+1},\, \dots,\, Y_{n+n}\bigr)^\top,$$

where Y_{n+i} is supposed to come from the conditional distribution $Y \,|\, \boldsymbol{Z} = \boldsymbol{z}_i$, $i = 1, \dots, n$, 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

$$\begin{aligned}
\mathbb{E}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{Z}_1 = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_n = \boldsymbol{z}_n\bigr) &= \quad \boldsymbol{\mu} \quad = \mathbb{E}\bigl(\boldsymbol{Y}_{new} \bigm| \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\bigr),\\
\mathsf{var}\bigl(\boldsymbol{Y} \bigm| \boldsymbol{Z}_1 = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_n = \boldsymbol{z}_n\bigr) &= \quad \sigma^2\mathbf{I}_n \quad = \mathsf{var}\bigl(\boldsymbol{Y}_{new} \bigm| \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\bigr),\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for some } \sigma^2 > 0,
\end{aligned}$$

  and

$$\boldsymbol{\mu} = \bigl(m(\boldsymbol{z}_1),\, \dots,\, m(\boldsymbol{z}_n)\bigr)^\top = \bigl(\mu_1,\, \dots,\, \mu_n\bigr)^\top.$$

Prediction of replicated response

Let

$$\widehat{\boldsymbol{Y}}_{new} = \bigl(\widehat{Y}_{n+1},\, \dots,\, \widehat{Y}_{n+n}\bigr)^\top$$

be the prediction of a vector $\boldsymbol{Y}_{new}$ based on the assumed regression model (10.9) estimated using the original data $\boldsymbol{Y}$ with $\boldsymbol{Z}_1 = \boldsymbol{z}_1,\, \dots,\, \boldsymbol{Z}_n = \boldsymbol{z}_n$. That is, $\widehat{\boldsymbol{Y}}_{new}$ is some statistic of $\boldsymbol{Y}$ (and $\mathbb{Z}$). Analogously to Section 7.3, we shall evaluate a quality of the prediction by the mean squared error of prediction (MSEP). Nevertheless, in contrast to Section 7.3, the following issues will be different:

(i) A value of a random vector rather than a value of a random variable (as in Section 7.3) 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 $\boldsymbol{Z}_i = \boldsymbol{Z}_{n+i} = \boldsymbol{z}_i$, $i = 1, \dots, n$). In contrast, variability of the covariates was taken into account in Section 7.3.

(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 7.3.

Definition 10.2 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*

$$\mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new}\big) = \sum_{i=1}^{n} \mathbb{E}\Big\{\big(\widehat{Y}_{n+i} - Y_{n+i}\big)^2 \,\Big|\, \mathbb{Z}\Big\}, \tag{10.10}$$

where the expectation is with respect to the $(n{+}n)$-dimensional conditional distribution of the vector $\big(\boldsymbol{Y}^\top, \boldsymbol{Y}_{new}^\top\big)^\top$ 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*

$$\mathsf{AMSEP}\big(\widehat{\boldsymbol{Y}}_{new}\big) = \frac{1}{n}\,\mathsf{MSEP}\big(\widehat{\boldsymbol{Y}}_{new}\big).$$

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 (10.10) is minimized for

$$\begin{aligned}\widehat{\boldsymbol{Y}}_{new} &= \widehat{\boldsymbol{Y}}, & \widehat{\boldsymbol{Y}} &= \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-}\mathbb{X}^\top\boldsymbol{Y}, \\ \widehat{Y}_{n+i} &= \widehat{Y}_i, & i &= 1, \dots, n.\end{aligned}$$

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 $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \boldsymbol{Z}_1 = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_n = \boldsymbol{z}_n\big) = \mathbb{E}\big(\boldsymbol{Y}_{new} \,\big|\, \boldsymbol{Z}_{n+1} = \boldsymbol{z}_1, \dots, \boldsymbol{Z}_{n+n} = \boldsymbol{z}_n\big)$.

[4] *střední čtvercová chyba predikce* [5] *průměrná střední čtvercová chyba predikce*

Lemma 10.4 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, $\widehat{Y}_{n+i} = \widehat{Y}_i$, $i = 1,\dots,n$),

$$\begin{aligned}
\mathbb{E}\big(\widehat{Y}_{n+i} - Y_{n+i}\big)^2 &= \mathbb{E}\big(\widehat{Y}_i - Y_{n+i}\big)^2 \\
&= \mathbb{E}\big\{\widehat{Y}_i - \mu_i - (Y_{n+i} - \mu_i)\big\}^2 \\
&= \mathbb{E}\big(\widehat{Y}_i - \mu_i\big)^2 + \mathbb{E}\big(Y_{n+i} - \mu_i\big)^2 - 2\;\underbrace{\mathbb{E}(\widehat{Y}_i - \mu_i)(Y_{n+i} - \mu_i)}_{\mathbb{E}(\widehat{Y}_i - \mu_i)\,\mathbb{E}(Y_{n+i} - \mu_i) \,=\, \mathbb{E}(\widehat{Y}_i - \mu_i)\cdot 0} \\
&= \mathbb{E}\big(\widehat{Y}_i - \mu_i\big)^2 + \mathbb{E}\big(Y_{n+i} - \mu_i\big)^2 \\
&= \mathsf{MSE}(\widehat{Y}_i) + \sigma^2.
\end{aligned}$$

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

$$\mathsf{MSE}\big(\widehat{Y}_i\big) = \mathbb{E}\Big\{\big(\widehat{Y}_i - \mu_i\big)^2\,\Big|\,\mathbb{Z}\Big\} = \mathsf{var}\big(\widehat{Y}_i\,\big|\,\mathbb{Z}\big), \qquad i = 1,\dots,n.$$

- 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

$$\begin{aligned}
\mathsf{MSE}\big(\widehat{Y}_i\big) &= \mathbb{E}\Big\{\big(\widehat{Y}_i - \mu_i\big)^2\,\Big|\,\mathbb{Z}\Big\} \\
&= \mathsf{var}\big(\widehat{Y}_i\,\big|\,\mathbb{Z}\big) + \Big\{\mathbb{E}\big(\widehat{Y}_i - \mu_i\,\big|\,\mathbb{Z}\big)\Big\}^2 = \mathsf{var}\big(\widehat{Y}_i\,\big|\,\mathbb{Z}\big) + \Big\{\mathsf{bias}\big(\widehat{Y}_i\big)\Big\}^2, \qquad i = 1,\dots,n.
\end{aligned}$$

[6] *střední čtvercová chyba*

- 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 $\boldsymbol{\mu}$.

10.2.3 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,

$$\begin{aligned}
\mathbb{E}\big(\widehat{\boldsymbol{\beta}}_{XV}\,\big|\,\mathbb{Z}\big) &= \boldsymbol{\beta},\\
\mathbb{E}\big(\widehat{\boldsymbol{Y}}_{XV}\,\big|\,\mathbb{Z}\big) &= \mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma} =: \boldsymbol{\mu},\\
\sum_{i=1}^{n}\mathsf{MSE}\big(\widehat{Y}_{XV,i}\big) &= \sum_{i=1}^{n}\mathsf{var}\big(\widehat{Y}_{XV,i}\,\big|\,\mathbb{Z}\big) \;=\; \mathsf{tr}\Big(\mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV}\,\big|\,\mathbb{Z}\big)\Big) \;=\; \mathsf{tr}\big(\sigma^2\,\mathbb{H}_{XV}\big) \qquad (10.11)\\
&= \sigma^2\,(k+l),\\
\mathbb{E}\big(\mathsf{MS}_{e,XV}\,\big|\,\mathbb{Z}\big) &= \sigma^2.
\end{aligned}$$

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.

Lemma 10.5 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 $\mathbb{Z}$.

$\mathbb{E}\big(\widehat{\boldsymbol{\beta}}_X \,\big|\, \mathbb{Z}\big)$

By Lemma 9.1: $\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X = -\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}$.

$$\begin{aligned}\text{Hence,}\quad \mathbb{E}\big(\widehat{\boldsymbol{\beta}}_X\big) &= \mathbb{E}\Big\{\widehat{\boldsymbol{\beta}}_{XV} + \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}\Big\}\\ &= \boldsymbol{\beta} + \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma},\\ \mathsf{bias}\big(\widehat{\boldsymbol{\beta}}_X\big) &= \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma}.\end{aligned}$$

$\mathbb{E}\big(\widehat{\boldsymbol{Y}}_X \,\big|\, \mathbb{Z}\big)$

By Lemma 9.1: $\widehat{\boldsymbol{Y}}_{XV} - \widehat{\boldsymbol{Y}}_X = \mathbb{X}\big(\widehat{\boldsymbol{\beta}}_{XV} - \widehat{\boldsymbol{\beta}}_X\big) + \mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}$.

$$\begin{aligned}\text{Hence,}\quad \mathbb{E}\big(\widehat{\boldsymbol{Y}}_X\big) &= \mathbb{E}\big(\widehat{\boldsymbol{Y}}_{XV} - \mathbb{X}\widehat{\boldsymbol{\beta}}_{XV} + \mathbb{X}\widehat{\boldsymbol{\beta}}_X - \mathbb{V}\widehat{\boldsymbol{\gamma}}_{XV}\big)\\ &= \boldsymbol{\mu} - \mathbb{X}\boldsymbol{\beta} + \mathbb{X}\boldsymbol{\beta} + \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\mathbb{V}\boldsymbol{\gamma} - \mathbb{V}\boldsymbol{\gamma}\\ &= \boldsymbol{\mu} + \Big\{\mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top - \mathbf{I}_n\Big\}\mathbb{V}\boldsymbol{\gamma}\\ &= \boldsymbol{\mu} - \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma},\\ \mathsf{bias}\big(\widehat{\boldsymbol{Y}}_X\big) &= -\,\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}.\end{aligned}$$

$\sum_{i=1}^n \mathsf{MSE}\big(\widehat{Y}_{X,i}\big)$

Let us first calculate $\mathsf{MSE}\big(\widehat{\boldsymbol{Y}}_X\big) = \mathbb{E}\Big\{\big(\widehat{\boldsymbol{Y}}_X - \boldsymbol{\mu}\big)\big(\widehat{\boldsymbol{Y}}_X - \boldsymbol{\mu}\big)^\top\Big\}$:

$$\begin{aligned}\mathsf{MSE}\big(\widehat{\boldsymbol{Y}}_X\big) &= \mathsf{var}\big(\widehat{\boldsymbol{Y}}_X\big) + \mathsf{bias}\big(\widehat{\boldsymbol{Y}}_X\big)\mathsf{bias}^\top\big(\widehat{\boldsymbol{Y}}_X\big)\\ &= \sigma^2\mathbb{H}_X + \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X.\end{aligned}$$

$$\begin{aligned}\text{Hence,}\quad \sum_{i=1}^n \mathsf{MSE}\big(\widehat{Y}_{X,i}\big) &= \mathsf{tr}\Big(\mathsf{MSE}\big(\widehat{\boldsymbol{Y}}_X\big)\Big)\\ &= \mathsf{tr}\big(\sigma^2\mathbb{H}_X + \mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\big)\\ &= \mathsf{tr}\big(\sigma^2\mathbb{H}_X\big) + \mathsf{tr}\big(\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\big)\\ &= \sigma^2\,k + \mathsf{tr}\big(\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big)\\ &= \sigma^2\,k + \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2.\end{aligned}$$

$\mathbb{E}\big(\mathsf{MS}_{e,X} \,\big|\, \mathbb{Z}\big)$

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 $\mathbb{E}\big(\mathsf{SS}_{e,X}\big) := \mathbb{E}\big(\mathsf{SS}_{e,X} \,\big|\, \mathbb{Z}\big)$. To do that, write the linear model M_{XV} using the error terms as

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \mathbf{0}_n, \quad \mathsf{var}\big(\boldsymbol{\varepsilon} \,\big|\, \mathbb{Z}\big) = \sigma^2\mathbf{I}_n.$$

$$
\begin{aligned}
\mathbb{E}\big(\mathsf{SS}_{e,X}\big) &= \mathbb{E}\big\|\mathbb{M}_X \boldsymbol{Y}\big\|^2 = \mathbb{E}\big\|\mathbb{M}_X(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})\big\|^2 \\
&= \mathbb{E}\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma} + \mathbb{M}_X\boldsymbol{\varepsilon}\big\|^2 \\
&= \mathbb{E}\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2 \;+\; \mathbb{E}\big\|\mathbb{M}_X\boldsymbol{\varepsilon}\big\|^2 \;+\; 2\,\underbrace{\mathbb{E}\big(\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\mathbb{M}_X\boldsymbol{\varepsilon}\big)}_{\boldsymbol{\gamma}^\top\mathbb{V}^\top\mathbb{M}_X\,\mathbb{E}\boldsymbol{\varepsilon}=0} \\
&= \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2 \;+\; \underbrace{\mathbb{E}\big(\boldsymbol{\varepsilon}^\top\mathbb{M}_X\boldsymbol{\varepsilon}\big)}_{\mathbb{E}\left(\mathsf{tr}(\boldsymbol{\varepsilon}^\top\mathbb{M}_X\boldsymbol{\varepsilon})\right)=\mathsf{tr}\left(\mathbb{E}(\mathbb{M}_X\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\right)=\mathsf{tr}\left(\sigma^2\,\mathbb{M}_X\right)=\sigma^2\,(n-k)} \\
&= \big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2 \;+\; \sigma^2\,(n-k).
\end{aligned}
$$

Hence,
$$
\begin{aligned}
\mathbb{E}\big(\mathsf{MS}_{e,X}\big) &= \mathbb{E}\left(\frac{\mathsf{SS}_{e,X}}{n-k}\right) \\
&= \sigma^2 \;+\; \frac{\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2}{n-k}, \\
\mathsf{bias}\big(\mathsf{MS}_{e,X}\big) &= \frac{\big\|\mathbb{M}_X\mathbb{V}\boldsymbol{\gamma}\big\|^2}{n-k}.
\end{aligned}
$$

❑

Least squares estimators

Lemma 10.5 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) $\underline{\mathbb{X}^\top\mathbb{V} = \mathbf{0}_{k\times l}}$, 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) $\underline{\mathbb{X}^\top\mathbb{V} \neq \mathbf{0}_{k\times l}}$

- $\widehat{\boldsymbol{\beta}}_X$ 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 (10.11)) together with results of Lemma 10.4 and Lemma 10.5 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

Lemma 10.5 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 $n \to \infty$.

10.2.4 Irrelevant regressors

In this section, we will assume that the correct model is model

$$\mathsf{M}_X:\quad \boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big).$$

This means, that also model

$$\mathsf{M}_{XV}:\quad \boldsymbol{Y}\,\big|\,\mathbb{Z} \sim \big(\mathbb{X}\boldsymbol{\beta} + \mathbb{V}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}_n\big)$$

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,

$$\begin{aligned}\mathbb{E}\big(\widehat{\boldsymbol{\beta}}_X \,\big|\, \mathbb{Z}\big) = \mathbb{E}\big(\widehat{\boldsymbol{\beta}}_{XV} \,\big|\, \mathbb{Z}\big) &= \boldsymbol{\beta},\\ \mathbb{E}\big(\widehat{\boldsymbol{Y}}_X \,\big|\, \mathbb{Z}\big) = \mathbb{E}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big) &= \mathbb{X}\boldsymbol{\beta} =: \boldsymbol{\mu},\\ \sum_{i=1}^{n}\mathsf{MSE}\big(\widehat{Y}_{X,i}\big) &= \sum_{i=1}^{n}\mathsf{var}\big(\widehat{Y}_{X,i}\,\big|\,\mathbb{Z}\big) \;=\; \mathsf{tr}\Big(\mathsf{var}\big(\widehat{\boldsymbol{Y}}_X \,\big|\, \mathbb{Z}\big)\Big) \;=\; \mathsf{tr}\big(\sigma^2\,\mathbb{H}_X\big)\\ &= \sigma^2\,k,\\ \sum_{i=1}^{n}\mathsf{MSE}\big(\widehat{Y}_{XV,i}\big) &= \sum_{i=1}^{n}\mathsf{var}\big(\widehat{Y}_{XV,i}\,\big|\,\mathbb{Z}\big) \;=\; \mathsf{tr}\Big(\mathsf{var}\big(\widehat{\boldsymbol{Y}}_{XV} \,\big|\, \mathbb{Z}\big)\Big) \;=\; \mathsf{tr}\big(\sigma^2\,\mathbb{H}_{XV}\big)\\ &= \sigma^2\,(k+l),\\ \mathbb{E}\big(\mathsf{MS}_{e,X} \,\big|\, \mathbb{Z}\big) = \mathbb{E}\big(\mathsf{MS}_{e,XV} \,\big|\, \mathbb{Z}\big) &= \sigma^2.\end{aligned}$$

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 Lemma 10.3, their quality expressed by the mean squared error which in this case coincide with the covariance matrix (may) differ since

$$\mathsf{MSE}\big(\widehat{\boldsymbol{\beta}}_{XV}\big) - \mathsf{MSE}\big(\widehat{\boldsymbol{\beta}}_{X}\big) = \mathbb{E}\Big\{\big(\widehat{\boldsymbol{\beta}}_{XV}-\boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}}_{XV}-\boldsymbol{\beta}\big)^\top \,\Big|\, \mathbb{Z}\Big\} - \mathbb{E}\Big\{\big(\widehat{\boldsymbol{\beta}}_{X}-\boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}}_{X}-\boldsymbol{\beta}\big)^\top \,\Big|\, \mathbb{Z}\Big\}$$

$$= \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV}\,\big|\,\mathbb{Z}\big) - \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X}\,\big|\,\mathbb{Z}\big) \geq 0.$$

In particular, we derived during the proof of Lemma 10.3 that

$$\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV}\,\big|\,\mathbb{Z}\big) - \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X}\,\big|\,\mathbb{Z}\big) = \sigma^2\Big[\Big\{\mathbb{X}^\top\mathbb{X} - \mathbb{X}^\top\mathbb{V}\big(\mathbb{V}^\top\mathbb{V}\big)^{-1}\mathbb{V}^\top\mathbb{X}\Big\}^{-1} - \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\Big].$$

Let us again consider two situations.

(i) $\underline{\mathbb{X}^\top\mathbb{V} = \mathbf{0}_{k\times l}}$, 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 $\mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{X}\,\big|\,\mathbb{Z}\big) = \mathsf{var}\big(\widehat{\boldsymbol{\beta}}_{XV}\,\big|\,\mathbb{Z}\big)$.
- 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) $\underline{\mathbb{X}^\top\mathbb{V} \neq \mathbf{0}_{k\times l}}$

- 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 10.4 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}.

10.2.5 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 10.2.3 and 10.2.4.

(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 10.2.3 and 10.2.4 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 11

Unusual Observations

In this chapter, we develop tools for identification of observations which are in a certain sense unusual with respect to the assumed linear model. First, in Section 11.2, we shall deal with so called *outliers* which are observations with unusual response values. Second, in Section 11.3, we shall talk about so called *leverage points*, which are observations with unusual covariate values. Finally, in Section 11.4, we discuss tools for identification of those observations which, in a certain sense, might have harmful influence on statistical inference based on the considered model.

In the whole chapter, we assume a full-rank linear model

$$\mathsf{M}: \ \boldsymbol{Y}\,\big|\,\mathbb{X} \sim \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big), \quad \mathsf{rank}(\mathbb{X}_{n\times k}) = k,$$

where standard notation is considered. That is,

- $\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y} = \big(\widehat{\beta}_0,\,\ldots,\,\widehat{\beta}_{k-1}\big)^\top$: LSE of the vector $\boldsymbol{\beta}$;
- $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top = \big(h_{i,t}\big)_{i,t=1,\ldots,n}$: the hat matrix;
- $\mathbb{M} = \mathbf{I}_n - \mathbb{H} = \big(m_{i,t}\big)_{i,t=1,\ldots,n}$: the residual projection matrix;
- $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \big(\widehat{Y}_1,\,\ldots,\,\widehat{Y}_n\big)^\top$: 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;
- $\mathsf{MS}_e = \frac{1}{n-k}\,\mathsf{SS}_e$ 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 identification 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.,

$$\mathbb{X} = \begin{pmatrix}\boldsymbol{X}_1^\top\\ \vdots\\ \boldsymbol{X}_n^\top\end{pmatrix} = \begin{pmatrix}\boldsymbol{x}_1^\top\\ \vdots\\ \boldsymbol{x}_n^\top\end{pmatrix}.$$

11.1 Leave-one-out and outlier model

Notation. For chosen $t \in \{1, \dots, n\}$, 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 $\bigl(0, \dots, 0, 1, 0, \dots, 0\bigr)^\top$ of length n with 1 on the tth place.

Definition 11.1 Leave-one-out model.
The tth leave-one-out model[1] *is a linear model*

$$\mathsf{M}_{(-t)}: \quad \boldsymbol{Y}_{(-t)} \,\big|\, \mathbb{X}_{(-t)} \sim \bigl(\mathbb{X}_{(-t)}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_{n-1}\bigr).$$

Definition 11.2 Outlier model.
The tth outlier model[2] *is a linear model*

$$\mathsf{M}_t^{out}: \quad \boldsymbol{Y} \,\big|\, \mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{j}_t\gamma_t^{out},\, \sigma^2\mathbf{I}_n\bigr).$$

Lemma 11.1 Three equivalent statements.
While assuming $\mathsf{rank}(\mathbb{X}_{n\times k}) = k$, the following three statements are equivalent:

(i) $\mathsf{rank}(\mathbb{X}) = \mathsf{rank}\bigl(\mathbb{X}_{(-t)}\bigr) = k$, *i.e.,* $\boldsymbol{x}_t \in \mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr)$*;*

(ii) $m_{t,t} > 0$*;*

(iii) $\mathsf{rank}\bigl(\mathbb{X},\, \boldsymbol{j}_t\bigr) = k + 1$.

Proof. **Proof/calculations were skipped and are not requested for the exam.**

- We will proof the lemma by showing non(i) ⇔ non(ii) ⇔ non(iii).
- non(i) means that $\boldsymbol{x}_t \notin \mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr) \subset \mathcal{M}\bigl(\mathbb{X}^\top\bigr)$.

$$\mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr) \subset \mathcal{M}\bigl(\mathbb{X}^\top\bigr) \quad \text{and}\ \mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr) \neq \mathcal{M}\bigl(\mathbb{X}^\top\bigr).$$

$$\Leftrightarrow \mathcal{M}\bigl(\mathbb{X}^\top\bigr)^\perp \subset \mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr)^\perp \quad \text{and}\ \mathcal{M}\bigl(\mathbb{X}^\top\bigr)^\perp \neq \mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr)^\perp.$$

- That is, $\Leftrightarrow \exists \boldsymbol{a} \in \mathcal{M}\bigl(\mathbb{X}_{(-t)}^\top\bigr)^\perp$ such that $\boldsymbol{a} \notin \mathcal{M}\bigl(\mathbb{X}^\top\bigr)^\perp$.

$$\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^k \ \text{such that}\ \boldsymbol{a}^\top\mathbb{X}_{(-t)}^\top = \mathbf{0}^\top \ \ \&\ \ \boldsymbol{a}^\top\mathbb{X}^\top \neq \mathbf{0}^\top.$$

$$\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^k \ \text{such that}\ \mathbb{X}_{(-t)}\boldsymbol{a} = \mathbf{0} \ \ \&\ \ \mathbb{X}\boldsymbol{a} \neq \mathbf{0}.$$

It must be

$$\mathbb{X}\boldsymbol{a} = \bigl(0, \dots, 0, c, 0, \dots, 0\bigr)^\top = c\,\boldsymbol{j}_t$$

for some $c \neq 0$.

[1] *model vynechaného tého pozorování* [2] *model tého odlehlého pozorování*

$$\Leftrightarrow \exists \boldsymbol{a} \in \mathbb{R}^k \ \text{ such that } \mathbb{X}\boldsymbol{a} = c\boldsymbol{j}_t, \ c \neq 0.$$

$$\Leftrightarrow \boldsymbol{j}_t \in \mathcal{M}\big(\mathbb{X}\big) \qquad \Leftrightarrow \text{non(iii)}$$

$$\Leftrightarrow \qquad \underbrace{\mathbb{M}\boldsymbol{j}_t}_{t\text{th column of } \mathbb{M}} \qquad = \mathbf{0}.$$

$$\Leftrightarrow \boldsymbol{m}_t = \mathbf{0}.$$

$$\Leftrightarrow \big\|\boldsymbol{m}_t\big\|^2 = m_{t,t} = 0.$$

$$\Leftrightarrow \text{non(ii)}.$$

$\boldsymbol{m}_t$ denotes the tth row of $\mathbb{M}$ (and also its t column since $\mathbb{M}$ is symmetric).

❑

Note. Under the assumption of either (i), (ii) or (iii) of Lemma 11.1, both the leave-one-out and the outlier model are full rank models.

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.,

$$\widehat{\boldsymbol{\beta}}_{(-t)}, \widehat{\boldsymbol{Y}}_{(-t)}, \mathsf{SS}_{e,(-t)}, \mathsf{MS}_{e,(-t)}, \ldots$$

- Quantities related to model M_t^{out} will be recognized by subscript t and superscript out, i.e.,

$$\widehat{\boldsymbol{\beta}}_t^{out}, \widehat{\boldsymbol{Y}}_t^{out}, \mathsf{SS}_{e,t}^{out}, \mathsf{MS}_{e,t}^{out}, \ldots$$

- Solutions to normal equations in model M_t^{out} (the LSE of $\big((\boldsymbol{\beta}_t^{out})^\top, \gamma_t^{out}\big)^\top$) will be denoted as

$$\big((\widehat{\boldsymbol{\beta}}_t^{out})^\top, \widehat{\gamma}_t^{out}\big)^\top.$$

Lemma 11.2 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* $\widehat{\boldsymbol{\beta}}_{(-t)}$ *solves the normal equations of model* $\mathsf{M}_{(-t)}$ *if and only if a vector* $\big((\widehat{\boldsymbol{\beta}}_t^{out})^\top, \widehat{\gamma}_t^{out}\big)^\top$ *solves the normal equations of model* M_t^{out}*, where*

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_t^{out} &= \widehat{\boldsymbol{\beta}}_{(-t)}, \\ \widehat{\gamma}_t^{out} &= Y_t - \boldsymbol{x}_t{}^\top \widehat{\boldsymbol{\beta}}_{(-t)}. \end{aligned}$$

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

$$\begin{aligned}\mathsf{SS}_t^{out}(\boldsymbol{\beta},\,\gamma_t^{out}) &= \big\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} - \boldsymbol{j}_t\gamma_t^{out}\big\|^2 && \text{separate the } t\text{th element of the sum}\\ &= \big\|\boldsymbol{Y}_{(-t)} - \mathbb{X}_{(-t)}\boldsymbol{\beta}\big\|^2 \;+\; \big(Y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta} - \gamma_t^{out}\big)^2\\ &= \mathsf{SS}_{(-t)}(\boldsymbol{\beta}) \;+\; \big(Y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta} - \gamma_t^{out}\big)^2,\end{aligned}$$

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 $\widehat{\boldsymbol{\beta}}_{(-t)} \in \mathbb{R}^k$ minimizes $\mathsf{SS}_{(-t)}(\boldsymbol{\beta})$ if and only if a vector

$$\Big(\underbrace{\widehat{\boldsymbol{\beta}}_{(-t)}^\top}_{\widehat{\boldsymbol{\beta}}_t^{out}},\;\underbrace{Y_t - \boldsymbol{x}_t^\top\widehat{\boldsymbol{\beta}}_{(-t)}}_{\widehat{\gamma}_t^{out}}\Big)^\top \in \mathbb{R}^{k+1}$$

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:

$$\widehat{Y}_{[t]} := \boldsymbol{x}_t^\top\widehat{\boldsymbol{\beta}}_{(-t)}, \quad t = 1, \ldots, n,$$

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 9.1, 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,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_t^{out} &= \widehat{\boldsymbol{\beta}} - \frac{U_t}{m_{t,t}} \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \boldsymbol{x}_t, \\
\widehat{\boldsymbol{Y}}_t^{out} &= \widehat{\boldsymbol{Y}} + \frac{U_t}{m_{t,t}} \boldsymbol{m}_t, \\
\mathsf{SS}_e - \mathsf{SS}_{e,t}^{out} &= \frac{U_t^2}{m_{t,t}} = \mathsf{MS}_e \left(U_t^{std}\right)^2,
\end{aligned}$$

where $\boldsymbol{m}_t$ denotes the tth column (and row as well) of the residual projection matrix $\mathbb{M}$.

Lemma 11.3 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.

$$\begin{aligned}
& & \widehat{\gamma}_t^{out} &= Y_t - \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_{(-t)} = Y_t - \widehat{Y}_{[t]} = \frac{U_t}{m_{t,t}}, \\
\widehat{\boldsymbol{\beta}}_{(-t)} &= \widehat{\boldsymbol{\beta}}_t^{out} & &= \widehat{\boldsymbol{\beta}} - \frac{U_t}{m_{t,t}} \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \boldsymbol{x}_t, \\
\mathsf{SS}_{e,(-t)} &= \mathsf{SS}_{e,t}^{out} & &= \mathsf{SS}_e - \frac{U_t^2}{m_{t,t}} = \mathsf{SS}_e - \mathsf{MS}_e \left(U_t^{std}\right)^2, \\
\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e} &= \frac{\mathsf{MS}_{e,t}^{out}}{\mathsf{MS}_e} & &= \frac{n-k-\left(U_t^{std}\right)^2}{n-k-1}.
\end{aligned} \tag{11.1}$$

Proof. Equality between the quantities of the outlier and the leave-one-out model follows from Lemma 11.2. Remaining expressions follow from previously conducted calculations.

To see the last equality in (11.1), remember that the residual degrees of freedom of both the outlier and the leave-one-out models are equal to $n-k-1$. That is, whereas in model M,

$$\mathsf{MS}_e = \frac{\mathsf{SS}_e}{n-k},$$

in the outlier and the leave-one-out model,

$$\mathsf{MS}_{e,(-t)} = \frac{\mathsf{SS}_{e,(-t)}}{n-k-1} = \frac{\mathsf{SS}_{e,t}^{out}}{n-k-1} = \mathsf{MS}_{e,t}^{out}.$$

❑

Notes.

- Expressions in Lemma 11.3 quantify the influence of the tth observation on
  - (i) the LSE of a vector $\boldsymbol{\beta}$ of the regression coefficients;
  - (ii) the estimate of the residual variance.
- Lemma 11.3 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 11.3 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.

11.2 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 $m_{t,t} > 0$, there exist the least squares estimator of the parameter γ_t^{out} in 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

$$\mathrm{H}_0 : \ \gamma_t^{out} = 0$$

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 regression 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 $n-k-1$ degrees of freedom (residual degrees of freedom of the outlier model).

From Section 11.1, 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),

$$\mathsf{var}\big(\widehat{\gamma}_t^{out} \,\big|\, \mathbb{X}\big) \;=\; \mathsf{var}\bigg(\frac{U_t}{m_{t,t}} \,\bigg|\, \mathbb{X}\bigg) \;=\; \frac{1}{m_{t,t}^2}\,\mathsf{var}\big(U_t \,\big|\, \mathbb{X}\big) \;\overset{(\star)}{=}\; \frac{1}{m_{t,t}^2}\,\sigma^2\, m_{t,t} \;=\; \frac{\sigma^2}{m_{t,t}}.$$

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,

$$T_t = \frac{\widehat{\gamma}_t^{out}}{\sqrt{\frac{\mathsf{MS}_{e,t}^{out}}{m_{t,t}}}}.$$

Two useful expressions of the statistic T_t are obtained by remembering from Section 11.1 (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 11.4 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 t*th* studentized residual[4] *of the model* M.

Notes.

- Using the last equality in (11.1), 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,
$$T_t = \sqrt{\frac{n-k-1}{n-k-\bigl(U_t^{std}\bigr)^2}}\;U_t^{std}.$$
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.

Lemma 11.4 On studentized residuals.

Let $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\mathbf{I}_n\bigr)$*, where* $\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k < n$*. Let further* $n > k+1$*. Let for given* $t \in \bigl\{1,\dots,n\bigr\}$ $m_{t,t} > 0$*. Then*

1. *The* t*th studentized residual* T_t *follows the Student t-distribution with* $n-k-1$ *degrees of freedom.*
2. *If additionally* $n > k+2$ *then* $\mathbb{E}\bigl(T_t\bigr) = 0$*.*
3. *If additionally* $n > k+3$ *then* $\mathsf{var}\bigl(T_t\bigr) = \dfrac{n-k-1}{n-k-3}$*.*

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 $\mathtt{t}_{n-k-1}$ distribution under the null hypothesis) of the test

H$_0$: $\gamma_t^{out} = 0$,

H$_1$: $\gamma_t^{out} \neq 0$

in the tth outlier model M_t^{out}: $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{j}_t\gamma_t^{out},\,\sigma^2\mathbf{I}_n\bigr)$.

The above testing problem can also be interpreted as a test of

[4] *studentizované reziduum*

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

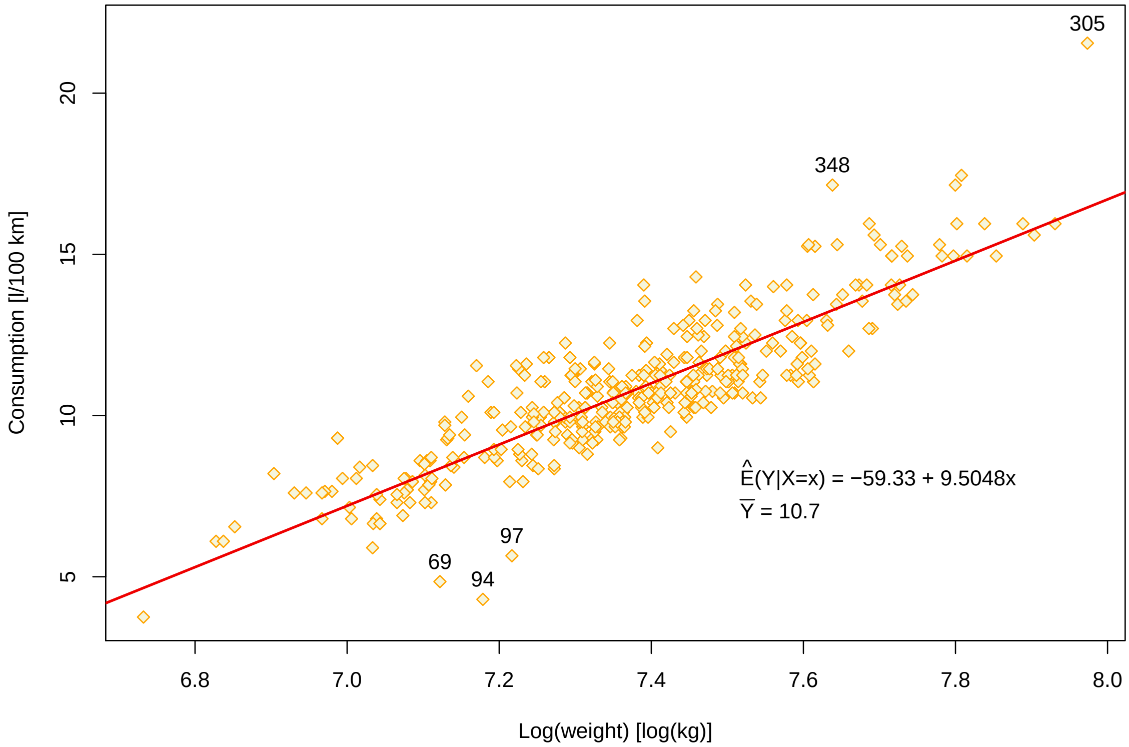
- The expected value of the tth observation is different from that given by model M;
- The observed value of Y_t is unusual under model M.

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-k-1} distribution are needed to keep the rate of falsely identified outliers under the requested level of α (see Chapter 14 for more details concerning the multiple testing problems). For example, Bonferroni adjustment can be used.

Illustrations

Cars2004 (subset, $n = 412$), `consumption` $\sim$ `log(weight)`

Observations with five highest absolute values of studentized residuals



Cars2004 (subset, $n = 412$), `consumption` $\sim$ `log(weight)`

Standardized, studentized and deleted residuals

Standardized residuals $U_1^{std}, \dots, U_n^{std}$

```
m1 <- lm(consumption ~ lweight, data = CarsUsed)
rstandard(m1)
```

```
          1           2            3            4            5            6
0.600003668  0.683558025 -0.237013632 -0.437157041 -0.237013632 -0.491068598  ...
```

Studentized residuals $T_1, \dots, T_n$

```
rstudent(m1)
```

```
          1           2            3            4            5            6
0.599534780  0.683113271 -0.236740634 -0.436725391 -0.236740634 -0.490613671  ...
```

Deleted residuals $\widehat{\gamma}_1^{out}, \dots, \widehat{\gamma}_n^{out}$

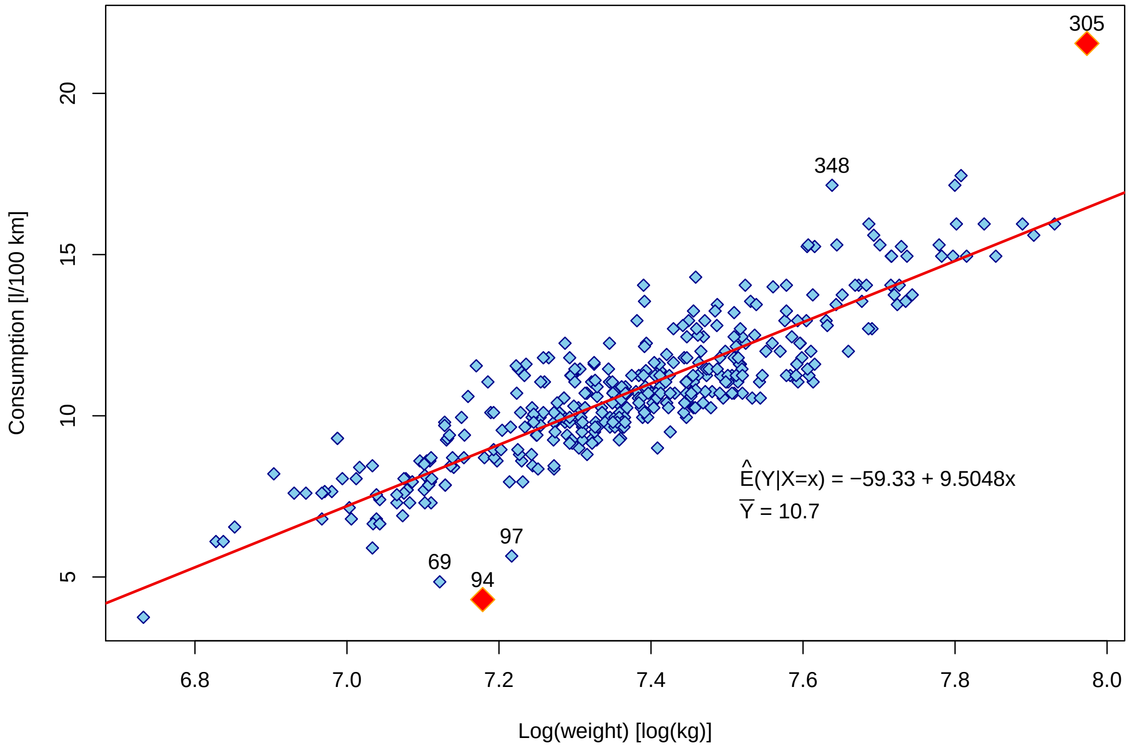
```
residuals(m1) / (1 - hatvalues(m1))
```

```
          1           2            3            4            5            6
0.646454917  0.736641641 -0.254845546 -0.469869858 -0.254845546 -0.528142442  ...
```

Illustrations

Cars2004 (subset, $n = 412$), consumption ∼ log(weight)

Identified outliers



Cars2004 (subset, $n = 412$), consumption ∼ log(weight)

Observations with five highest absolute values of studentized residuals

```
                                vname fhybrid consumption  lweight weight
305 Hummer.H2                              No       21.55 7.973500   2903
94  Toyota.Prius.4dr.(gas/electric)       Yes        4.30 7.178545   1311
348 Land.Rover.Discovery.SE                No       17.15 7.638198   2076
97  Volkswagen.Jetta.GLS.TDI.4dr           No        5.65 7.216709   1362
69  Honda.Civic.Hybrid                    Yes        4.85 7.122060   1239
    .4dr.manual.(gas/electric)

                                vname     gamma         Tt     PvalUnadj  PvalBonf
305 Hummer.H2                          5.223712   4.953073      0.000001  0.000441
94  Toyota.Prius.4dr.(gas/electric)   -4.618542  -4.396641      0.000014  0.005782
348 Land.Rover.Discovery.SE            3.910233   3.693509      0.000251  0.103499
97  Volkswagen.Jetta.GLS.TDI.4dr      -3.623890  -3.420244      0.000689  0.283692
69  Honda.Civic.Hybrid                -3.531883  -3.327145      0.000957  0.394186
    .4dr.manual.(gas/electric)
```

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)?
  - $\vdots$
- ***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.

11.3 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 $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top$ 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 $\mathbb{H} = \mathbb{X}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top$ can also be calculated using the matrix $\big(\mathbf{1},\, \widetilde{\mathbb{X}}\big)$, where we can use additional property $\mathbf{1}_n^\top\widetilde{\mathbb{X}} = \mathbf{0}_{k-1}^\top$:

$$\mathbb{H} = \big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\Big\{\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)^\top\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\Big\}^{-1}\big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)^\top$$

$$= \big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\begin{pmatrix} \underbrace{\mathbf{1}_n^\top\mathbf{1}_n}_{n} & \underbrace{\mathbf{1}_n^\top\widetilde{\mathbb{X}}}_{\mathbf{0}_{k-1}^\top} \\ \underbrace{\widetilde{\mathbb{X}}^\top\mathbf{1}_n}_{\mathbf{0}_{k-1}} & \widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{1}_n^\top \\ \widetilde{\mathbb{X}}^\top \end{pmatrix}$$

$$= \big(\mathbf{1}_n,\, \widetilde{\mathbb{X}}\big)\begin{pmatrix} \frac{1}{n} & \mathbf{0}_{k-1}^\top \\ \mathbf{0}_{k-1} & \big(\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}}\big)^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{1}_n^\top \\ \widetilde{\mathbb{X}}^\top \end{pmatrix}$$

[5] *vzdálená pozorování*

$$= \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top + \widetilde{\mathbb{X}}\bigl(\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}}\bigr)^{-1}\widetilde{\mathbb{X}}^\top.$$

That is, the tth leverage equals

$$h_{t,t} \;=\; \frac{1}{n} + \bigl(x_{t,1} - \overline{x}^1,\;\ldots,\;x_{t,k-1} - \overline{x}^{k-1}\bigr)\bigl(\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}}\bigr)^{-1}\bigl(x_{t,1} - \overline{x}^1,\;\ldots,\;x_{t,k-1} - \overline{x}^{k-1}\bigr)^\top.$$

The second term is then a square of the generalized distance between the non-intercept regressors $\bigl(x_{t,1},\ldots,x_{t,k-1}\bigr)^\top$ 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

$$\sum_{i=1}^{n} h_{i,i} = \mathsf{tr}(\mathbb{H}) = \mathsf{tr}\bigl(\mathbb{Q}\mathbb{Q}^\top\bigr) = \mathsf{tr}\bigl(\mathbb{Q}^\top\mathbb{Q}\bigr) = \mathsf{tr}(\mathbf{I}_k) = k.$$

That is,

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{i,i} = \frac{k}{n}. \tag{11.2}$$

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 (11.2), a reasonable bound is a value higher than $\frac{k}{n}$. For example, the R function `influence.measures` marks the tth observation as a leverage point if

$$h_{t,t} > \frac{3\,k}{n}.$$

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:

$$\mathsf{var}\bigl(U_t \,\big|\, \mathbb{X}\bigr) = \sigma^2\, m_{t,t} = \sigma^2\,(1 - h_{t,t}), \qquad t = 1,\ldots,n.$$

Remind that $U_t = Y_t - \widehat{Y}_t$ and hence also

$$\mathsf{var}\bigl(Y_t - \widehat{Y}_t \,\big|\, \mathbb{X}\bigr) = \sigma^2\,(1 - h_{t,t}), \qquad t = 1,\ldots,n.$$

That is, $\mathsf{var}\bigl(U_t \,\big|\, \mathbb{X}\bigr) = \mathsf{var}\bigl(Y_t - \widehat{Y}_t \,\big|\, \mathbb{X}\bigr)$ 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.

Illustrations

Cars2004 (subset, $n = 412$), consumption $\sim$ log(weight)

Leverages and influence measures

Leverages $h_{1,1}, \ldots, h_{n,n}$

```
m1 <- lm(consumption ~ lweight, data = CarsUsed)
hatvalues(m1)
```

```
          1           2           3           4           5           6
0.011453373 0.011892770 0.007436292 0.006688146 0.007436292 0.007916965    ...
```

Influence measures

```
influence.measures(m1)
```

```
Influence measures of
         lm(formula = consumption ~ lweight, data = CarsUsed) :

       dfb.1_  dfb.lwgh     dffit cov.r   cook.d     hat inf
1    5.81e-02 -5.73e-02  0.064533 1.015 2.09e-03 0.01145   *
2    6.78e-02 -6.69e-02  0.074943 1.015 2.81e-03 0.01189   *
3   -1.71e-02  1.68e-02 -0.020491 1.012 2.10e-04 0.00744
4   -2.92e-02  2.86e-02 -0.035836 1.011 6.43e-04 0.00669
5   -1.71e-02  1.68e-02 -0.020491 1.012 2.10e-04 0.00744
6   -3.71e-02  3.65e-02 -0.043827 1.012 9.62e-04 0.00792
7   -4.59e-02  4.50e-02 -0.055070 1.010 1.52e-03 0.00732
8    7.70e-03 -7.56e-03  0.009196 1.012 4.24e-05 0.00749
9   -2.15e-02  2.11e-02 -0.025596 1.012 3.28e-04 0.00758
...
```

Cars2004 (subset, $n = 412$), consumption $\sim$ log(weight)

Potentially influential observations

```
summary(influence.measures(m1))
```

```
Potentially influential observations of
         lm(formula = consumption ~ lweight, data = CarsUsed) :

    dfb.1_ dfb.lwgh dffit   cov.r   cook.d hat
1    0.06  -0.06     0.06    1.01_*  0.00   0.01
2    0.07  -0.07     0.07    1.01_*  0.00   0.01
17   0.07  -0.07     0.07    1.01_*  0.00   0.01
39  -0.01   0.01    -0.01    1.02_*  0.00   0.01
47   0.07  -0.07     0.07    1.02_*  0.00   0.02_*
48   0.09  -0.09     0.10    1.02_*  0.00   0.02_*
49   0.06  -0.06     0.06    1.02_*  0.00   0.02_*
69  -0.21   0.20    -0.26_*  0.96_*  0.03   0.01
70  -0.14   0.14    -0.14    1.03_*  0.01   0.03_*
94  -0.21   0.20    -0.30_*  0.92_*  0.04   0.00
97  -0.13   0.13    -0.21_*  0.95_*  0.02   0.00
204 -0.05   0.06     0.14    0.98_*  0.01   0.00
270  0.20  -0.20     0.22_*  0.99    0.02   0.01
271  0.20  -0.20     0.22_*  0.99    0.02   0.01
278  0.05  -0.04     0.12    0.98_*  0.01   0.00
294  0.21  -0.21     0.23_*  1.00    0.03   0.02_*
295 -0.02   0.02     0.02    1.02_*  0.00   0.01
301  0.00   0.00    -0.01    1.02_*  0.00   0.01
302  0.00   0.00     0.00    1.01_*  0.00   0.01
...
```

Illustrations

Cars2004 (subset, $n = 412$), consumption ~ log(weight)

Leverage points

$\frac{3\,k}{n} = 0.0146$

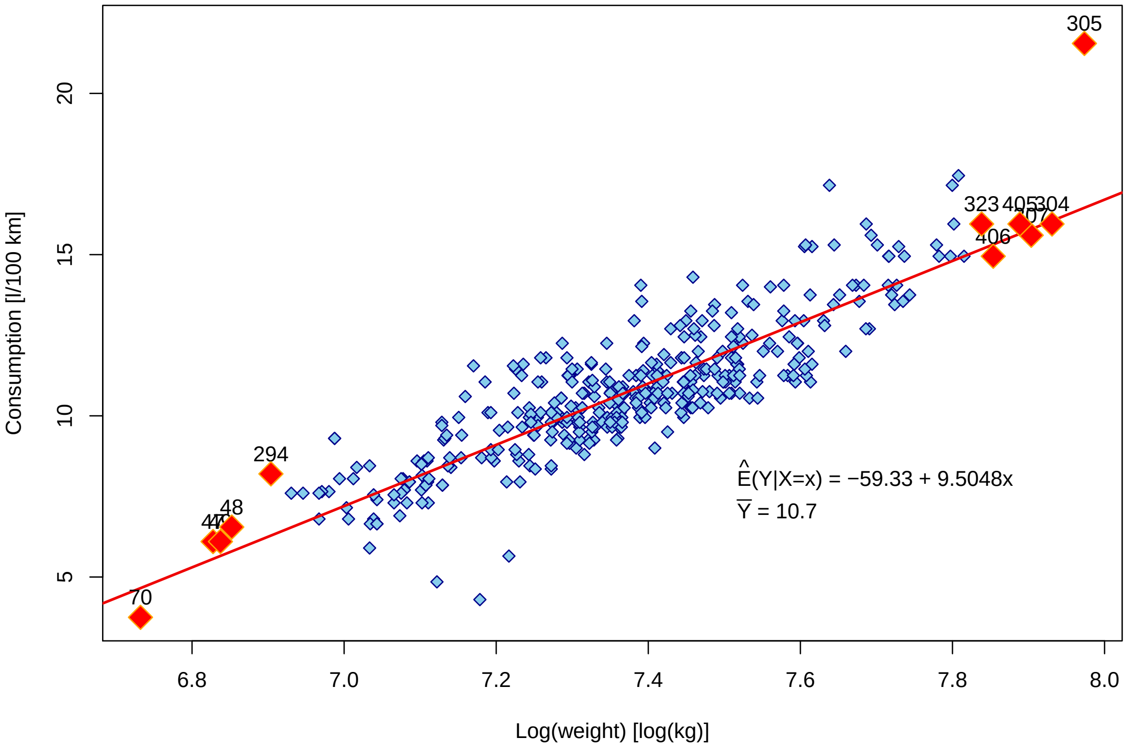
```
sum(hatvalues(m1) > 3 * k / n)
```

```
[1] 11
```

```
                                vname consumption weight  lweight          h
47              Toyota.Echo.2dr.manual        6.10    923 6.827629 0.01992471
48                Toyota.Echo.2dr.auto        6.55    946 6.852243 0.01836889
49                     Toyota.Echo.4dr        6.10    932 6.837333 0.01930270
70    Honda.Insight.2dr.(gas/electric)        3.75    839 6.732211 0.02664081
294 Toyota.MR2.Spyder.convertible.2dr         8.20    996 6.903747 0.01534760
304                GMC.Yukon.XL.2500.SLT     15.95   2782 7.930925 0.02132481
305                             Hummer.H2     21.55   2903 7.973500 0.02429502
307             Lincoln.Navigator.Luxury     15.60   2707 7.903596 0.01953240
323                         Lexus.LX.470     15.95   2536 7.838343 0.01561382
405                Cadillac.Escalade.EXT     15.95   2667 7.888710 0.01859360
406             Chevrolet.Avalanche.1500     14.95   2575 7.853605 0.01648470
```

Cars2004 (subset, $n = 412$), consumption ~ log(weight)

Leverage points



11.4 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 ($t = 1, \dots, n$) 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.,

$$\mathsf{M}:\ \boldsymbol{Y}\,\big|\,\mathbb{X} \ \sim\ \big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big), \qquad \mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k,$$

and the quantity based on the leave-one-out model

$$\mathsf{M}_{(-t)}:\ \boldsymbol{Y}_{(-t)}\,\big|\,\mathbb{X}_{(-t)} \ \sim\ \big(\mathbb{X}_{(-t)}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_{n-1}\big).$$

It will overally be assumed, that $m_{t,t} > 0$ which implies (see Lemma 11.1) $\mathsf{rank}\big(\mathbb{X}_{(-t)}\big) = \mathsf{rank}\big(\mathbb{X}\big) = k$.

11.4.1 DFBETAS

The LSE's of the vector of regression coefficients based on the two models are

$$\begin{aligned}
\mathsf{M}: \qquad \widehat{\boldsymbol{\beta}} \ &=\ \big(\widehat{\beta}_0,\, \dots,\, \widehat{\beta}_{k-1}\big)^\top &&=\ \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y},\\
\mathsf{M}_{(-t)}: \qquad \widehat{\boldsymbol{\beta}}_{(-t)} \ &=\ \big(\widehat{\beta}_{(-t),0},\, \dots,\, \widehat{\beta}_{(-t),k-1}\big)^\top &&=\ \big(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\big)^{-1}\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}.
\end{aligned}$$

Using (11.1):

$$\widehat{\boldsymbol{\beta}} \ -\ \widehat{\boldsymbol{\beta}}_{(-t)} \ =\ \frac{U_t}{m_{t,t}}\,\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{x}_t, \tag{11.3}$$

which quantifies influence of the tth observation on the LSE of the regression coefficients. In the following, let $\boldsymbol{v}_0 = \big(v_{0,0},\, \dots,\, v_{0,k-1}\big)^\top,\ \dots,\ \boldsymbol{v}_{k-1} = \big(v_{k-1,0},\, \dots,\, v_{k-1,k-1}\big)^\top$ be the rows of the matrix $\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}$, i.e.,

$$\big(\mathbb{X}^\top\mathbb{X}\big)^{-1} = \begin{pmatrix} \boldsymbol{v}_0^\top \\ \vdots \\ \boldsymbol{v}_{k-1}^\top \end{pmatrix} = \begin{pmatrix} v_{0,0} & \dots & v_{0,k-1} \\ \vdots & \vdots & \vdots \\ v_{k-1,0} & \dots & v_{k-1,k-1} \end{pmatrix}.$$

Expression (11.3) written elementwise lead to a quantities called **DFBETA**:

$$\text{DFBETA}_{t,j} \ :=\ \widehat{\beta}_j \ -\ \widehat{\beta}_{(-t),j} \ =\ \frac{U_t}{m_{t,t}}\,\boldsymbol{v}_t^\top\boldsymbol{x}_t, \qquad t = 1, \dots, n,\ j = 0,\, \dots, k-1.$$

Note that $\text{DFBETA}_{t,j}$ 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"):

$$\text{DFBETAS}_{t,j} \ :=\ \frac{\widehat{\beta}_j \ -\ \widehat{\beta}_{(-t),j}}{\sqrt{\mathsf{MS}_{e,(-t)}\, v_{j,j}}} \ =\ \frac{U_t}{m_{t,t}\,\sqrt{\mathsf{MS}_{e,(-t)}\, v_{j,j}}}\,\boldsymbol{v}_t^\top\boldsymbol{x}_t,$$

$$t = 1, \dots, n,\ j = 0,\, \dots, k-1.$$

The reason for using $\sqrt{\mathsf{MS}_{e,(-t)}\, v_{j,j}}$ as a scale factor is that $\mathsf{MS}_{e,(-t)}$ 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. 11.1). On the other hand, a value of $v_{(-t),j,j}$ (that fits with the leave-one-out residual mean square $\mathsf{MS}_{e,(-t)}$ 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

$$\left|\text{DFBETAS}_{t,j}\right| > 1.$$

Illustrations

`Cars2004` **(subset,** $n = 412$**),** `consumption` $\sim$ `log(weight)`

DFBETAS

DFBETAS

```
dfbetas(m1)
```

```
    (Intercept)      lweight
1    0.058079251 -0.057288572
2    0.067760218 -0.066859700
3   -0.017131716  0.016817978
4   -0.029182966  0.028603518
5   -0.017131716  0.016817978
6   -0.037145548  0.036495821
7   -0.045873896  0.045023905
8    0.007702297 -0.007562061
9   -0.021494294  0.021106330
10   0.009424138 -0.009254036
...
```

Maximal absolute values of DFBETAS for each regressor

```
apply(abs(dfbetas(m1)), 2, max)
```

```
(Intercept)     lweight
  0.7344821   0.7415123
```

11.4.2 DFFITS

The LSE's of $\mu_t := \mathbb{E}\bigl(Y_t \bigm| \boldsymbol{X}_t = \boldsymbol{x}_t\bigr) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$ in the two models are

$$\begin{aligned} \mathsf{M}: &\quad \widehat{Y}_t = \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}, \\ \mathsf{M}_{(-t)}: &\quad \widehat{Y}_{[t]} = \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_{(-t)}. \end{aligned}$$

Using (11.1):

$$\widehat{Y}_{[t]} \;=\; \boldsymbol{x}_t^\top \Bigl\{\widehat{\boldsymbol{\beta}} - \frac{U_t}{m_{t,t}} \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1} \boldsymbol{x}_t\Bigr\} \;=\; \widehat{Y}_t - \frac{U_t}{m_{t,t}} \boldsymbol{x}_t^\top \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1} \boldsymbol{x}_t \;=\; \widehat{Y}_t - U_t\,\frac{h_{t,t}}{m_{t,t}}.$$

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:

$$\mathrm{DFFIT}_t \;:=\; \widehat{Y}_t - \widehat{Y}_{[t]} \;=\; U_t\,\frac{h_{t,t}}{m_{t,t}}, \qquad t = 1, \dots, n.$$

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, $\mathsf{S.E.}\bigl(\widehat{Y}_t\bigr) = \sqrt{\mathsf{MS}_e\, h_{t,t}}$) leading to a quantity called **DFFITS**:

$$\begin{aligned} \mathrm{DFFITS}_t \;&:=\; \frac{\widehat{Y}_t - \widehat{Y}_{[t]}}{\sqrt{\mathsf{MS}_{e,(-t)}\, h_{t,t}}} \\ &=\; \frac{h_{t,t}}{m_{t,t}}\,\frac{U_t}{\sqrt{\mathsf{MS}_{e,(-t)}\, h_{t,t}}} \;=\; \sqrt{\frac{h_{t,t}}{m_{t,t}}}\,\frac{U_t}{\sqrt{\mathsf{MS}_{e,(-t)}\, m_{t,t}}} \;=\; \sqrt{\frac{h_{t,t}}{m_{t,t}}}\;T_t, \qquad t = 1, \dots, n, \end{aligned}$$

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

$$\bigl|\mathrm{DFFITS}_t\bigr| > 3\sqrt{\frac{k}{n-k}}.$$

Illustrations

`Cars2004` **(subset, $n = 412$),** `consumption` $\sim$ `log(weight)`

DFFITS

DFFITS

```
dffits(m1)
```

```
           1            2             3             4             5
0.0645330957  0.0749431929 -0.0204914092 -0.0358359160 -0.0204914092  ...
```

$$3\sqrt{\frac{k}{n-k}} = 0.2095$$

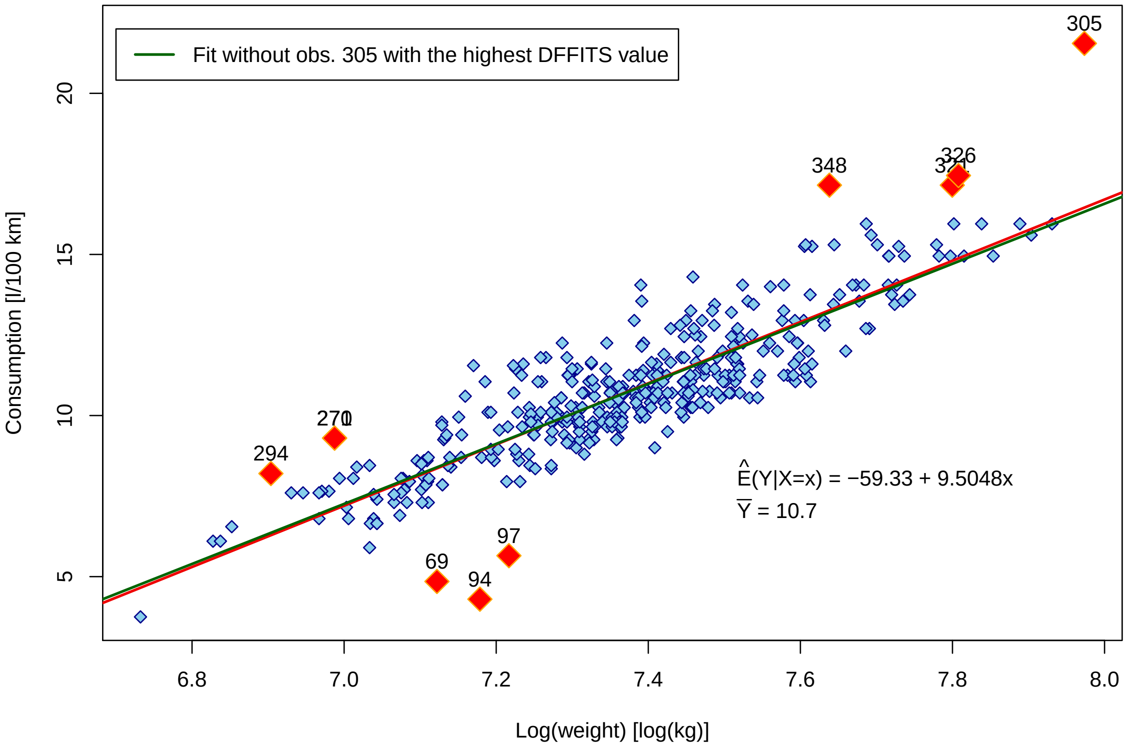
```
sum(abs(dffits(m1)) > 3 * sqrt(k / (n-k)))
```

```
[1] 10
```

```
                                   vname consumption weight  lweight     dffits
69                Honda.Civic.Hybrid.4dr        4.85   1239 7.122060 -0.2598440
                   manual.(gas/electric)
94       Toyota.Prius.4dr.(gas/electric)        4.30   1311 7.178545 -0.2984834
97          Volkswagen.Jetta.GLS.TDI.4dr        5.65   1362 7.216709 -0.2114462
270       Mazda.MX-5.Miata.convertible.2dr        9.30   1083 6.987490  0.2216790
271  Mazda.MX-5.Miata.LS.convertible.2dr        9.30   1083 6.987490  0.2216790
294     Toyota.MR2.Spyder.convertible.2dr        8.20    996 6.903747  0.2254823
305                             Hummer.H2       21.55   2903 7.973500  0.7815812
321           Land.Rover.Range.Rover.HSE       17.15   2440 7.799753  0.2597672
326                   Mercedes-Benz.G500       17.45   2460 7.807917  0.2892681
348              Land.Rover.Discovery.SE       17.15   2076 7.638198  0.3049335
```

`Cars2004` **(subset, $n = 412$),** `consumption` $\sim$ `log(weight)`

Large DFFITS values



11.4.3 Cook distance

In this Section, we concentrate on evaluation of the influence of the tth observation on the LSE of a vector parameter $\boldsymbol{\mu} := \mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr) = \mathbb{X}\boldsymbol{\beta}$. As in Section 11.4.2, let $\widehat{\boldsymbol{\beta}} = \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ be any solution to normal equations in model M and let $\widehat{\boldsymbol{\beta}}_{(-t)} = \bigl(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\bigr)^{-}\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}$ 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

$$\begin{aligned} \mathsf{M}: \qquad & \widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} \quad = \mathbb{H}\boldsymbol{Y}, \\ \mathsf{M}_{(-t)}: \quad & \widehat{\boldsymbol{Y}}_{(-t\bullet)} := \mathbb{X}\widehat{\boldsymbol{\beta}}_{(-t)}. \end{aligned}$$

Note. Remind that $\widehat{\boldsymbol{Y}}_{(-t\bullet)}$, $\widehat{\boldsymbol{Y}}_{[\bullet]}$ and $\widehat{\boldsymbol{Y}}_{(-t)}$ are three different quantities. Namely,

$$\widehat{\boldsymbol{Y}}_{(-t\bullet)} = \mathbb{X}\widehat{\boldsymbol{\beta}}_{(-t)} = \begin{pmatrix} \boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}_{(-t)} \\ \vdots \\ \boldsymbol{x}_n^\top\widehat{\boldsymbol{\beta}}_{(-t)} \end{pmatrix}, \qquad \widehat{\boldsymbol{Y}}_{[\bullet]} = \begin{pmatrix} \widehat{Y}_{[1]} \\ \vdots \\ \widehat{Y}_{[n]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}_{(-1)} \\ \vdots \\ \boldsymbol{x}_n^\top\widehat{\boldsymbol{\beta}}_{(-n)} \end{pmatrix}.$$

Finally, $\widehat{\boldsymbol{Y}}_{(-t)} = \mathbb{X}_{(-t)}\widehat{\boldsymbol{\beta}}_{(-t)}$ 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

$$\bigl\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)}\bigr\|^2.$$

Let us remind from Lemma 11.3:

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-t)} = \frac{U_t}{m_{t,t}}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_t.$$

Hence,

$$\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)} = \mathbb{X}\bigl(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-t)}\bigr) = \frac{U_t}{m_{t,t}}\mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_t.$$

Then

$$\begin{aligned} \bigl\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)}\bigr\|^2 &= \Bigl\|\frac{U_t}{m_{t,t}}\mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_t\Bigr\|^2 \\ &= \frac{U_t^2}{m_{t,t}^2}\,\boldsymbol{x}_t^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_t \\ &= \frac{U_t^2}{m_{t,t}^2}\,h_{t,t}. \end{aligned} \tag{11.4}$$

The equality (11.4) follows from noting that $\boldsymbol{x}_t^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_t$ is the tth diagonal element of matrix $\mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top = \mathbb{X}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top = \mathbb{H}$.

The so called **Cook distance** of the tth observation is (11.4) modified to get a unit-free quantity. Namely, the Cook distance is defined as

$$D_t := \frac{1}{k\,\mathsf{MS}_e}\bigl\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-t\bullet)}\bigr\|^2.$$

Expression (11.4) 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

$$D_t = \frac{1}{k}\,\frac{h_{t,t}}{m_{t,t}}\,\frac{U_t^2}{\mathsf{MS}_e\,m_{t,t}} = \frac{1}{k}\,\frac{h_{t,t}}{m_{t,t}}\,\bigl(U_t^{std}\bigr)^2.$$

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.

Further, 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

$$\widehat{\boldsymbol{\beta}}_{(-t)} \in \mathcal{C}(\alpha) \qquad \text{if and only if} \qquad D_t < \mathcal{F}_{k,n-k}(1-\alpha). \tag{11.5}$$

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

$$D_t > \mathcal{F}_{k,n-k}(0.50).$$

Illustrations

Cars2004 **(subset,** $n = 412$**),** consumption $\sim$ log(weight)

Cook distance

Cook distance

```
cooks.distance(m1)
```

```
           1            2            3            4            5
0.0020855185 0.0028118990 0.0002104334 0.0006433764 0.0002104334   ...
```

$$\mathcal{F}_{k,n-k}(0.50) = 0.6943$$

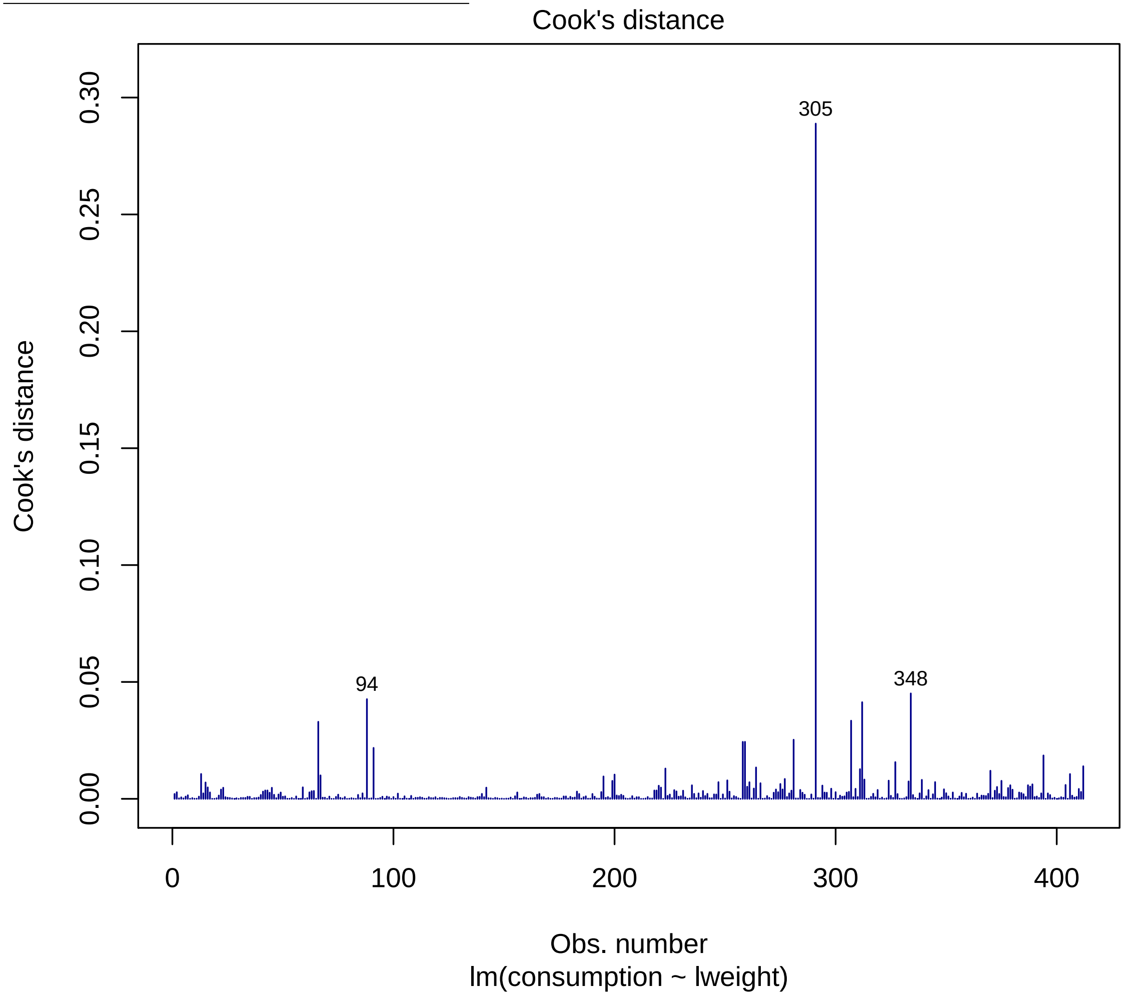
Maximal Cook distance

```
max(cooks.distance(m1))
```

```
[1] 0.288855
```

Cars2004 **(subset,** $n = 412$**),** consumption $\sim$ log(weight)

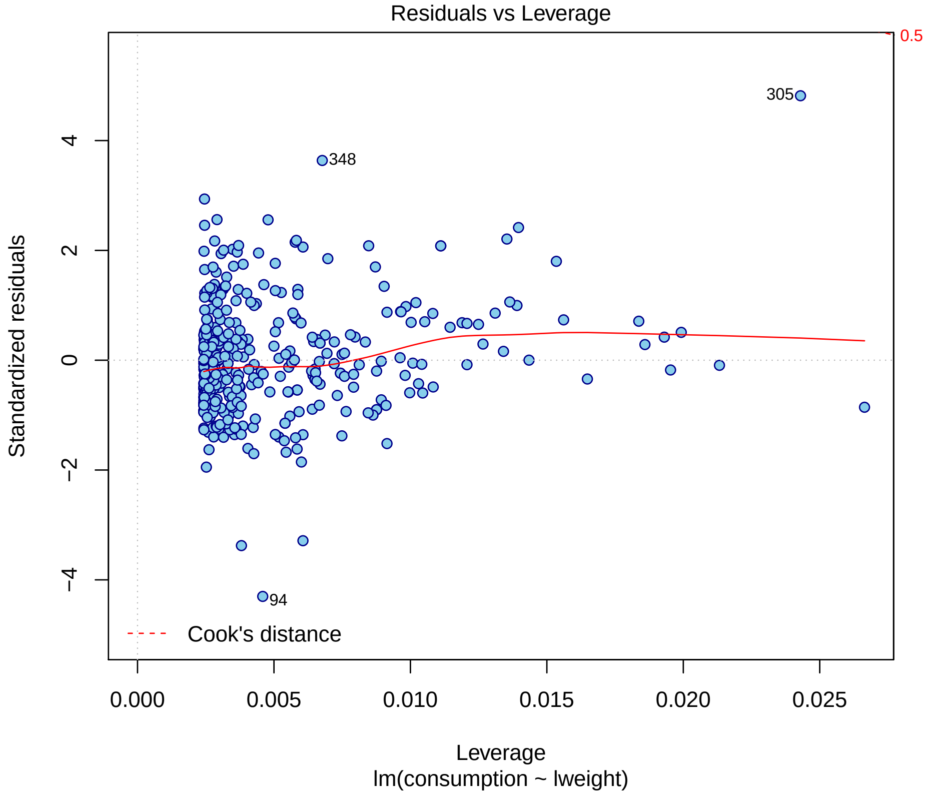
R diagnostic plot (plot(m1, which = 4))



Illustrations

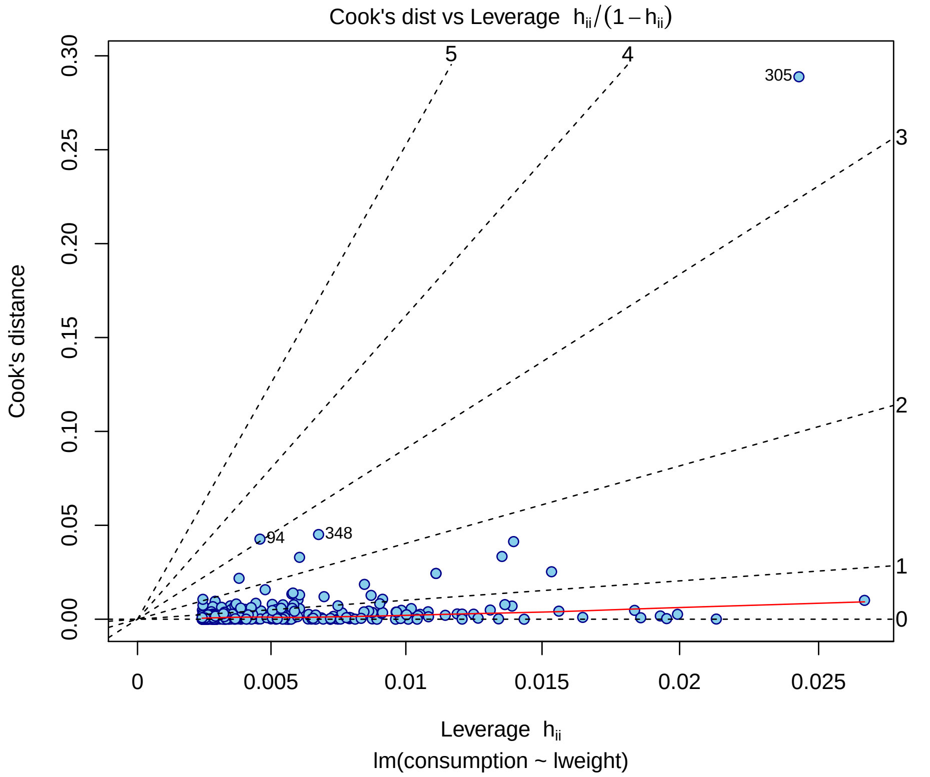
Cars2004 (subset, $n = 412$), `consumption` $\sim$ `log(weight)`

R diagnostic plot (`plot(m1, which = 5)`)



Cars2004 (subset, $n = 412$), `consumption` $\sim$ `log(weight)`

R diagnostic plot (`plot(m1, which = 6)`)



The x-axis shows values of $h_{i,i}/(1 - h_{i,i})$ and not $h_{i,i}$. Contours are related to the values of $U_t^{std}/\sqrt{k}$.

11.4.4 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

$$\begin{aligned} \mathsf{M}: \qquad \widehat{\boldsymbol{\beta}} &= \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\boldsymbol{Y}, \\ \mathsf{M}_{(-t)}: \qquad \widehat{\boldsymbol{\beta}}_{(-t)} &= \bigl(\mathbb{X}_{(-t)}{}^\top\mathbb{X}_{(-t)}\bigr)^{-1}\mathbb{X}_{(-t)}{}^\top\boldsymbol{Y}_{(-t)}. \end{aligned}$$

The estimated covariance matrices of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(-t)}$, respectively, are

$$\begin{aligned} \widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{X}\bigr) &= \mathsf{MS}_e\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}, \\ \widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}}_{(-t)}\,\big|\,\mathbb{X}\bigr) &= \mathsf{MS}_{e,(-t)}\,\bigl(\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\bigr)^{-1}. \end{aligned}$$

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

$$\mathrm{COVRATIO}_t = \frac{\det\Bigl\{\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}}_{(-t)}\,\big|\,\mathbb{X}\bigr)\Bigr\}}{\det\Bigl\{\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol{\beta}}\,\big|\,\mathbb{X}\bigr)\Bigr\}}, \qquad t = 1,\dots,n.$$

After some calculation (see below), it can be shown that

$$\mathrm{COVRATIO}_t = \frac{1}{m_{t,t}}\left\{\frac{n-k-\bigl(U_t^{std}\bigr)^2}{n-k-1}\right\}^k, \qquad t = 1,\dots,n.$$

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

$$\bigl|1 - \mathrm{COVRATIO}_t\bigr| > 3\,\frac{k}{n-k}.$$

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

$$\begin{vmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{C} & \mathbb{D} \end{vmatrix} = \bigl|\mathbb{A}\bigr|\cdot\bigl|\mathbb{D} - \mathbb{C}\mathbb{A}^{-1}\mathbb{B}\bigr| = \bigl|\mathbb{D}\bigr|\cdot\bigl|\mathbb{A} - \mathbb{B}\mathbb{D}^{-1}\mathbb{C}\bigr|.$$

Use twice the above identity:

$$\begin{aligned} \begin{vmatrix} \mathbb{X}^\top\mathbb{X} & \boldsymbol{x}_t \\ \boldsymbol{x}_t^\top & 1 \end{vmatrix} &= \bigl|\mathbb{X}^\top\mathbb{X}\bigr|\cdot\bigl|\underbrace{1 - \boldsymbol{x}_t^\top\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{x}_t}_{1-h_{t,t}=m_{t,t}}\bigr| = \bigl|\mathbb{X}^\top\mathbb{X}\bigr|\,m_{t,t}, \\ &= |1|\cdot\bigl|\mathbb{X}^\top\mathbb{X} - \boldsymbol{x}_t\boldsymbol{x}_t^\top\bigr| = \bigl|\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\bigr|. \end{aligned}$$

So that, $\qquad m_{t,t}\,\bigl|\mathbb{X}^\top\mathbb{X}\bigr| = \bigl|\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\bigr|.$

Then,

$$\frac{\det\Bigl\{\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol\beta}_{(-t)}\,\big|\,\mathbb{X}\bigr)\Bigr\}}{\det\Bigl\{\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol\beta}\,\big|\,\mathbb{X}\bigr)\Bigr\}} = \frac{\Bigl|\mathsf{MS}_{e,(-t)}\,\bigl(\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\bigr)^{-1}\Bigr|}{\Bigl|\mathsf{MS}_{e}\,\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\Bigr|}$$

$$= \left(\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e}\right)^k \cdot \frac{\bigl|\mathbb{X}_{(-t)}^\top\mathbb{X}_{(-t)}\bigr|^{-1}}{\bigl|\mathbb{X}^\top\mathbb{X}\bigr|^{-1}} = \left(\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e}\right)^k \cdot \frac{1}{m_{t,t}}.$$

Expression (11.1):

$$\frac{\mathsf{MS}_{e,(-t)}}{\mathsf{MS}_e} = \frac{n-k-\bigl(U_t^{std}\bigr)^2}{n-k-1}.$$

Hence, $$\frac{\det\Bigl\{\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol\beta}_{(-t)}\,\big|\,\mathbb{X}\bigr)\Bigr\}}{\det\Bigl\{\widehat{\mathsf{var}}\bigl(\widehat{\boldsymbol\beta}\,\big|\,\mathbb{X}\bigr)\Bigr\}} = \frac{1}{m_{t,t}}\left(\frac{n-k-\bigl(U_t^{std}\bigr)^2}{n-k-1}\right)^k.$$

Illustrations

Cars2004 (subset, $n = 412$), consumption ~ log(weight)

COVRATIO

COVRATIO

```
covratio(m1)
```

```
       1        2        3        4        5        6 
1.014754 1.014674 1.012147 1.010719 1.012147 1.011724   ...
```

$3\,\frac{k}{n-k} = 0.0146$

```
sum(abs(1 - covratio(m1)) > 3 * (k / (n-k)))
```

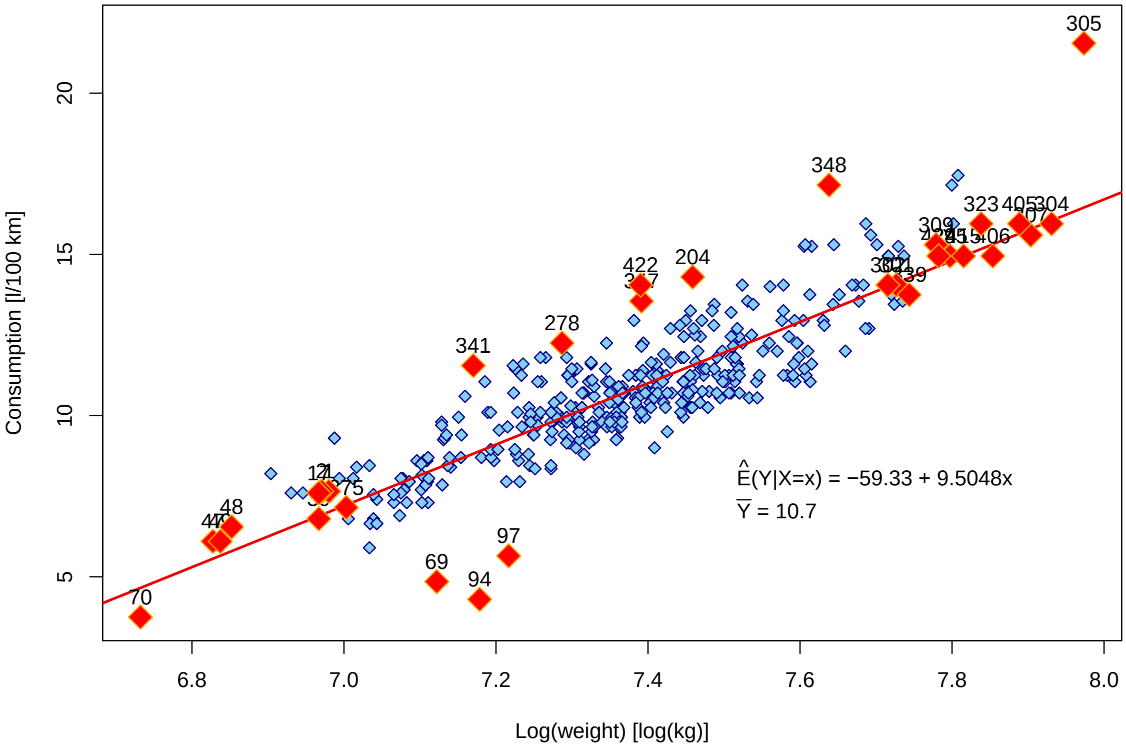
```
[1] 31
```

```
                              vname consumption weight  lweight  covratio
1                Chevrolet.Aveo.4dr        7.65   1075 6.980076 1.0147544
2       Chevrolet.Aveo.LS.4dr.hatch        7.65   1065 6.970730 1.0146741
17      Hyundai.Accent.GT.2dr.hatch        7.60   1061 6.966967 1.0149481
39               Scion.xA.4dr.hatch        6.80   1061 6.966967 1.0171433
47           Toyota.Echo.2dr.manual        6.10    923 6.827629 1.0240384
48             Toyota.Echo.2dr.auto        6.55    946 6.852243 1.0211810
49                  Toyota.Echo.4dr        6.10    932 6.837333 1.0237925
69               Honda.Civic.Hybrid        4.85   1239 7.122060 0.9584411
         .4dr.manual.(gas/electric)
70 Honda.Insight.2dr.(gas/electric)        3.75    839 6.732211 1.0287100
...
305                       Hummer.H2       21.55   2903 7.973500 0.9166531
...
```

Illustrations

`Cars2004` **(subset,** $n = 412$**),** `consumption` $\sim$ `log(weight)`

COVRATIO value far from 1



11.4.5 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”!

Chapter 12

Model Building

Plagiarism notice: Some parts of the text of this chapter are only a mild modification of Section 2.9 from *Course notes to NMST432 Advanced Regression Models* (version dated May 22, 2020) by Michal Kulich.

In this chapter we concentrate on a situation when a (larger) set of covariates is available, i.e., the generic covariate vector $\boldsymbol{Z}$ equals $\bigl(Z_1, \dots, Z_p\bigr)^\top$ with $p > 1$ (often $p \gg 1$) and the task is to propose a regression function to express $\mathbb{E}\bigl(Y \bigm| \boldsymbol{Z} = \boldsymbol{z}\bigr)$ for a given outcome variable Y. In particular, we will now not consider a situation where $p = 1$, $Z \equiv Z_1$ is numeric and $\mathbb{E}\bigl(Y \bigm| Z = z\bigr)$ is a real function of one variable. Developing a reasonable model in this case corresponds to finding a suitable parameterization for the numeric covariate Z and this problem was quite extensively covered already by Section 4.3.

12.1 General principles

When a larger set of covariates is available ($p \gg 1$), most numeric covariates are usually parameterized by identity (included in the model "as they are"), categorical covariates are still parameterized by a suitable choice of (pseudo)contrast (the reference group pseudocontrasts, i.e., the dummy variables dominate in practical applications for easy interpretation of related regression coefficients). Hierarchically well formulated models are then considered while including at most two-way interactions in the regression function. The principal and quite challenging task is to decide which model terms (which interactions, which main effects) should be included in the final model which is then used to solve the research problem that triggers the regression analyzis.

The primary tool for model building in context of a linear model are submodel tests (see Chapter 8) comparing a larger model with a submodel. If the submodel test is significant it means that the terms in the larger model cannot be removed without a significant decrease in the quality of model fit.

Since the development of the final model usually involves repeated applications of submodel tests, each performed on a selected level α (usually $\alpha = 0.05$), it is clear that the overall procedure does not preserve the desired level. If many tests are done then the final model is likely to include terms that in fact do not affect the response at all (overfitting). There is no universal and reliable method for adjusting the levels of the individual tests so that the overall probability of including irrelevant terms is under control. Nevertheless the analyst should be aware of this problem and should not interpret the p-values of submodel tests too dogmatically.

Approaches for developing reasonable models vary with the nature of the problem, structure of the data and questions to be addressed by the analysis. There is **no universal solution** to be recommended. Each problem requires careful consideration by the analyst taking into account the nature of the problem, the data-collection methods and tools, the meaning of the variables included in the dataset, their mutual relationships, and the goals of the analysis. Experience of the analyst is one of the most important determinants of success in model building. That is, it is only possible to learn on how to build the model by doing it (repeatedly). Passive reading of textbooks or course notes will not develop necessary expertise.

Even though no universal solution exists, there are two broad classes of problems requiring a bit different strategies which will now be briefly discussed.

12.2 Prediction

If ***prediction*** is the primary goal, it is useful to consider rich and flexible models. Omission of an important term from the model or its inclusion with an inappropriate transformation may have detrimental biasing effects on the predictions (Section 10.2.3). If unnecessary covariates are left in, the variability in the predicted response is increased but the predictions are not biased (Section 10.2.4). Interpretation of regression parameters is usually not that important. That is, parameterization of some of numeric covariates by splines or polynomials is not a problem and is recommended if it appears that the effect of a particular covariate is not linear.

In prediction analyses, validation of the prediction model should be performed either by dividing the data set into disjoint training (used for model building) and validation (used for evaluation of the predictions) subsets or at least by cross-validation (predictions of each observation by a model fitted on data excluding that observation compared to observed values). Validation is also a very useful tool for selection of the best prediction model out of several candidates.

Even though multicollinearity is not a problem for fitted values (Section 10.1) which we considered as predictions that were used to evaluate a prediction quality of the model, it might be a problem for external validity of the developed prediction model. It means that multicollinearity may invalidate predictions for new observations with a combination of the covariate values not being exactly present in a dataset used to build the prediction model. Such a problem would also be revealed by (cross-)validation. Finally, even if prediction is the primary goal, it is usually of interest to also know which covariates out of a set of available p covariates $Z_1, \ldots, Z_p$ are "significantly" associated with the outcome (= those included in the final model). For this reason, it is highly recommended to treat multicollinearity even when the model is primarily built for prediction purposes.

12.3 Evaluation of a covariate effect

The second typical problem, often encountered with data coming from *observational studies*, is to ***evaluate covariate effects***. That is, there is one covariate (of few covariates) of primary interest (let us denote it as X) and the task is to evaluate how this covariate affects the mean of the response. Quite often, the analyst also hypothesize some causal pathways and tries to show that changes in the X variables *cause* changes in the expected outcome.

When evaluating the covariate effect, one should be aware of results derived in Chapter 10. Multicollinearity is a huge problem, especially if we include covariates being collinear with the primary covariate X. The collinear covariates may completely diminish the effect of the primary covariate in a final model. On the other hand, omission of important explanatory variables will bias estimation of the effects of the primary covariate X (10.2.3). In this respect, so called *confounding* is the most frequent problem which requires careful pre-analyzis of not only available data but also deep understanding of the problem at hand. It may easily happen that important explanatory variables (e.g., *confounders*) are not directly available in a dataset to be analyzed. The fact that they are not available to the analyst does not remove possible bias from estimation of effects of the primary covariate.

When building a model for purpose of evaluation of a covariate effecf, one must be really careful about several things.

- First, the covariate of primary interest must be kept in the model even if it is not significant – otherwise its effect cannot be evaluated.
- Second, the regression parameters expressing the influence of the covariate of interest should have a straightforward interpretation. Thus, we cannot afford to model the effect of X by a complicated

function that cannot be easily summarized (splines of degree > 1, polynomials, . . .), or to use complex transformations of the response that are difficult to interpret.

- Third, there might be covariates that should be kept in the model regardless of their significance (suspected *confounders*) and/or covariates that should not be included in the model no matter how significant they are (variables on the causal pathway between X and Y, variables that are influenced by the value of Y).

Thus, making reasonable decisions about which covariates should be included in the model and which should be dropped is not based solely on significance tests but also on external expert knowledge of the problem to be analyzed. ***It is precisely this issue that makes automated computer-based algorithms (unsupervised stepwise regression, regression trees, neural networks, deep learning, etc.) unable to solve certain problems acceptably.***

12.4 Model building strategy

The common problem in model-building strategies is the inclusion of interactions, especially when the number of covariates that can be considered for interactions is quite large. The strategy that starts with a model that includes a lot of main effects as well as all possible two-way interactions between them, and tries to gradually eliminate the superfluous terms usually does not lead to a good model. With this approach, we are likely to end up with a model that suffers from overfitting, keeps a lot of unnecessary interactions and is hard to interpret. It is better to fit only the main effects first and eliminate those that are not contributing to the model. As soon as the total amount of available covariates has been eliminated in this way, either all two-way interactions based only on the included effects may be added to the model and then eliminated by another sequence of submodel F-tests. Or one may try to add two-way interactions of the remaining terms one by one. This strategy is much more likely to end up only with interactions that really matter. Considering higher order interactions (three-way, four-way, . . .) is usually a hopeless task. It is better not to consider them at all, except in analyses where, for some reason, such interactions are among the terms of interest.

There is one principle about building models with interactions, which is almost universally valid and the analyst should take care not to violate it. The models should be built hierarchically, meaning that if a covariate is present in a higher-order interaction, then all its corresponding lower-order interactions as well as the main effects should be included in the model as well, no matter if they are significant or not. This principle should be ignored only in analyses where there is a sound justification for its violation.

12.5 Conclusion

This brief exposition of model-building strategies cannot be complete and should be understood in the whole context of the particular task to be done. As noted earlier, each problem should be carefully considered in order to choose a tailor-made strategy that works well for it. This requires practical experience. The analyst should be aware that there is no such thing as the true model and that his task is not to discover it. All models are wrong – we are only looking for an acceptable model that provides satisfactory answers to the questions of interest.

Chapter 13

Analysis of Variance

In this chapter, we examine few specific issues of linear models where all covariates are *categorical*. That is, the covariate vector $\boldsymbol{Z}$ is $\boldsymbol{Z} = \bigl(Z_1, \dots, Z_p\bigr)^\top$, $Z_j \in \mathcal{Z}_j$, $j = 1, \dots, p$, 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.

13.1 One-way classification

One-way classification corresponds to situation of one categorical covariate $Z \in \mathcal{Z} = \{1, \ldots, G\}$, see also Section 4.4. A linear model can then used to parameterize a set of G (conditional) response expectations $\mathbb{E}\big(Y \,\big|\, Z = 1\big)$, ..., $\mathbb{E}\big(Y \,\big|\, Z = G\big)$ that we call as *one-way classified group means*:

$$m(g) = \mathbb{E}\big(Y \,\big|\, Z = g\big) =: m_g, \qquad g = 1, \ldots, G.$$

Without loss of generality, we can assume that the response random variables $Y_1, \ldots, Y_n$ are sorted such that

$$\begin{array}{lcccccl} Z_1 & = & \cdots & = & Z_{n_1} & = & 1, \\ Z_{n_1+1} & = & \cdots & = & Z_{n_1+n_2} & = & 2, \\ & & & & & \vdots & \\ Z_{n_1+\cdots+n_{G-1}+1} & = & \cdots & = & Z_n & = & G. \end{array}$$

As in Section 4.4, it is useful (for notational clarity in theoretical derivations) 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 $\boldsymbol{Y}_g$:

$$\begin{array}{lll} Z = 1: & \boldsymbol{Y}_1 = \big(Y_{1,1}, \ldots, Y_{1,n_1}\big)^\top & = \big(Y_1, \ldots, Y_{n_1}\big)^\top, \\ \vdots & \qquad\vdots & \quad\vdots \\ Z = G: & \boldsymbol{Y}_G = \big(Y_{G,1}, \ldots, Y_{G,n_G}\big)^\top & = \big(Y_{n_1+\cdots+n_{G-1}+1}, \ldots, Y_n\big)^\top. \end{array}$$

The full response vector is $\boldsymbol{Y}$ and its (conditional, given $\mathbb{Z} = \big(Z_1, \ldots, Z_n\big)^\top$) mean are

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_G \end{pmatrix}, \qquad \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \begin{pmatrix} m_1 \, \mathbf{1}_{n_1} \\ \vdots \\ m_G \, \mathbf{1}_{n_G} \end{pmatrix} =: \boldsymbol{\mu}. \tag{13.1}$$

By using a linear model, we just use a suitable parameterization of the mean vector $\boldsymbol{\mu}$ given by (13.1). By using a standard linear model, we additionally assume that

$$\mathsf{var}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big) = \sigma^2 \, \mathbf{I}_n. \tag{13.2}$$

As in Section 4.4, 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.

13.1.1 Parameters of interest

Differences between the group means

The principal inferential interest with one-way classification lies in estimation of and tests on parameters

$$\theta_{g,h} := m_g - m_h, \qquad g,\, h = 1, \dots, G,\ g \neq h,$$

which are the differences between the group means. Since each $\theta_{g,h}$ is a linear combination of the elements of the mean vector $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$. The LSE of each $\theta_{g,h}$ is trivially 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

$$\mathrm{H}_0:\ m_1 = \cdots = m_G,$$

which written in terms of the differences between the group means is

$$\mathrm{H}_0:\ \theta_{g,h} = 0, \quad g,\, h = 1, \dots, G,\ g \neq h.$$

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 13.1 Factor effects in a one-way classification.
By factor effects *in case of a one-way classification we understand the quantities $\eta_1, \dots, \eta_G$ defined as*

$$\eta_g = m_g - \overline{m}, \qquad g = 1, \dots, G,$$

where $\overline{m} = \dfrac{1}{G}\sum_{h=1}^{G} m_h$ is the mean of the group means.

Notes.

- The factor effects are again linear combinations of the elements of the mean vector $\boldsymbol{\mu} = \mathbb{E}\big(\boldsymbol{Y} \,\big|\, \mathbb{Z}\big)$ and hence for all of them the LSE is 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
$$\mathrm{H}_0: \eta_g = 0, \quad g = 1, \dots, G,$$
is equivalent to the null hypothesis $\mathrm{H}_0: m_1 = \cdots = m_G$ on the equality of the group means.

13.1.2 One-way ANOVA model

As a reminder from Section 4.4.2, 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 Section 4.4.3, we introduced class of full-rank parameterizations of this regression space and of the response mean vector $\boldsymbol{\mu}$ as $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$, $\boldsymbol{\beta} \in \mathbb{R}^k$:

$$m_g \;=\; \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, \qquad g = 1, \ldots, G$$

with $k = G$, $\boldsymbol{\beta} = \bigl(\beta_0,\; \underbrace{\boldsymbol{\beta}^Z}_{(\beta_1,\,\ldots,\,\beta_{G-1})^\top}\;\bigr)^\top$,

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 sum contrasts (see expression 4.23) are used then

$$\begin{aligned}
\alpha_0 &:= \beta_0 &&= \frac{1}{G}\sum_{g=1}^{G} m_g &&= \overline{m},\\
\alpha_g &:= \beta_g &&= m_g - \frac{1}{H}\sum_{h=1}^{H} m_h &&= m_g - \overline{m} = \eta_g, \qquad g = 1, \ldots, G-1,\\
\alpha_G &:= -\sum_{h=1}^{G-1}\beta_h &&= m_G - \frac{1}{H}\sum_{h=1}^{H} m_h &&= m_G - \overline{m} = \eta_G.
\end{aligned}$$

That is, parameters $\alpha_1,\,\ldots,\,\alpha_G$ are equal to the factor effects.

Terminology. Related linear model is referred to as *one-way ANOVA model*[1]

Notes.

- Depending on chosen parameterization the differences between the group means, parameters $\theta_{g,h}$, are expressed as
$$\theta_{g,h} = \alpha_g - \alpha_h = \bigl(\boldsymbol{c}_g - \boldsymbol{c}_h\bigr)^\top \boldsymbol{\beta}^Z, \qquad g \neq h.$$
The null hypothesis $\mathrm{H}_0 : m_1 = \cdots m_G$ on equality of the group means is the expressed as $\mathrm{H}_0 : \beta_1 = 0 \;\&\; \ldots \;\&\; \beta_{G-1} = 0$, i.e., $\mathrm{H}_0 : \boldsymbol{\beta}^Z = \mathbf{0}_{G-1}$.
- If a normal linear model is assumed, test on a value of subvector of regression parameters or a submodel test which compares the one-way ANOVA model with the intercept-only model can be used to test the above null hypothesis. The corresponding F-test is indeed a well known one-way ANOVA F-test.

[1] *model analýzy rozptylu jednoduchého třídění*

13.1.3 Least squares estimation

In case of a one-way ANOVA linear model, explicit formulas for the LSE related quantities can easily be derived.

Lemma 13.1 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:

$$\widehat{m}_g = \widehat{Y}_{g,j} = \frac{1}{n_g}\sum_{l=1}^{n_g} Y_{g,l} =: \overline{Y}_{g\bullet}, \qquad g = 1,\dots,G,\; j = 1,\dots,n_g.$$

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* $\boldsymbol{\mu} = \big(m_1\,\mathbf{1}_{n_1}^\top,\,\dots,\, m_G\,\mathbf{1}_{n_G}^\top\big)^\top$, *then* $\widehat{\boldsymbol{m}}\,|\,\mathbb{Z} \sim \mathcal{N}_G\big(\boldsymbol{m},\, \sigma^2\,\mathbb{V}\big)$, *where*

$$\mathbb{V} = \begin{pmatrix}\frac{1}{n_1} & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \frac{1}{n_G}\end{pmatrix}.$$

Proof. Use a full-rank parameterization $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}$ with

$$\mathbb{X} = \begin{pmatrix}\mathbf{1}_{n_1} & \dots & \mathbf{0}_{n_1}\\ \vdots & \vdots & \vdots\\ \mathbf{0}_{n_G} & \vdots & \mathbf{1}_{n_G}\end{pmatrix}, \qquad \boldsymbol{\beta} = \big(m_1,\,\dots,m_G\big)^\top.$$

We have

$$\mathbb{X}^\top\mathbb{X} = \begin{pmatrix}n_1 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & n_G\end{pmatrix}, \quad \mathbb{X}^\top\boldsymbol{Y} = \begin{pmatrix}\sum_{j=1}^{n_1} Y_{1,j}\\ \vdots\\ \sum_{j=1}^{n_G} Y_{G,j}\end{pmatrix}, \quad \big(\mathbb{X}^\top\mathbb{X}\big)^{-1} = \begin{pmatrix}\frac{1}{n_1} & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \frac{1}{n_G}\end{pmatrix},$$

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{m}} = \big(\widehat{m}_1,\,\dots,\,\widehat{m}_G\big)^\top = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y} = \big(\overline{Y}_{1\bullet},\,\dots,\,\overline{Y}_{G\bullet}\big)^\top.$$

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 their linear combinations

With a full-rank parameterization, a vector $\boldsymbol{m}$ is linked to the regression coefficients $\boldsymbol{\beta} = \bigl(\beta_0,\, \boldsymbol{\beta}^Z\bigr)^\top$, $\boldsymbol{\beta}^Z = \bigl(\beta_1,\, \dots,\, \beta_{G-1}\bigr)^\top$, 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 $\widehat{\boldsymbol{\beta}} = \bigl(\widehat{\beta}_0,\, \widehat{\boldsymbol{\beta}}^Z\bigr)^\top$ 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

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}^Z \end{pmatrix} = \Bigl(\mathbf{1}_G,\, \mathbb{C}\Bigr)^{-1} \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{G\bullet} \end{pmatrix}.$$

That is, the LSE of the regression coefficients is always a linear combination of the group sample means. The same then holds, of course, also for any linear combination of regression coefficients. 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

$$\widehat{\eta}_g = \overline{Y}_{g\bullet} - \frac{1}{G}\sum_{h=1}^{G} \overline{Y}_{h\bullet}, \qquad g = 1, \dots, G.$$

13.1.4 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

$$\begin{aligned}\mathsf{SS}_e &= \big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\big(Y_{g,j} - \widehat{Y}_{g,j}\big)^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\big(Y_{g,j} - \overline{Y}_{g\bullet}\big)^2,\\ \nu_e &= n - G,\end{aligned}$$

$$\begin{aligned}\mathsf{SS}_R &= \big\|\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}_n\big\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\big(\widehat{Y}_{g,j} - \overline{Y}\big)^2 = \sum_{g=1}^{G} n_g\,\big(\overline{Y}_{g\bullet} - \overline{Y}\big)^2,\\ \nu_R &= G - 1.\end{aligned}$$

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 $\boldsymbol{Y}\,|\,\mathbb{Z} \sim \mathcal{N}_n\big(\mathbf{1}_n\beta_0,\, \sigma^2\mathbf{I}_n\big)$ of the one-way ANOVA model. A residual sum of squares of the submodel is

$$\mathsf{SS}_e^0 = \mathsf{SS}_T = \big\|\boldsymbol{Y} - \overline{Y}\mathbf{1}_n\big\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g}\big(Y_{g,j} - \overline{Y}\big)^2.$$

Breakdown of the total sum of squares (Lemma 7.3) 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

$$F = \frac{\frac{\mathsf{SS}_R}{G-1}}{\frac{\mathsf{SS}_e}{n-G}} = \frac{\mathsf{MS}_R}{\mathsf{MS}_e}, \tag{13.3}$$

where

$$\mathsf{MS}_R = \frac{\mathsf{SS}_R}{G-1}, \qquad \mathsf{MS}_e = \frac{\mathsf{SS}_e}{n-G}.$$

The F-statistic (13.3) 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:

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		

[2] *vnitroskupinový součet čtverců* [3] *meziskupinový součet čtverců*

Consider a terminology introduced in Section 5.5, and denote as Z main effect terms that correspond to the covariate Z. We have $\mathsf{SS}_R = \mathsf{SS}\bigl(\mathsf{Z}\bigm|\mathsf{1}\bigr)$ 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.

13.2 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 13.1, 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 13.1.

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 13.1, 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

$$n_{\bullet h} = \sum_{g=1}^{G} n_{g,h}, \qquad h = 1, \ldots, H$$

denote the number of datapoints with $W = h$. Finally, we will denote various means of the group means as follows.

$$\overline{m} \;:=\; \frac{1}{G \cdot H} \sum_{g=1}^{G} \sum_{h=1}^{H} m_{g,h},$$

[4] *dvojné třídění*

Table 13.1: Two-way classification: Covariate matrix and overall response vector.

$$\begin{pmatrix} Z_1 & W_1\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ Z_n & W_n \end{pmatrix} = \left(\begin{array}{cc} Z_{1,1,1} & W_{1,1,1}\\ \vdots & \vdots\\ Z_{1,1,n_{1,1}} & W_{1,1,n_{1,1}}\\ \hdashline \vdots & \vdots\\ \hdashline Z_{G,1,1} & W_{G,1,1}\\ \vdots & \vdots\\ Z_{G,1,n_{G,1}} & W_{G,1,n_{G,1}}\\ \hdashline \vdots & \vdots\\ \vdots & \vdots\\ \hdashline Z_{1,H,1} & W_{1,H,1}\\ \vdots & \vdots\\ Z_{1,H,n_{1,H}} & W_{1,H,n_{1,H}}\\ \hdashline \vdots & \vdots\\ \hdashline Z_{G,H,1} & W_{G,H,1}\\ \vdots & \vdots\\ Z_{G,H,n_{G,H}} & W_{G,H,n_{G,H}} \end{array}\right) = \left(\begin{array}{cc} 1 & 1\\ \vdots & \vdots\\ 1 & 1\\ \hdashline \vdots & \vdots\\ \hdashline G & 1\\ \vdots & \vdots\\ G & 1\\ \hdashline \vdots & \vdots\\ \vdots & \vdots\\ \hdashline 1 & H\\ \vdots & \vdots\\ 1 & H\\ \hdashline \vdots & \vdots\\ \hdashline G & H\\ \vdots & \vdots\\ G & H \end{array}\right), \quad \boldsymbol{Y} = \begin{pmatrix} Y_1\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ Y_n \end{pmatrix} = \left(\begin{array}{c} Y_{1,1,1}\\ \vdots\\ Y_{1,1,n_{1,1}}\\ \hdashline \vdots\\ \hdashline Y_{G,1,1}\\ \vdots\\ Y_{G,1,n_{G,1}}\\ \hdashline \vdots\\ \vdots\\ \hdashline Y_{1,H,1}\\ \vdots\\ Y_{1,H,n_{1,H}}\\ \hdashline \vdots\\ \hdashline Y_{G,H,1}\\ \vdots\\ Y_{G,H,n_{G,H}} \end{array}\right).$$

$$\begin{aligned} \overline{m}_{g\bullet} &:= \frac{1}{H}\sum_{h=1}^{H} m_{g,h}, \qquad && g = 1, \ldots, G,\\ \overline{m}_{\bullet h} &:= \frac{1}{G}\sum_{g=1}^{G} m_{g,h}, \qquad && h = 1, \ldots, H. \end{aligned}$$

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 13.2.

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).
- When interpreting the means of the group means, it must be taken into account that in general, it is not necessarily true that $\overline{m}_{g\bullet} = \mathbb{E}\big(Y \bigm| Z = g\big)$ $(g = 1, \ldots, G)$, $\overline{m}_{\bullet h} = \mathbb{E}\big(Y \bigm| W = h\big)$ $(h = 1, \ldots, H)$, 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 13.2 one after each other.

Table 13.2: Two-way classification: Response variables, group means, sample sizes in a tabular display.

Response variables

		W	
Z	1	$\ldots$	H
1	$\boldsymbol{Y}_{1,1} = \bigl(Y_{1,1,1},\,\ldots,\,Y_{1,1,n_{1,1}}\bigr)^\top$	$\vdots$	$\boldsymbol{Y}_{1,H} = \bigl(Y_{1,H,1},\,\ldots,\,Y_{1,H,n_{1,H}}\bigr)^\top$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$\boldsymbol{Y}_{G,1} = \bigl(Y_{G,1,1},\,\ldots,\,Y_{G,1,n_{G,1}}\bigr)^\top$	$\vdots$	$\boldsymbol{Y}_{G,H} = \bigl(Y_{G,H,1},\,\ldots,\,Y_{G,H,n_{G,H}}\bigr)^\top$

Group means

		W		
Z	1	$\ldots$	H	$\bullet$
1	$m_{1,1}$	$\vdots$	$m_{1,H}$	$\overline{m}_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$m_{G,1}$	$\vdots$	$m_{G,H}$	$\overline{m}_{G\bullet}$
$\bullet$	$\overline{m}_{\bullet 1}$	$\ldots$	$\overline{m}_{\bullet H}$	$\overline{m}$

Sample sizes

		W		
Z	1	$\ldots$	H	$\bullet$
1	$n_{1,1}$	$\vdots$	$n_{1,H}$	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$n_{G,1}$	$\vdots$	$n_{G,H}$	$n_{G\bullet}$
$\bullet$	$n_{\bullet 1}$	$\ldots$	$n_{\bullet H}$	n

The full response vector is $\boldsymbol{Y}$ and its (conditional, given $\mathbb{Z}$ and $\mathbb{W}$) mean is

$$\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z},\,\mathbb{W}\bigr) = \begin{pmatrix} m_{1,1}\,\mathbf{1}_{n_{1,1}} \\ \vdots \\ m_{G,1}\,\mathbf{1}_{n_{G,1}} \\ \vdots \\ m_{1,H}\,\mathbf{1}_{n_{1,H}} \\ \vdots \\ m_{G,H}\,\mathbf{1}_{n_{G,H}} \end{pmatrix} =: \boldsymbol{\mu}. \tag{13.4}$$

By a linear model, we can use a suitable parameterization of the mean vector $\boldsymbol{\mu}$ given by (13.4). At the same time, if a linear model is used, it is additionally assumed that

$$\mathsf{var}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{Z},\,\mathbb{W}\bigr) = \sigma^2\,\mathbf{I}_n. \tag{13.5}$$

13.2.1 Parameters of interest

Various quantities, all being linear combinations of the two-way classified group meansare 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

$$\overline{m}_{1\bullet},\,\ldots,\,\overline{m}_{G\bullet}, \qquad\qquad \text{and} \qquad \overline{m}_{\bullet 1},\,\ldots,\,\overline{m}_{\bullet H}.$$

- 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 $\overline{m}_{\bullet h}$ $(h = 1,\ldots,H)$ 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}.

(iv) Factor main effects, see Definition 13.2.

Definition 13.2 Factor main 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 $\eta^Z_1,\ldots,\eta^Z_G$ defined as*

$$\eta^Z_g \;:=\; \overline{m}_{g\bullet} \;-\; \overline{m}, \qquad g = 1,\ldots,G.$$

By main effects *of the factor W, we understand quantities $\eta^W_1,\ldots,\eta^W_H$ defined as*

$$\eta^W_h \;:=\; \overline{m}_{\bullet h} \;-\; \overline{m}, \qquad h = 1,\ldots,H.$$

Note. Each difference between the means of the means can also be written as difference between two main effects:

$$\begin{aligned}
\theta_{g_1,g_2\bullet} &= \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet} = \eta^Z_{g_1} - \eta^Z_{g_2}, \qquad g_1,\, g_2 = 1,\dots,G,\ g_1 \neq g_2,\\
\theta_{\bullet h_1,h_2} &= \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2} = \eta^W_{h_1} - \eta^W_{h_2}, \qquad h_1,\, h_2 = 1,\dots,H,\ h_1 \neq h_2.
\end{aligned}$$

13.2.2 Two-way ANOVA models

Depending on whether additivity can be assumed or not, or whether either of the two factors has an effect on the response expectation or not different models can be considered in a context of two-way classification. Each model just corresponds to different structure for the two-way classified group means. The models are as follows.

Interaction model

No structure is imposed on the group means. Under the assumption that (almost surely) $n_{g,h} > 0$ for each $g = 1, \ldots, G$, $h = 1, \ldots, H$, the corresponding regression space has (almost surely) a vector dimension $G \cdot H$ and the full-rank parameterization of the group means is achieved (see Section 5.4) by choosing the two (pseudo)contrast matrices $\mathbb{C}$ and $\mathbb{D}$ having the rows $\boldsymbol{c}_g^\top$, $g = 1, \ldots, G$ and $\boldsymbol{d}_h^\top$, $h = 1, \ldots, H$, respectively. The group means are then parameterized as

$$m_{g,h} \;=\; \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z + \boldsymbol{d}_h^\top \boldsymbol{\beta}^W + \bigl(\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top\bigr)\boldsymbol{\beta}^{ZW}, \qquad g = 1, \ldots, H,\ h = 1, \ldots, H, \tag{13.6}$$

where

$$\begin{aligned} &\beta_0, \quad \boldsymbol{\beta}^Z = \bigl(\beta_1^Z, \ldots, \beta_{G-1}^Z\bigr)^\top, \quad \boldsymbol{\beta}^W = \bigl(\beta_1^W, \ldots, \beta_{H-1}^W\bigr)^\top, \\ &\boldsymbol{\beta}^{ZW} = \bigl(\beta_{1,1}^{ZW}, \ldots, \beta_{G-1,1}^{ZW}, \ \ldots, \ \beta_{1,H-1}^{ZW}, \ldots, \beta_{G-1,H-1}^{ZW}\bigr)^\top \end{aligned}$$

are the regression parameters.

As in Section 5.4, it is useful to view the parameterization (13.6) as

$$m_{g,h} \;=\; \alpha_0 + \alpha_g^Z + \alpha_h^W + \alpha_{g,h}^{ZW}, \qquad g = 1, \ldots, H,\ h = 1, \ldots, H, \tag{13.7}$$

where

$$\begin{aligned} \alpha_0 &= \beta_0, & & \\ \alpha_g^Z &= \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z, & & g = 1, \ldots, G, \\ \alpha_h^W &= \boldsymbol{d}_h^\top \boldsymbol{\beta}^W, & & h = 1, \ldots, H, \\ \alpha_{g,h}^{ZW} &= \bigl(\boldsymbol{d}_h^\top \otimes \boldsymbol{c}_g^\top\bigr)\boldsymbol{\beta}^{ZW}, & & g = 1, \ldots, G,\ h = 1, \ldots, H. \end{aligned}$$

It has also been derived in Section 5.4 that with a choice of *sum contrasts* (see page 97), the "α" parameters attain the following interpretation:

$$\begin{array}{llll} \alpha_0 = \beta_0 & = \overline{m}, & & \\ \alpha_g^Z = \beta_g^Z & = \overline{m}_{g\bullet} - \overline{m} & = \eta_g^Z, & g = 1, \ldots, G-1, \\ \alpha_G^Z = -\displaystyle\sum_{g=1}^{G-1} \beta_g^Z & = \overline{m}_{G\bullet} - \overline{m} & = \eta_G^Z, & \\ \alpha_h^W = \beta_h^W & = \overline{m}_{\bullet h} - \overline{m}, & = \eta_h^W & h = 1, \ldots, H-1, \\ \alpha_H^W = -\displaystyle\sum_{h=1}^{H-1} \beta_h^W & = \overline{m}_{\bullet H} - \overline{m} & = \eta_H^W, & \\ \alpha_{g,h}^{ZW} & = m_{g,h} - \overline{m}_{g\bullet} - \overline{m}_{\bullet h} + \overline{m}, & & g = 1, \ldots, G,\ h = 1, \ldots, H. \end{array}$$

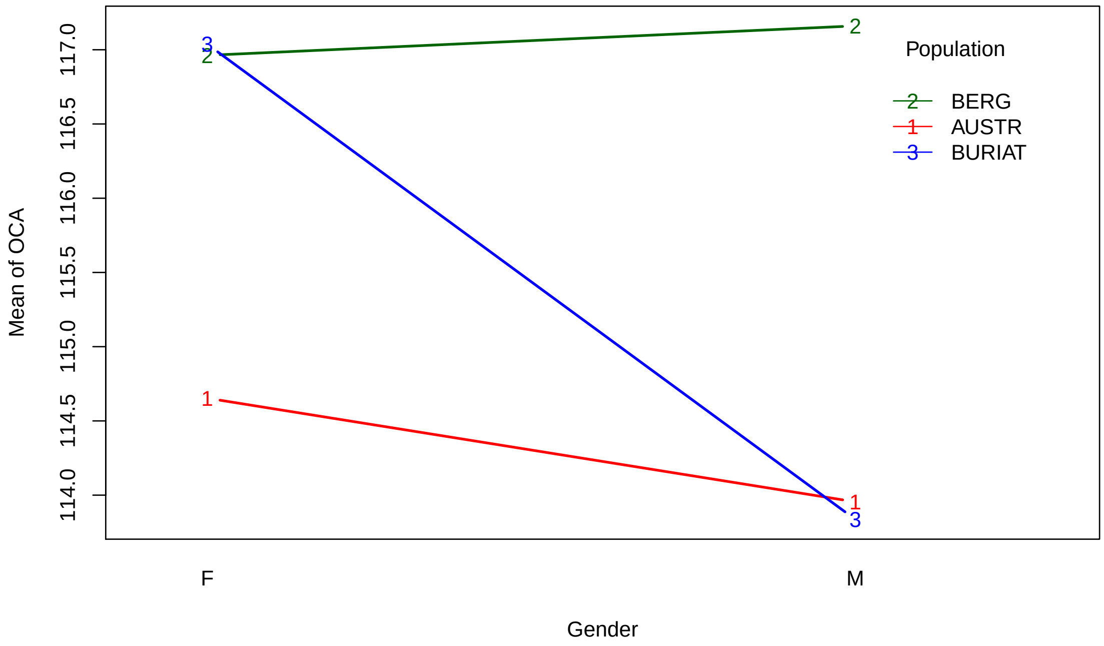
In the following, let symbols Z and W denote the terms in the model matrix that correspond to the coefficients $\boldsymbol{\alpha}^Z$ or $\boldsymbol{\beta}^Z$, and $\boldsymbol{\alpha}^W$ or $\boldsymbol{\beta}^W$, respectively. Let further Z : W denote the terms corresponding to the interaction coefficients $\boldsymbol{\alpha}^{ZW}$ or $\boldsymbol{\beta}^{ZW}$. The interaction model will then symbolically be written as

$$\mathsf{M}_{ZW}\colon \ \sim \mathsf{Z} + \mathsf{W} + \mathsf{Z}\!:\!\mathsf{W}.$$

Illustrations

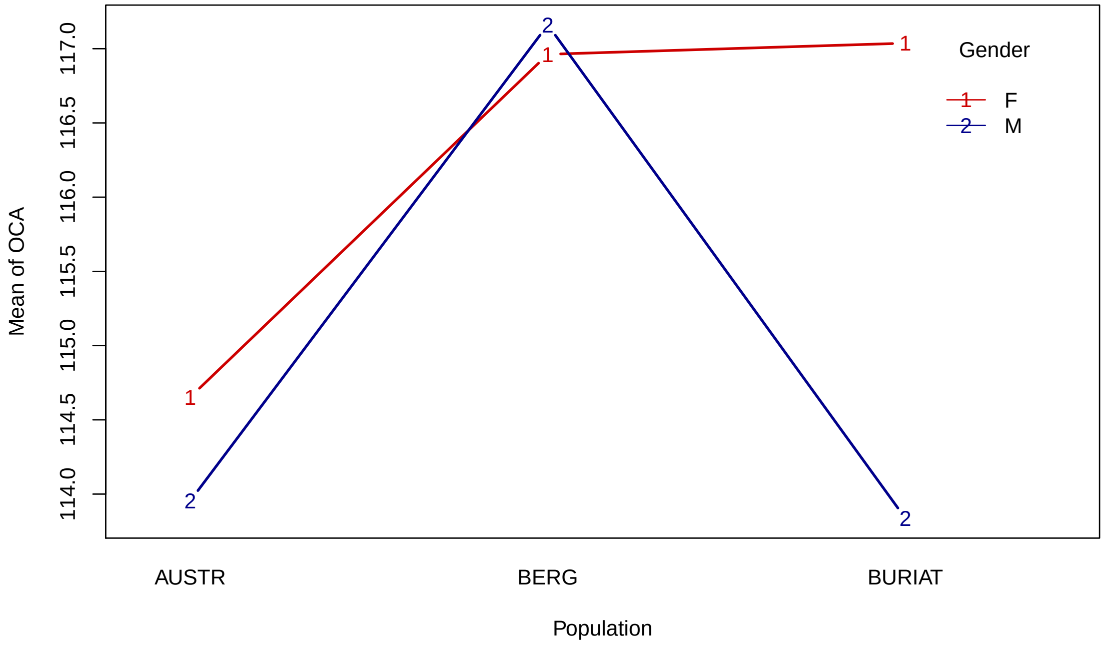
`Howells` ($n = 289$)

`oca` **(occipital angle)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$)



`Howells` ($n = 289$)

`oca` **(occipital angle)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$)



Additive model

It is obtained as a submodel of the interaction model (13.6) where it is requested

$$\alpha^{ZW}_{1,1} = \cdots = \alpha^{ZW}_{G,H},$$

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

$$\begin{aligned} m_{g,h} &= \alpha_0 + \alpha^Z_g + \alpha^W_h, \\ &= \beta_0 + \boldsymbol{c}_g^\top \boldsymbol{\beta}^Z + \boldsymbol{d}_h^\top \boldsymbol{\beta}^W, \qquad g = 1,\dots,H,\ h = 1,\dots,H. \end{aligned} \tag{13.8}$$

Under the condition that $n_{g,h} > 0$ (almost surely) for all g, h, the rank of the linear model with the two-way classified group means that satisfy (13.8), is $G + H - 1$ (almost surely). The additive model will symbolically be written as

$$\mathsf{M}_{Z+W}:\ \sim \mathsf{Z} + \mathsf{W}.$$

Note. It can easily be shown that $n_{g\bullet} > 0$ for all $g = 1,\dots,G$ and $n_{\bullet h} > 0$ for all $h = 1,\dots,H$ suffice to get a rank of the related linear model being still $G + H - 1$. This guarantees, among the other things, that all linear combinations of the regression coefficients of the additive model can still be estimated by a method of least squares under a weaker requirement $n_{g\bullet}$ for all $g = 1,\dots,G$ and $n_{\bullet h}$ for all $h = 1,\dots,H$. 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 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 (factor settings).

The additive model implies the following properties for the two-way classified group means:

(i) for each $g_1 \neq g_2$, $g_1, g_2 \in \{1,\dots,G\}$, the difference $m_{g_1,h} - m_{g_2,h}$ does not depend on a value of $h \in \{1,\dots,H\}$ and is equal to the difference between the corresponding means of the means by the first factor, i.e.,

$$m_{g_1,h} - m_{g_2,h} = \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet} = \eta^Z_{g_1} - \eta^Z_{g_2} = \theta_{g_1,g_2\bullet},$$

which is expressed using the parameterizations (13.8) as

$$\theta_{g_1,g_2\bullet} = \alpha^Z_{g_1} - \alpha^Z_{g_2} = \big(\boldsymbol{c}_{g_1} - \boldsymbol{c}_{g_2}\big)^\top \boldsymbol{\beta}^Z;$$

(ii) for each $h_1 \neq h_2$, $h_1, h_2 \in \{1,\dots,H\}$, the difference $m_{g,h_1} - m_{g,h_2}$ does not depend on a value of $g \in \{1,\dots,G\}$ and is equal to the difference between the corresponding means of the means by the second factor, i.e.,

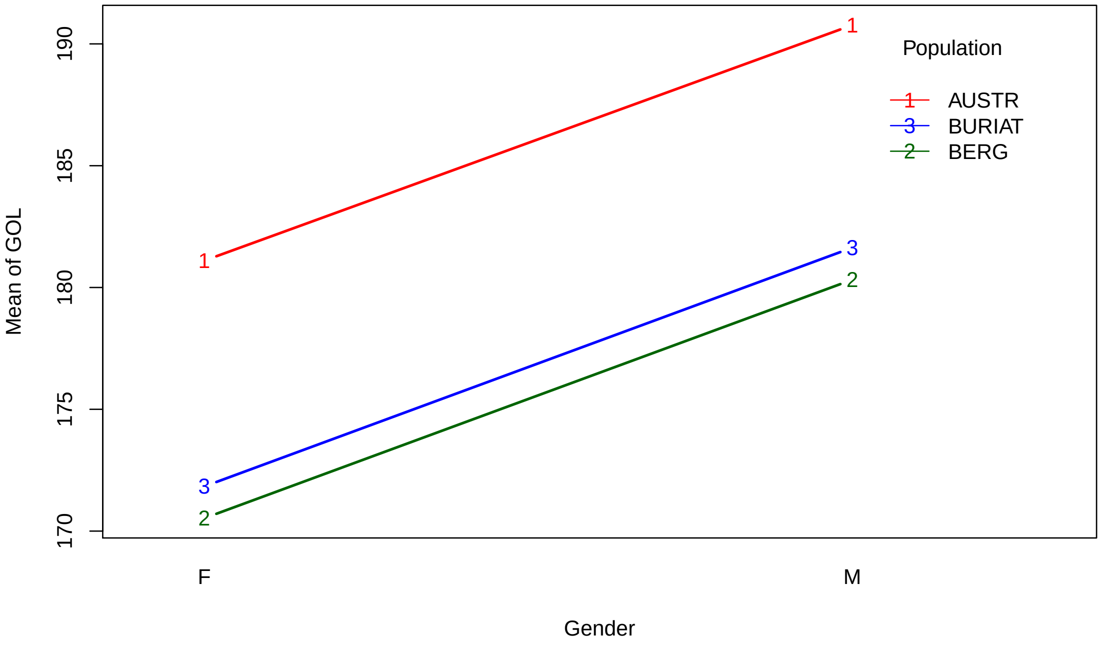
$$m_{g,h_1} - m_{g,h_2} = \overline{m}_{\bullet h_1} - \overline{m}_{\bullet h_2} = \eta^W_{h_1} - \eta^W_{h_2} = \theta_{\bullet h_1,h_2},$$

which is expressed using the parameterizations (13.8) as

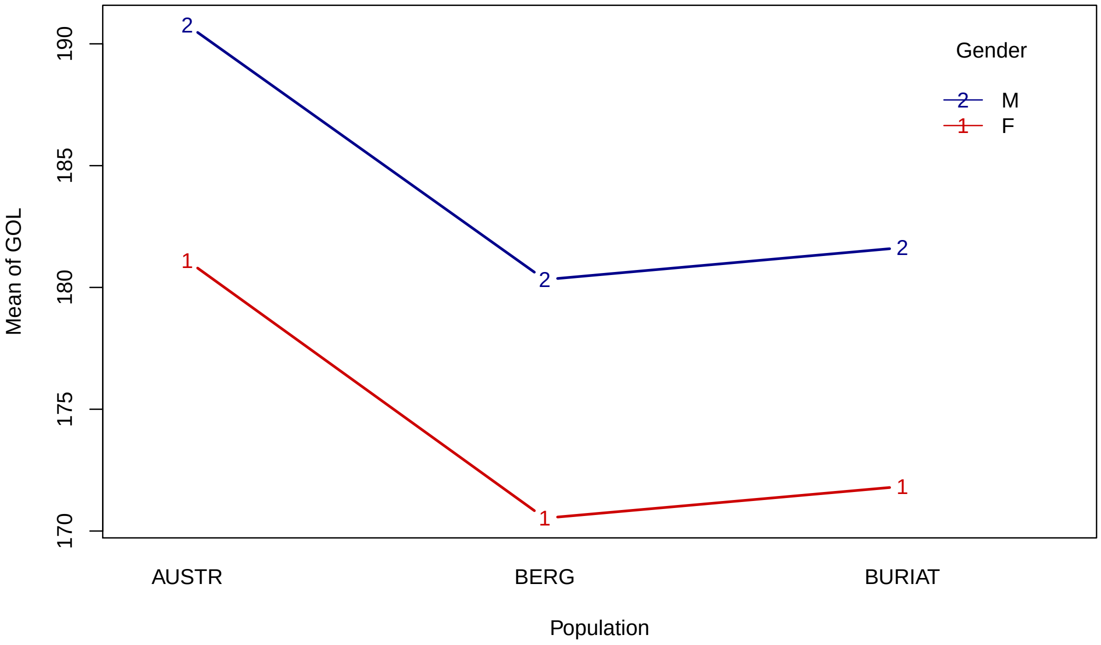
$$\theta_{\bullet h_1,h_2} = \alpha^W_{h_1} - \alpha^W_{h_2} = \big(\boldsymbol{d}_{h_1} - \boldsymbol{d}_{h_2}\big)^\top \boldsymbol{\beta}^W.$$

Illustrations

Howells ($n = 289$)
gol **(glabell-occipital length)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$)



Howells ($n = 289$)
gol **(glabell-occipital length)** $\sim$ **gender** ($G = 2$) **and population** ($H = 3$)



Finally, remember that it has been derived in Section 5.4 that with a choice of *sum contrasts* (see page 97), the "α" parameters attain the following interpretation:

$$\begin{aligned}
\alpha_0 &= \beta_0 &&= \overline{m}, && \\
\alpha_g^Z &= \beta_g^Z &&= \overline{m}_{g\bullet} - \overline{m} \;= \eta_g^Z, && g = 1,\dots,G-1, \\
\alpha_G^Z &= -\sum_{g=1}^{G-1}\beta_g^Z &&= \overline{m}_{G\bullet} - \overline{m} \;= \eta_G^Z, && \\
\alpha_h^W &= \beta_h^W &&= \overline{m}_{\bullet h} - \overline{m}, = \eta_h^W && h = 1,\dots,H-1, \\
\alpha_H^W &= -\sum_{h=1}^{H-1}\beta_h^W &&= \overline{m}_{\bullet H} - \overline{m} = \eta_H^W. &&
\end{aligned}$$

Model of effect of Z only

It is obtained as a submodel of the additive model (13.8) by requesting

$$\alpha_1^W = \cdots = \alpha_H^W,$$

which in the full-rank parameterization corresponds to requesting

$$\boldsymbol{\beta}^W = \mathbf{0}_{H-1}.$$

Hence the group means can be written as

$$\begin{aligned}
m_{g,h} &= \alpha_0 + \;\; \alpha_g^Z, \\
&= \beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta}^Z, \qquad g = 1,\dots,H,\; h = 1,\dots,H.
\end{aligned} \tag{13.9}$$

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 $g = 1,\dots,G$. The model of effect of Z only will symbolically be written as

$$\mathsf{M}_Z: \sim \mathsf{Z}.$$

The two-way classified group means then satisfy

(i) For each $g = 1,\dots,G$, $m_{g,1} = \cdots = m_{g,H} = \overline{m}_{g\bullet}$.

(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 (13.8) 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

$$\begin{aligned}
m_{g,h} &= \alpha_0 + \;\; \alpha_h^W, \\
&= \beta_0 + \boldsymbol{d}_h^\top\boldsymbol{\beta}^W, \qquad g = 1,\dots,H,\; h = 1,\dots,H.
\end{aligned} \tag{13.10}$$

The model of effect of W only will symbolically be written as

$$\mathsf{M}_W: \sim \mathsf{W}.$$

Table 13.3: 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 $g = 1, \dots, G$, $h = 1, \dots, H$
$\mathsf{M}_{Z+W}\colon\ \sim \mathsf{Z} + \mathsf{W}$	$G + H - 1$	$n_{g\bullet} > 0$ for all $g = 1, \dots, G$, $n_{\bullet h} > 0$ for all $h = 1, \dots, H$
$\mathsf{M}_{Z}\colon\ \sim \mathsf{Z}$	G	$n_{g\bullet} > 0$ for all $g = 1, \dots, G$
$\mathsf{M}_{W}\colon\ \sim \mathsf{W}$	H	$n_{\bullet h} > 0$ for all $h = 1, \dots, H$
$\mathsf{M}_{0}\colon\ \sim 1$	1	$n > 0$

Intercept only model

This is a submodel of either the model (13.9) 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 (13.10) 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

$$\begin{aligned} m_{g,h} &= \alpha_0, \\ &= \beta_0, \qquad g = 1, \dots, H,\ h = 1, \dots, H. \end{aligned}$$

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 13.3. The considered models form two sequences 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 8 (F-tests on submodels).

13.2.3 Least squares estimation

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).*

$$\overline{Y}_{g,h\bullet} := \frac{1}{n_{g,h}}\sum_{j=1}^{n_{g,h}} Y_{g,h,j}, \qquad g=1,\dots,G,\ h=1,\,\dots,H,$$

$$\overline{Y}_{g\bullet} := \frac{1}{n_{g\bullet}}\sum_{h=1}^{H}\sum_{j=1}^{n_{g,h}} Y_{g,h,j} = \frac{1}{n_{g\bullet}}\sum_{h=1}^{H} n_{g,h}\,\overline{Y}_{g,h\bullet}, \qquad g=1,\dots,G,$$

$$\overline{Y}_{\bullet h} := \frac{1}{n_{\bullet h}}\sum_{g=1}^{G}\sum_{j=1}^{n_{g,h}} Y_{g,h,j} = \frac{1}{n_{\bullet h}}\sum_{g=1}^{G} n_{g,h}\,\overline{Y}_{g,h\bullet}, \qquad h=1,\dots,H,$$

$$\overline{Y} := \frac{1}{n}\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{j=1}^{n_{g,h}} Y_{g,h,j} = \frac{1}{n}\sum_{g=1}^{G} n_{g\bullet}\,\overline{Y}_{g\bullet} = \frac{1}{n}\sum_{h=1}^{H} n_{\bullet h}\,\overline{Y}_{\bullet h}.$$

As usual, $\widehat{m}_{g,h}$, $g=1,\dots,G$, $h=1,\dots,H$, denote the LSE of the two-way classified group means and $\widehat{\boldsymbol{m}} = \bigl(\widehat{m}_{1,1},\,\dots,\,\widehat{m}_{G,H}\bigr)^\top$.

Lemma 13.2 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 $g=1,\dots,G$, $h=1,\dots,H$, $j=1,\dots,n_{g,h}$).

(i) ***Interaction model*** $\mathsf{M}_{ZW}:\ \sim \mathsf{Z}+\mathsf{W}+\mathsf{Z}{:}\mathsf{W}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{g,h\bullet}.$$

(ii) ***Additive model*** $\mathsf{M}_{Z+W}:\ \sim \mathsf{Z}+\mathsf{W}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y},$$

but only in case of balanced *data*[5] *($n_{g,h} = J$ for all $g=1,\dots,G$, $h=1,\dots,H$).*

(iii) ***Model of effect of Z only*** $\mathsf{M}_{Z}:\ \sim \mathsf{Z}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{g\bullet}.$$

(iv) ***Model of effect of W only*** $\mathsf{M}_{W}:\ \sim \mathsf{W}$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}_{\bullet h}.$$

(v) ***Intercept only model*** $\mathsf{M}_{0}:\ \sim 1$

$$\widehat{m}_{g,h} = \widehat{Y}_{g,h,j} = \overline{Y}.$$

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.

[5] *vyvážená data*

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 (minimizing the sum of squares) corresponding to the parameterization

$$m_{g,h} = \alpha_0 + \alpha_g^Z + \alpha_h^W, \qquad g = 1, \dots, G,\ h = 1, \dots, H.$$

while imposing the constraints

$$\sum_{g=1}^{G} \alpha_g^Z = 0, \qquad \sum_{h=1}^{H} \alpha_h^W = 0.$$

This corresponds to the full-rank parameterization while using the *sum contrasts* parameterization.

For the *additive* model with the *balanced* data ($n_{g,h} = J$ for all $g = 1, \dots, G$, $h = 1, \dots, H$):

- Sum of squares to be minimized

$$\mathsf{SS}(\boldsymbol{\alpha}) = \sum_g \sum_h \sum_j \bigl(Y_{g,h,j} - \alpha_0 - \alpha_g^Z - \alpha_h^W\bigr)^2.$$

- Normal equations $\equiv$ derivatives of $\mathsf{SS}(\boldsymbol{\alpha})$ divided by (-2) and set to zero:

$$\begin{aligned}
\sum_g \sum_h \sum_j Y_{g,h,j} - GHJ\alpha_0 - HJ\sum_g \alpha_g^Z - GJ\sum_h \alpha_h^W &= 0, \\
\sum_h \sum_j Y_{g,h,j} - HJ\alpha_0 - HJ\alpha_g^Z - J\sum_h \alpha_h^W &= 0, \qquad g = 1, \dots, G, \\
\sum_g \sum_j Y_{g,h,j} - GJ\alpha_0 - J\sum_g \alpha_g^Z - GJ\alpha_h^W &= 0, \qquad h = 1, \dots, H.
\end{aligned}$$

- After exploiting the identifying constraints:

$$\begin{aligned}
\sum_g \sum_h \sum_j Y_{g,h,j} - GHJ\alpha_0 &= 0, \\
\sum_h \sum_j Y_{g,h,j} - HJ\alpha_0 - HJ\alpha_g^Z &= 0, \qquad g = 1, \dots, G, \\
\sum_g \sum_j Y_{g,h,j} - GJ\alpha_0 - GJ\alpha_h^W &= 0, \qquad h = 1, \dots, H.
\end{aligned}$$

- Hence $\widehat{\alpha}_0 = \overline{Y}$,

$$\begin{aligned}
\widehat{\alpha}_g^Z &= \overline{Y}_{g\bullet} - \overline{Y}, \qquad g = 1, \dots, G, \\
\widehat{\alpha}_h^W &= \overline{Y}_{\bullet h} - \overline{Y}, \qquad h = 1, \dots, H.
\end{aligned}$$

- And then $\widehat{m}_{g,h} = \widehat{\alpha}_0 + \widehat{\alpha}_g^Z + \widehat{\alpha}_h^W = \overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y}$,

$$g = 1, \dots, G,\ h = 1, \dots, H.$$

❑

Consequence of Lemma 13.2: 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 $g = 1, \dots, G$, $h = 1, \dots, H$), the LSE of the means of the means by the first factor (parameters $\overline{m}_{1\bullet}, \dots, \overline{m}_{G\bullet}$) or by the second factor (parameters $\overline{m}_{\bullet 1}, \dots, \overline{m}_{\bullet H}$) satisfy in both the interaction and the additive two-way ANOVA linear models the following:

$$\widehat{\overline{m}}_{g\bullet} = \overline{Y}_{g\bullet}, \qquad g = 1, \dots, G,$$
$$\widehat{\overline{m}}_{\bullet h} = \overline{Y}_{\bullet h}, \qquad h = 1, \dots, H.$$

If additionally normality is assumed then $\widehat{\overline{\boldsymbol{m}}}^Z := \bigl(\widehat{\overline{m}}_{1\bullet}, \dots, \widehat{\overline{m}}_{G\bullet}\bigr)^\top$ and $\widehat{\overline{\boldsymbol{m}}}^W := \bigl(\widehat{\overline{m}}_{\bullet 1}, \dots, \widehat{\overline{m}}_{\bullet H}\bigr)^\top$ satisfy

$$\widehat{\overline{\boldsymbol{m}}}^Z \mid \mathbb{Z}, \mathbb{W} \sim \mathcal{N}_G\bigl(\overline{\boldsymbol{m}}^Z, \sigma^2\,\mathbb{V}^Z\bigr), \qquad \widehat{\overline{\boldsymbol{m}}}^W \mid \mathbb{Z}, \mathbb{W} \sim \mathcal{N}_H\bigl(\overline{\boldsymbol{m}}^W, \sigma^2\,\mathbb{V}^W\bigr),$$

where

$$\overline{\boldsymbol{m}}^Z = \begin{pmatrix} \overline{m}_{1\bullet} \\ \vdots \\ \overline{m}_{G\bullet} \end{pmatrix}, \quad \mathbb{V}^Z = \begin{pmatrix} \frac{1}{JH} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{JH} \end{pmatrix},$$

$$\overline{\boldsymbol{m}}^W = \begin{pmatrix} \overline{m}_{\bullet 1} \\ \vdots \\ \overline{m}_{\bullet H} \end{pmatrix}, \quad \mathbb{V}^W = \begin{pmatrix} \frac{1}{JG} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{JG} \end{pmatrix}.$$

Proof. All parameters $\overline{m}_{g\bullet}$, $g = 1, \dots, G$, and $\overline{m}_{\bullet h}$, $h = 1, \dots, H$ are linear combinations of the group means (of the response mean vector $\boldsymbol{\mu} = \mathbb{E}\bigl(\boldsymbol{Y} \bigm| \mathbb{Z}, \mathbb{W}\bigr)$) and hence their LSE is 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 $\overline{m}_{g\bullet}$, $g = 1, \dots, G$):

(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**

$$\widehat{\overline{m}}_{g\bullet} = \frac{1}{H}\sum_{h=1}^{H} \widehat{m}_{g,h} = \frac{1}{H}\sum_{h=1}^{H}\bigl(\overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y}\bigr)$$
$$= \overline{Y}_{g\bullet} + \frac{1}{H}\sum_{h=1}^{H} \overline{Y}_{\bullet h} - \overline{Y} = \overline{Y}_{g\bullet} + \frac{1}{H\,G\,J}\sum_{h=1}^{H} G\,J\overline{Y}_{\bullet h} - \overline{Y}$$
$$= \overline{Y}_{g\bullet} + \underbrace{\frac{1}{n}\sum_{h=1}^{H} n_{\bullet h}\overline{Y}_{\bullet h}}_{\overline{Y}} - \overline{Y} = \overline{Y}_{g\bullet}.$$

Further, $\mathbb{E}\bigl(\overline{Y}_{g\bullet} \bigm| \mathbb{Z}, \mathbb{W}\bigr) = \overline{m}_{g\bullet}$ follows from properties of the LSE which are unbiased or from direct calculation. Next,

$$\mathsf{var}\bigl(\overline{Y}_{g\bullet} \bigm| \mathbb{Z}, \mathbb{W}\bigr) = \mathsf{var}\biggl[\frac{1}{J\,H}\sum_{h=1}^{H}\sum_{j=1}^{J} Y_{g,h,j} \biggm| \mathbb{Z}, \mathbb{W}\biggr] = \frac{\sigma^2}{J\,H}$$

follows from the linear model assumption $\mathsf{var}\big(\boldsymbol{Y}\,\big|\,\mathbb{Z},\,\mathbb{W}\big) = \sigma^2\,\mathbf{I}_n$. This also implies $\mathsf{cov}\big(\overline{Y}_{g_1\bullet},\,\overline{Y}_{g_2\bullet}\,\big|\,\mathbb{Z},\,\mathbb{W}\big) = 0$ for $g_1 \neq g_2$ and zero off-diagonal elements of the matrix $\mathbb{V}^Z$.

Finally, normality of $\overline{Y}_{g\bullet}$ in case of a normal linear model, follows from the general LSE theory.

❑

13.2.4 Sums of squares and ANOVA tables with balanced data

Sums of squares

As already mentioned in Section 13.2.2, 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 8.1). In particular, in case of *balanced* data ($n_{g,h} = J$, $g = 1, \dots, G$, $h = 1, \dots, H$), we get

$$\mathsf{SS}\big(\mathsf{Z}+\mathsf{W}+\mathsf{Z}{:}\mathsf{W} \,\big|\, \mathsf{Z}+\mathsf{W}\big) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g,h\bullet} - \overline{Y}_{g\bullet} - \overline{Y}_{\bullet h} + \overline{Y}\right)^2,$$

$$\mathsf{SS}\big(\mathsf{Z}+\mathsf{W} \,\big|\, \mathsf{W}\big) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y} - \overline{Y}_{\bullet h}\right)^2 = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet} - \overline{Y}\right)^2,$$

$$\mathsf{SS}\big(\mathsf{Z}+\mathsf{W} \,\big|\, \mathsf{Z}\big) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet} + \overline{Y}_{\bullet h} - \overline{Y} - \overline{Y}_{g\bullet}\right)^2 = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{\bullet h} - \overline{Y}\right)^2,$$

$$\mathsf{SS}\big(\mathsf{Z} \,\big|\, 1\big) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{g\bullet} - \overline{Y}\right)^2,$$

$$\mathsf{SS}\big(\mathsf{W} \,\big|\, 1\big) = \sum_{g=1}^{G}\sum_{h=1}^{H} J\left(\overline{Y}_{\bullet h} - \overline{Y}\right)^2.$$

We see,

$$\mathsf{SS}\big(\mathsf{Z}+\mathsf{W} \,\big|\, \mathsf{W}\big) = \mathsf{SS}\big(\mathsf{Z} \,\big|\, 1\big),$$
$$\mathsf{SS}\big(\mathsf{Z}+\mathsf{W} \,\big|\, \mathsf{Z}\big) = \mathsf{SS}\big(\mathsf{W} \,\big|\, 1\big).$$

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 13.3 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}	$\star$		

13.3 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 4.4. 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 or in Seber and Lee (2003, Section 8.6).

Chapter 14

Simultaneous Inference in a Linear Model

As usual, we will assume that data are represented by a set of n random vectors $\bigl(Y_i,\, \boldsymbol{X}_i^\top\bigr)^\top$, $\boldsymbol{X}_i = \bigl(X_{i,0}, \dots, X_{i,k-1}\bigr)^\top$, $i = 1, \dots, n$, that satisfy a linear model. Throughout the whole chapter, full rank and *normality* will also be assumed. That is, we assume that

$$\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr), \qquad \mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k < n,$$

where $\boldsymbol{Y} = \bigl(Y_1, \dots, Y_n\bigr)^\top$, $\mathbb{X}$ is a matrix with vectors $\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_n^\top$ in its rows and $\boldsymbol{\beta} = \bigl(\beta_0, \dots, \beta_{k-1}\bigr)^\top \in \mathbb{R}^k$ and $\sigma^2 > 0$ are unknown parameters. Further, we will assume that a matrix $\mathbb{L}_{m\times k}$ $(m > 1)$ with rows $\mathbf{l}_1^\top, \dots, \mathbf{l}_m^\top$ (all non-zero vectors) is given:

$$\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta} = \bigl(\mathbf{l}_1^\top\boldsymbol{\beta}, \dots, \mathbf{l}_m^\top\boldsymbol{\beta}\bigr)^\top = \bigl(\theta_1, \dots, \theta_m\bigr)^\top$$

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 the null hypothesis H$_0$: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ for given $\boldsymbol{\theta}^0 \in \mathbb{R}^m$.

14.1 Basic simultaneous inference

If matrix $\mathbb{L}_{m \times k}$ is such that

(i) $m \leq k$;

(ii) its rows, i.e., vectors $\mathbf{l}_1, \dots, \mathbf{l}_m \in \mathbb{R}^k$ 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 6.2 (Least squares estimators under the normality). It provides a confidence region for $\boldsymbol{\theta}$ with a coverage of $1 - \alpha$ which is

$$\Bigl\{\boldsymbol{\theta} \in \mathbb{R}^m : \bigl(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\bigr)^\top \Bigl\{\mathsf{MS}_e\, \mathbb{L}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{L}^\top\Bigr\}^{-1} \bigl(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\bigr) \;<\; m\,\mathcal{F}_{m,n-k}(1-\alpha)\Bigr\}, \tag{14.1}$$

where $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\theta}$. The null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ is tested using the statistic

$$Q_0 = \frac{1}{m}\bigl(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\bigr)^\top \Bigl\{\mathsf{MS}_e\, \mathbb{L}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{L}^\top\Bigr\}^{-1} \bigl(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\bigr), \tag{14.2}$$

which under the null hypothesis follows an $\mathcal{F}_{m,n-k}$ distribution and the critical region of a test on the level of α is

$$C(\alpha) = \bigl[\mathcal{F}_{m,n-k}(1-\alpha),\ \infty\bigr). \tag{14.3}$$

The P-value if $Q_0 = q_0$ is then given as $p = 1 - \mathsf{CDF}_{\mathcal{F},\,m,\,n-k}(q_0)$. Note that the confidence region (14.1) and the test based on the statistic Q_0 and the critical region (14.3) 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 (14.1) with a coverage $1 - \alpha$.

14.2 Multiple comparison procedures

14.2.1 Multiple testing

The null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ ($\boldsymbol{\theta}^0 = \bigl(\theta_1^0, \ldots, \theta_m^0\bigr)^\top$) on a vector parameter $\boldsymbol{\theta} = \bigl(\theta_1, \ldots, \theta_m\bigr)^\top$ 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 14.1 Multiple testing problem, elementary null hypotheses, global null hypothesis.

A testing problem with the null hypothesis

$$H_0: \quad \theta_1 = \theta_1^0 \quad \& \quad \ldots \quad \& \quad \theta_m = \theta_m^0, \tag{14.4}$$

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 14.1 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 14.1 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 (14.4) will always be given by a complement of the parameter space under the global null hypothesis, i.e.,

  $$\begin{aligned} \mathrm{H}_0^{\complement}: \ \theta_1 \neq \theta_1^0 \quad \mathrm{OR} \quad \ldots \quad \mathrm{OR} \quad \theta_m \neq \theta_m^0, \\ \equiv \quad \mathrm{H}_1^{\complement} \quad \mathrm{OR} \quad \ldots \quad \mathrm{OR} \quad \mathrm{H}_m^{\complement}, \end{aligned}$$

  where $\mathrm{H}_j^{\complement}: \ \theta_j \neq \theta_j^0$, $j = 1, \ldots, m$. We will also write

  $$\mathrm{H}_0^{\complement} = \bigcup_{j=1}^{m} \mathrm{H}_j^{\complement}.$$

[1] *problém vícenásobného testování* [2] *elementární hypotézy*

- Different ways of indexing the elementary null hypotheses will also be used (e.g., a double subscript) depending on a problem at hand.

Example 14.1 (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 $\mathcal{Z} = \{1,\,\dots,\,G\}$, where the regression space $\mathcal{M}\big(\mathbb{X}\big)$ of a vector dimension G parameterizes the one-way classified group means

$$m_1 := \mathbb{E}\big(Y \,\big|\, Z = 1\big),\ \dots,\ m_G = \mathbb{E}\big(Y \,\big|\, Z = G\big).$$

If we restrict ourselves to full-rank parameterizations (see Section 4.4.3), the regression coefficients vector is $\boldsymbol{\beta} = \big(\beta_0,\, \boldsymbol{\beta}^{Z^\top}\big)^\top$, $\boldsymbol{\beta}^Z = \big(\beta_1,\,\dots,\,\beta_{G-1}\big)^\top$ and the group means are parameterized as

$$m_g = \beta_0 + \boldsymbol{c}_g^\top\boldsymbol{\beta}^Z, \qquad g = 1,\dots,G,$$

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):

$$H_{1,2}\colon m_1 = m_2, \quad \dots, \quad H_{G-1,G}\colon m_{G-1} = m_G.$$

The elementary null hypotheses can now be written in terms of a vector estimable parameter

$$\boldsymbol{\theta} = \big(\theta_{1,2},\,\dots,\,\theta_{G-1,G}\big)^\top,$$

$$\theta_{g,h} = m_g - m_h = \big(\boldsymbol{c}_g - \boldsymbol{c}_h\big)^\top\boldsymbol{\beta}^Z, \qquad g = 1,\dots,G-1,\ h = g+1,\dots,G$$

as

$$H_{1,2}\colon \theta_{1,2} = 0, \quad \dots, \quad H_{G-1,G}\colon \theta_{G-1,G} = 0,$$

or written directly in term of the group means as

$$H_{1,2}\colon m_1 - m_2 = 0, \quad \dots, \quad H_{G-1,G}\colon m_{G-1} - m_G = 0,$$

The global null hypothesis is H_0: $\boldsymbol{\theta} = \boldsymbol{0}$, 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 (14.1) for $\boldsymbol{\theta}$;

(ii) nor use the test statistic (14.2) to test H_0: $\boldsymbol{\theta} = \boldsymbol{0}$.

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.

14.2.2 Simultaneous confidence intervals

Suppose that a distribution of the random vector $\boldsymbol{D}$ depends on a (vector) parameter $\boldsymbol{\theta} = \bigl(\theta_1, \ldots, \theta_m\bigr)^\top \in \Theta_1 \times \cdots \times \Theta_m = \Theta \subseteq \mathbb{R}^m$.

Definition 14.2 Simultaneous confidence intervals.

(Random) intervals $\bigl(\theta_j^L,\, \theta_j^U\bigr)$, $j = 1, \ldots, m$, *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* $\boldsymbol{\theta}^0 = \bigl(\theta_1^0, \ldots, \theta_m^0\bigr)^\top \in \Theta$,

$$\mathsf{P}\Bigl(\bigl(\theta_1^L,\, \theta_1^U\bigr) \times \cdots \times \bigl(\theta_m^L,\, \theta_m^U\bigr) \ni \boldsymbol{\theta}^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \geq 1 - \alpha.$$

Notes.

- The condition in the definition can also be written as

$$\mathsf{P}\Bigl(\forall\, j = 1, \ldots, m : \bigl(\theta_j^L,\, \theta_j^U\bigr) \ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \geq 1 - \alpha.$$

- The product of the simultaneous confidence intervals indeed forms a confidence region in a classical sense.

Example 14.2 (Bonferroni simultaneous confidence intervals).

Let for each $j = 1, \ldots, m$, $\bigl(\theta_j^L,\, \theta_j^U\bigr)$ *be a classical confidence interval for* θ_j *with a coverage of* $1 - \frac{\alpha}{m}$. *That is,*

$$\forall\, j = 1, \ldots, m,\ \forall\, \theta_j^0 \in \Theta_j : \quad \mathsf{P}\Bigl(\bigl(\theta_j^L,\, \theta_j^U\bigr) \ni \theta_j^0;\ \theta_j = \theta_j^0\Bigr) \geq 1 - \frac{\alpha}{m}.$$

We then have

$$\forall\, j = 1, \ldots, m,\ \forall\, \theta_j^0 \in \Theta_j : \quad \mathsf{P}\Bigl(\bigl(\theta_j^L,\, \theta_j^U\bigr) \not\ni \theta_j^0;\ \theta_j = \theta_j^0\Bigr) \leq \frac{\alpha}{m}.$$

Further, using elementary property of a probability (for any $\boldsymbol{\theta}^0 \in \Theta$*)*

$$\mathsf{P}\Bigl(\exists\, j = 1, \ldots, m : \quad \bigl(\theta_j^L,\, \theta_j^U\bigr) \not\ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \ \leq\ \sum_{j=1}^m \mathsf{P}\Bigl(\bigl(\theta_j^L,\, \theta_j^U\bigr) \not\ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr)$$

$$\leq\ \sum_{j=1}^m \frac{\alpha}{m}\ =\ \alpha.$$

Hence,

$$\mathsf{P}\Bigl(\forall\, j = 1, \ldots, m : \quad \bigl(\theta_j^L,\, \theta_j^U\bigr) \ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr)\ \geq\ 1 - \alpha.$$

That is, intervals $\bigl(\theta_j^L,\, \theta_j^U\bigr)$, $j = 1, \ldots, m$, *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* $1 - \alpha$.

[3] *simultánní intervaly spolehlivosti*

14.2.3 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 $\boldsymbol{\theta} = \bigl(\theta_1, \ldots, \theta_m\bigr)^\top \in \Theta_1 \times \cdots \times \Theta_m = \Theta \subseteq \mathbb{R}^m$. Let for each $0 < \alpha < 1$ a procedure be given to construct the simultaneous confidence intervals $\bigl(\theta_j^L(\alpha), \theta_j^U(\alpha)\bigr)$, $j = 1, \ldots, m$, 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

$$1 - \alpha_1 < 1 - \alpha_2 \qquad \Longrightarrow \qquad \bigl(\theta_j^L(\alpha_1), \theta_j^U(\alpha_1)\bigr) \subseteq \bigl(\theta_j^L(\alpha_2), \theta_j^U(\alpha_2)\bigr).$$

Definition 14.3 Multiple comparison procedure.

Multiple comparison procedure (MCP)[4] *for a multiple testing problem with the elementary null hypotheses* H_j: $\theta_j = \theta_j^0$, $j = 1, \ldots, m$, *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* $\mathrm{H}_0\colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$ *if and only if*

$$\bigl(\theta_1^L(\alpha), \theta_1^U(\alpha)\bigr) \times \cdots \times \bigl(\theta_m^L(\alpha), \theta_m^U(\alpha)\bigr) \not\ni \boldsymbol{\theta}^0;$$

(ii) *for* $j = 1, \ldots, m$, *rejects the* j*th elementary hypothesis* $H_j\ :\ \theta_j = \theta_j^0$ *if and only if*

$$\bigl(\theta_j^L(\alpha), \theta_j^U(\alpha)\bigr) \not\ni \theta_j^0.$$

Note. Since $\bigl(\theta_1^L(\alpha), \theta_1^U(\alpha)\bigr) \times \cdots \times \bigl(\theta_m^L(\alpha), \theta_m^U(\alpha)\bigr) \not\ni \boldsymbol{\theta}^0$ if and only if there exists $j = 1, \ldots, m$, such that $\bigl(\theta_j^L(\alpha), \theta_j^U(\alpha)\bigr) \not\ni \theta_j^0$, the MCP rejects, for given $0 < \alpha < 1$, the global null hypothesis $\mathrm{H}_0\colon\ \boldsymbol{\theta} = \boldsymbol{\theta}^0$ 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.,

$$\mathsf{P}\bigl(\mathrm{H}_0 \text{ rejected; } \boldsymbol{\theta} = \boldsymbol{\theta}^0\bigr) \le \alpha;$$

(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.,

$$\mathsf{P}\bigl(\mathrm{H}_j \text{ rejected; } \theta_j = \theta_j^0\bigr) \le \alpha.$$

Definition 14.4 P-values adjusted for multiple comparison.

P-values adjusted for multiple comparison *for a multiple testing problem with the elementary null hypotheses* H_j: $\theta_j = \theta_j^0$, $j = 1, \ldots, m$, *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*

$$p_j^{adj} = \inf\Bigl\{\alpha:\ \bigl(\theta_j^L(\alpha), \theta_j^U(\alpha)\bigr) \not\ni \theta_j^0\Bigr\}, \qquad j = 1, \ldots, m.$$

[4] *procedura vícenásobného srovnávání*

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

$$\min\bigl\{p_1^{adj},\, \ldots,\, p_m^{adj}\bigr\} \leq \alpha.$$

That is,

$$p^{adj} := \min\bigl\{p_1^{adj},\, \ldots,\, p_m^{adj}\bigr\}$$

is the P-value of a test of the global null hypothesis based on the considered MCP.

Example 14.3 (Bonferroni multiple comparison procedure, Bonferroni adjusted P-values).

Let for $0 < \alpha < 1$, $\bigl(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\bigr)$, $j = 1, \ldots, m$, be the confidence intervals for parameters $\theta_1,\, \ldots,\, \theta_m$, each with a (univariate) coverage of $1 - \frac{\alpha}{m}$. That is,

$$\forall\, j = 1, \ldots, m,\ \forall\, \theta_j^0 \in \Theta_j\,:\quad \mathsf{P}\Bigl(\bigl(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\bigr) \ni \theta_j^0;\, \theta_j = \theta_j^0\Bigr) \geq 1 - \frac{\alpha}{m}.$$

As shown in Example 14.2, $\bigl(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\bigr)$, $j = 1, \ldots, m$, are the Bonferroni simultaneous confidence intervals for parameter $\boldsymbol{\theta} = \bigl(\theta_1,\, \ldots,\, \theta_m\bigr)^\top$ 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 $\bigl(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\bigr)$. That is,

$$p_j^{uni} = \inf\Bigl\{\frac{\alpha}{m}\,:\ \bigl(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\bigr) \not\ni \theta_j^0\Bigr\}.$$

Hence,

$$\min\bigl\{m\, p_j^{uni},\, 1\bigr\} = \inf\Bigl\{\alpha\,:\ \bigl(\theta_j^L(\alpha),\, \theta_j^U(\alpha)\bigr) \not\ni \theta_j^0\Bigr\}.$$

That is, the P-values adjusted for multiple comparison based on the Bonferroni simultaneous confidence intervals are

$$p_j^B = \min\bigl\{m\, p_j^{uni},\, 1\bigr\}, \qquad j = 1, \ldots, m.$$

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.

14.2.4 Bonferroni simultaneous inference in a normal linear model

Consider a linear model

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

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 a vector of linear combinations of regression coefficients. 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 $m > k$ and also for possibly linearly dependent vectors $\mathbf{l}_1,\,\ldots,\,\mathbf{l}_m$.

As usual, let $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}} = \big(\mathbf{l}_1^\top\widehat{\boldsymbol{\beta}},\,\ldots,\,\mathbf{l}_m^\top\widehat{\boldsymbol{\beta}}\big)^\top = \big(\widehat{\theta}_1,\,\ldots,\,\widehat{\theta}_m\big)^\top$ 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

$$\begin{aligned}
\theta_j^L(\alpha) &= \mathbf{l}_j^\top\widehat{\boldsymbol{\beta}} - \sqrt{\mathsf{MS}_e\,\mathbf{l}_j^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}_j}\ \ \mathsf{t}_{n-k}\Big(1 - \frac{\alpha}{2\,m}\Big),\\
\theta_j^U(\alpha) &= \mathbf{l}_j^\top\widehat{\boldsymbol{\beta}} + \sqrt{\mathsf{MS}_e\,\mathbf{l}_j^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}_j}\ \ \mathsf{t}_{n-k}\Big(1 - \frac{\alpha}{2\,m}\Big), \qquad j = 1, \ldots, m.
\end{aligned} \tag{14.5}$$

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 (14.5) is dual to the (single) test of the (jth elementary) hypothesis $\mathrm{H}_j:\ \theta_j = \theta_j^0$ based on the statistic

$$T_j(\theta_j^0) = \frac{\mathbf{l}_j^\top\widehat{\boldsymbol{\beta}} - \theta_j^0}{\sqrt{\mathsf{MS}_e\,\mathbf{l}_j^\top\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbf{l}_j}},$$

(which under the hypothesis H_j follows the Student t_{n-k} distribution) while having the critical region of the test on a level of α/m as

$$C_j = \bigg(-\infty,\ -\mathsf{t}_{n-k}\Big(1 - \frac{\alpha}{2\,m}\Big)\bigg] \bigcup \bigg[\mathsf{t}_{n-k}\Big(1 - \frac{\alpha}{2\,m}\Big),\ \infty\bigg).$$

The related univariate P-values are then calculated as

$$p_j^{uni} = 2\,\mathsf{CDF}_{t,\,n-k}\big(-\,|t_{j,0}|\big),$$

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 $\mathrm{H}_j:\ \theta_j = \theta_j^0$, $j = 1, \ldots, m$, are

$$p_j^B = \min\Big\{2\,m\,\mathsf{CDF}_{t,\,n-k}\big(-\,|t_{j,0}|\big),\ 1\Big\}, \qquad j = 1, \ldots, m.$$

14.3 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).

14.3.1 Tukey's pairwise comparisons theorem

Lemma 14.1 Studentized range.

Let $T_1, \dots, T_m$ be a random sample from $\mathcal{N}(\mu,\, \sigma^2)$, $\sigma^2 > 0$. Let

$$R = \max_{j=1,\dots,m} T_j \;-\; \min_{j=1,\dots,m} T_j$$

be the range of the sample. Let S^2 be the estimator of σ^2 such that S^2 and $\boldsymbol{T} = \bigl(T_1, \dots, T_m\bigr)^\top$ 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:

$$\frac{R}{S} = \frac{\dfrac{1}{\sigma}\Bigl\{\max\limits_j (T_j - \mu) - \min\limits_j (T_j - \mu)\Bigr\}}{\dfrac{S}{\sigma}} = \frac{\max\limits_j \Bigl(\dfrac{T_j - \mu}{\sigma}\Bigr) - \min\limits_j \Bigl(\dfrac{T_j - \mu}{\sigma}\Bigr)}{\dfrac{S}{\sigma}}.$$

- Distribution of both the numerator and the denominator depends on neither μ, nor σ since
  - For all $j = 1, \dots, m$ $\quad \dfrac{T_j - \mu}{\sigma} \sim \mathcal{N}(0,\, 1)$.
  - Distribution of $\dfrac{S}{\sigma}$ is a transformation of the χ^2_ν distribution.
- At the same time, numerator and denominator are independent and hence also their joint distribution depends on neither μ, nor σ. Consequently, also distribution of their ratio depends on neither μ, nor σ. ❑

Note. The distribution of the random variable $Q = \frac{R}{S}$ from Lemma 14.1 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 14.5 Studentized range.

The random variable $Q = \dfrac{R}{S}$ *from Lemma 14.1 will be called* studentized range[5] *of a sample of size* m *with* ν *degrees of freedom and its distribution will be denoted as* $\mathsf{q}_{m,\nu}$.

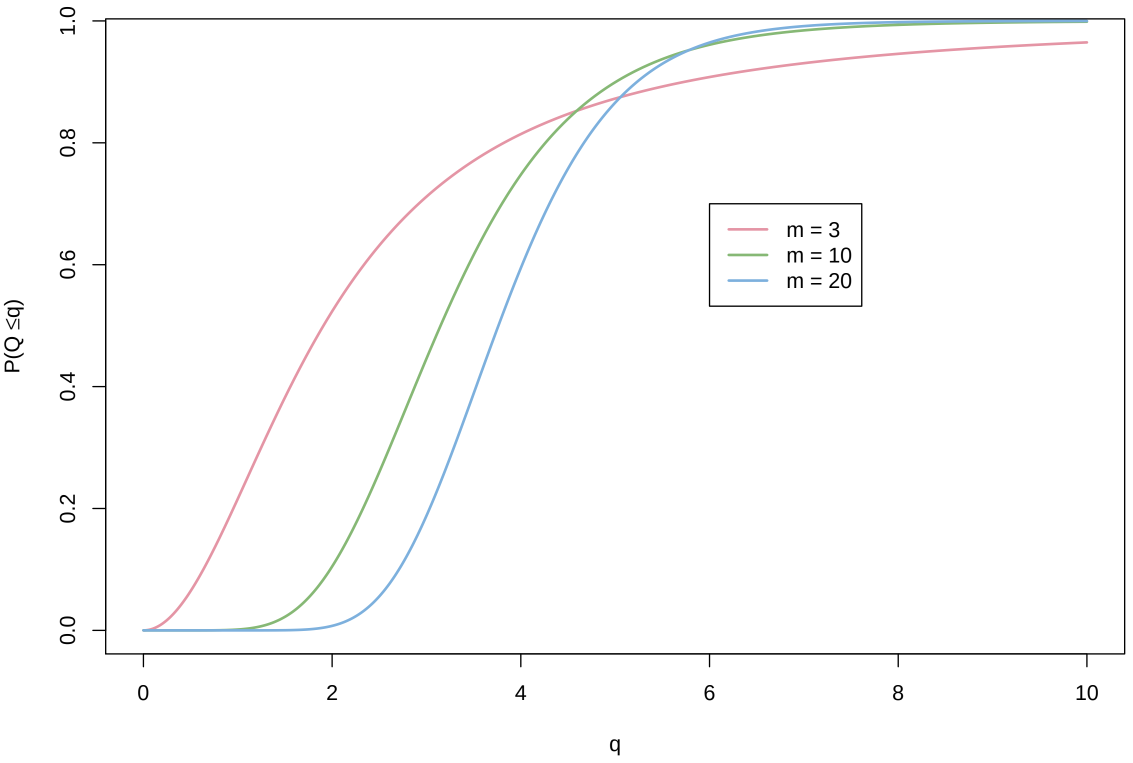
Notation.

- 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 $\mathsf{CDF}_{\mathsf{q},m,\nu}(\cdot)$.

Illustrations

Studentized range: distribution functions

For $m = 3, 10, 20$ **and** $\nu = m - 1$**, R:** `ptukey(q, m, nu)`



[5] *studentizované rozpětí*

Illustrations

Studentized range: selected quantiles

For $m = 3, 10, 20$ and $\nu = m - 1$, R: `qtukey(p, m, nu)`

```
p <- c(0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.975)
quants <- data.frame(p = p,
                     q3  = round(qtukey(p, 3, 2), 4),
                     q10 = round(qtukey(p, 10, 9), 4),
                     q20 = round(qtukey(p, 20, 19), 4))
colnames(quants) <- c("p", paste("m = ", c(3, 10, 20), sep = ""))
print(quants)
```

```
      p     m = 3   m = 10   m = 20
1 0.025    0.3050   1.5291   2.2698
2 0.050    0.4370   1.7270   2.4650
3 0.100    0.6351   1.9800   2.7087
4 0.250    1.1007   2.4726   3.1664
5 0.500    1.9082   3.1494   3.7626
6 0.750    3.3080   4.0107   4.4724
7 0.900    5.7326   5.0067   5.2315
8 0.950    8.3308   5.7384   5.7518
9 0.975   11.9365   6.4790   6.2498
```

Theorem 14.2 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 $\boldsymbol{T} = \bigl(T_1, \ldots, T_m\bigr)^\top$ are* independent *and*

$$\frac{\nu\, S^2}{\sigma^2} \sim \chi^2_\nu \qquad \textit{for some} \qquad \nu > 0.$$

Then

$$\mathsf{P}\Bigl(\textit{for all } j \neq l\textit{:}\quad \bigl|T_j - T_l\, -\, (\mu_j - \mu_l)\bigr|\; <\; \mathsf{q}_{m,\nu}(1-\alpha)\,\sqrt{v^2\ S^2}\Bigr) \;=\; 1 - \alpha.$$

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 $R = \max\limits_j\left(\dfrac{T_j - \mu_j}{v}\right) - \min\limits_j\left(\dfrac{T_j - \mu_j}{v}\right)$.
$$\Rightarrow\ \frac{R}{S} \sim q_{m,\,\nu}.$$
- Hence for any $0 < \alpha < 1$ ($\mathsf{q}_{m,\nu}$ is a continuous distribution):
$$1 - \alpha = \mathsf{P}\left(\frac{\max\limits_j\left(\dfrac{T_j - \mu_j}{v}\right) - \min\limits_j\left(\dfrac{T_j - \mu_j}{v}\right)}{S} < q_{m,\nu}(1-\alpha)\right)$$

$$
\begin{aligned}
&= \mathsf{P}\left(\frac{\max\limits_j(T_j-\mu_j) \;-\; \min\limits_j(T_j-\mu_j)}{v\,S} \;<\; q_{m,\nu}(1-\alpha)\right)\\
&= \mathsf{P}\Big(\max_j(T_j-\mu_j) \;-\; \min_j(T_j-\mu_j) \;<\; v\,S\,q_{m,\nu}(1-\alpha)\Big)\\
&= \mathsf{P}\Big(\text{for all } j\neq l \quad \big|(T_j-\mu_j) \;-\; (T_l-\mu_l)\big| \;<\; v\,S\,q_{m,\nu}(1-\alpha)\Big)\\
&= \mathsf{P}\Big(\text{for all } j\neq l \quad \big|T_j-T_l \;-\; (\mu_j-\mu_l)\big| \;<\; q_{m,\nu}(1-\alpha)\sqrt{v^2\,S^2}\Big).
\end{aligned}
$$

❑

Theorem 14.3 Tukey's pairwise comparisons theorem, general version.

Let $T_1,\dots,T_m$ *be* independent *random variables and let* $T_j \sim \mathcal{N}(\mu_j,\, v_j^2\sigma^2)$, $j=1,\dots,m$, *where* $v_j^2>0$, $j=1,\dots,m$ *are* known *constants. Let* S^2 *be the estimator of* σ^2 *such that* S^2 *and* $\boldsymbol{T}=\big(T_1,\dots,T_m\big)^\top$ *are* independent *and*

$$
\frac{\nu\,S^2}{\sigma^2} \sim \chi^2_\nu \qquad \textit{for some} \qquad \nu>0.
$$

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 14.3 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 14.3 was called as Tukey–Kramer conjecture.
- The proof is not an easy adaptation of the proof of the balanced version.

14.3.2 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

$$\boldsymbol{T} = \bigl(T_1,\,\ldots,\,T_m\bigr)^\top \sim \mathcal{N}_m(\boldsymbol{\mu},\,\sigma^2\,\mathbb{V}),$$

where

- $\boldsymbol{\mu} = \bigl(\mu_1,\,\ldots,\,\mu_m\bigr)^\top \in \mathbb{R}^m$ 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

$$\theta_{j,l} = \mu_j - \mu_l, \qquad j = 1,\,\ldots,\,m-1,\ l = j+1,\,\ldots,\,m,$$

$$\boldsymbol{\theta} = \bigl(\theta_{1,2},\,\theta_{1,3},\,\ldots,\,\theta_{m-1,m}\bigr)^\top.$$

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 $\boldsymbol{\theta}^0 = \bigl(\theta^0_{1,2},\,\theta^0_{1,3},\,\ldots,\,\theta^0_{m-1,m}\bigr)^\top \in \mathbb{R}^{m^\star}$. 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 14.2 and 14.3), we have (for chosen $0 < \alpha < 1$):

$$\mathsf{P}\Biggl(\text{for all } j \neq l \quad \bigl|T_j - T_l\ -\ (\mu_j - \mu_l)\bigr|\ <\ \mathsf{q}_{m,\nu}(1-\alpha)\,\sqrt{\frac{v_j^2 + v_l^2}{2}\,S^2}\Biggr) \quad \geq \quad 1-\alpha,$$

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,

$$\mathsf{P}\left(\text{for all } j \neq l \quad \left|\frac{T_j - T_l - (\mu_j - \mu_l)}{\sqrt{\frac{v_j^2+v_l^2}{2}\,S^2}}\right| < \mathsf{q}_{m,\nu}(1-\alpha)\right) \quad \geq \quad 1-\alpha.$$

Let for $j \neq l$ and for $\theta^0_{j,l} \in \mathbb{R}$

$$T_{j,l}(\theta^0_{j,l}) := \frac{T_j - T_l - \theta^0_{j,l}}{\sqrt{\dfrac{v_j^2 + v_l^2}{2}\; S^2}}.$$

That is

$$1 - \alpha \;\leq\; \mathsf{P}\Bigl(\text{for all } j \neq l \quad \bigl|T_{j,l}(\theta^0_{j,l})\bigr| < \mathsf{q}_{m,\nu}(1-\alpha);\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \tag{14.6}$$

$$= \mathsf{P}\left(\text{for all } j \neq l \quad \left|\frac{T_j - T_l - \theta^0_{j,l}}{\sqrt{\frac{v_j^2+v_l^2}{2}\; S^2}}\right| < \mathsf{q}_{m,\nu}(1-\alpha);\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\right)$$

$$= \mathsf{P}\Bigl(\text{for all } j \neq l \quad \bigl(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\bigr) \ni \theta^0_{j,l};\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr), \tag{14.7}$$

where

$$\begin{aligned} \theta^{TL}_{j,l}(\alpha) &= T_j - T_l - \mathsf{q}_{m,\nu}(1-\alpha)\sqrt{\tfrac{v_j^2+v_l^2}{2}\; S^2}, \\ \theta^{TU}_{j,l}(\alpha) &= T_j - T_l + \mathsf{q}_{m,\nu}(1-\alpha)\sqrt{\tfrac{v_j^2+v_l^2}{2}\; S^2}, \qquad j < l. \end{aligned} \tag{14.8}$$

Theorem 14.4 Tukey's honest significance differences.

Random intervals given by (14.8) 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 $\boldsymbol{\theta}^0 \in \mathbb{R}^{m^\star}$

$$\mathsf{P}\Bigl(\textit{for all } j \neq l \quad \bigl(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\bigr) \ni \theta^0_{j,l};\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \;=\; 1 - \alpha.$$

Related P-values for a multiple testing problem with elementary hypotheses $H_{j,l}:\; \theta_{j,l} = \theta^0_{j,l}$, $\theta^0_{j,l} \in \mathbb{R}$, $j < l$, adjusted for multiple comparison are given by

$$p^T_{j,l} = 1 \;-\; \mathsf{CDF}_{\mathsf{q},m,\nu}\Bigl(\bigl|t^0_{j,l}\bigr|\Bigr), \qquad j < l,$$

where $t^0_{j,l}$ is a value of $T_{j,l}(\theta^0_{j,l}) = \frac{T_j - T_l - \theta^0_{j,l}}{\sqrt{\frac{v_j^2+v_l^2}{2}\; S^2}}$ attained with given data.

Proof.

The fact that $\bigl(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\bigr)$, $j < l$, are simultaneous confidence intervals for parameters $\theta_{j,l} = \mu_j - \mu_l$ with a coverage of $1 - \alpha$ follows from (14.7).

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 (14.6) 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 $\mathrm{H}_{j,l}:\; \theta_{j,l} = \theta^0_{j,l}$, $j < l$, follows from noting the following (for each $j < l$):

$$\bigl(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\bigr) \not\ni \theta^0_{j,l} \quad\Longleftrightarrow\quad \Bigl|T_{j,l}\bigl(\theta^0_{j,l}\bigr)\Bigr| \geq \mathsf{q}_{m,\,\nu}(1-\alpha)$$

It now follows from monotonicity of the quantiles of a continuous Studentized range distribution that

$$p^T_{j,l} = \inf\Bigl\{\alpha :\; \bigl(\theta^{TL}_{j,l}(\alpha),\; \theta^{TU}_{j,l}(\alpha)\bigr) \not\ni \theta^0_{j,l}\Bigr\} = \inf\Bigl\{\alpha :\; \Bigl|T_{j,l}\bigl(\theta^0_{j,l}\bigr)\Bigr| \geq \mathsf{q}_{m,\,\nu}(1-\alpha)\Bigr\}$$

is attained for $p^T_{j,l}$ satisfying

$$\Big|T_{j,l}\big(\theta^0_{j,l}\big)\Big| = \mathsf{q}_{m,\,\nu}\big(1 - p^T_{j,l}\big).$$

That is, if $t^0_{j,l}$ is a value of the statistic $T_{j,l}\big(\theta^0_{j,l}\big)$ attained with given data, we have

$$p^T_{j,l} = 1 \;-\; \mathsf{CDF}_{\mathsf{q},m,\nu}\Big(\big|t^0_{j,l}\big|\Big).$$

❑

14.3.3 Tukey's HSD in a linear model

In context of a normal linear model $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \mathcal{N}_n\big(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\big)$, $\mathsf{rank}\big(\mathbb{X}_{n\times k}\big) = k < n$, the Tukey's honest significance differences are applicable in the following situation.

- $\mathbb{L}_{m\times k}$ is a matrix with non-zero rows $\mathbf{l}_1^\top, \ldots, \mathbf{l}_m^\top$,

$$\boldsymbol{\eta} := \mathbb{L}\boldsymbol{\beta} = \big(\mathbf{l}_1^\top\boldsymbol{\beta},\, \ldots,\, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \big(\eta_1,\, \ldots,\, \eta_m\big)^\top.$$

- Matrix $\mathbb{L}$ is such that

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

  is a diagonal matrix with $v_j^2 := v_{j,j}$, $j = 1, \ldots, m$.

With $\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}$ and the residual mean square MS_e of the fitted linear model, we have (conditionally, given the model matrix $\mathbb{X}$):

$$\boldsymbol{T} := \widehat{\boldsymbol{\eta}} = \big(\mathbf{l}_1^\top\widehat{\boldsymbol{\beta}},\, \ldots,\, \mathbf{l}_m^\top\widehat{\boldsymbol{\beta}}\big)^\top = \mathbb{L}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_m\big(\boldsymbol{\eta},\, \sigma^2\mathbb{V}\big), \qquad \frac{(n-k)\mathsf{MS}_e}{\sigma^2} \sim \chi^2_{n-k},$$

$$\widehat{\boldsymbol{\eta}} \text{ and } \mathsf{MS}_e \text{ independent.}$$

Hence the Tukey's T-procedure can be used for a multiple comparison problem on (also estimable) parameters

$$\theta_{j,l} = \eta_j - \eta_l = \big(\mathbf{l}_j - \mathbf{l}_l\big)^\top\boldsymbol{\beta}, \qquad j < l.$$

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

$$\begin{aligned} \theta^{TL}_{j,l}(\alpha) &= \widehat{\eta}_j - \widehat{\eta}_l - \mathsf{q}_{m,n-k}(1-\alpha)\,\sqrt{\tfrac{v_j^2+v_l^2}{2}\;\mathsf{MS}_e}, \\ \theta^{TU}_{j,l}(\alpha) &= \widehat{\eta}_j - \widehat{\eta}_l + \mathsf{q}_{m,n-k}(1-\alpha)\,\sqrt{\tfrac{v_j^2+v_l^2}{2}\;\mathsf{MS}_e}, \qquad j < l. \end{aligned}$$

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with elementary hypotheses

$$\mathrm{H}_{j,l}\colon\; \theta_{j,l} = \theta^0_{j,l}, \qquad j < l,$$

for chosen $\theta^0_{j,l} \in \mathbb{R}$, is based on statistics

$$T_{j,l}(\theta^0_{j,l}) = \frac{\widehat{\eta}_j - \widehat{\eta}_l \;-\; \theta^0_{j,l}}{\sqrt{\dfrac{v_j^2 + v_l^2}{2}\;\mathsf{MS}_e}}, \qquad j < l.$$

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 $\boldsymbol{Y} = \bigl(Y_{1,1}, \, \dots, \, Y_{G,n_G}\bigr)^\top$, $n = \sum_{g=1}^{G} n_g$, and

$$\begin{aligned} &Y_{g,j} \;\sim\; \mathcal{N}(m_g, \, \sigma^2), \\ &Y_{g,j} \text{ independent for } g = 1, \dots, G, \; j = 1, \dots, n_g, \end{aligned}$$

We then have (see Lemma 13.1, with random covariates conditionally given the covariate values)

$$\boldsymbol{T} := \begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_G \end{pmatrix} \sim \mathcal{N}_G \left(\begin{pmatrix} m_1 \\ \vdots \\ m_G \end{pmatrix}, \quad \sigma^2 \begin{pmatrix} \frac{1}{n_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{n_G} \end{pmatrix} \right).$$

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 $\boldsymbol{m} = \bigl(m_1, \, \dots, \, m_G\bigr)^\top$). Hence the Tukey's simultaneous confidence intervals for $\theta_{g,h} = m_g - m_h$, $g = 1, \dots, G-1$, $h = g+1, \dots, G$ 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 ($n_1 = \dots = n_G$), 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 $\boldsymbol{Y} = \bigl(Y_{1,1,1}, \, \dots, \, Y_{G,H,n_{G,H}}\bigr)^\top$, $n = \sum_{g=1}^{G} \sum_{h=1}^{H} n_{g,h}$, and

$$\begin{aligned} &Y_{g,h,j} \;\sim\; \mathcal{N}(m_{g,h}, \, \sigma^2), \\ &Y_{g,h,j} \text{ independent for } g = 1, \dots, G, \; h = 1, \dots, H, \; j = 1, \dots, n_{g,h}, \end{aligned}$$

Let, as usual,

$$n_{g\bullet} = \sum_{h=1}^{H} n_{g,h}, \qquad \overline{Y}_{g\bullet} = \frac{1}{n_{g\bullet}} \sum_{h=1}^{H} \sum_{j=1}^{n_{g,h}} Y_{g,h,j},$$

$$\overline{m}_{g\bullet} = \frac{1}{H} \sum_{h=1}^{H} m_{g,h}, \qquad \overline{m}^{wt}_{g\bullet} = \frac{1}{n_{g\bullet}} \sum_{h=1}^{H} n_{g,h} m_{g,h}, \qquad g = 1, \dots, G.$$

Balanced data

In case of *balanced data* ($n_{g,h} = J$ for all g, h), we have $n_{g\bullet} = J\,H$, $\overline{m}_g^{wt} = \overline{m}_g$. Further,

$$\boldsymbol{T} := \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{G\bullet} \end{pmatrix} \sim \mathcal{N}_G\left(\begin{pmatrix} \overline{m}_{1\bullet} \\ \vdots \\ \overline{m}_{G\bullet} \end{pmatrix},\; \sigma^2 \begin{pmatrix} \frac{1}{J\,H} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{J\,H} \end{pmatrix} \right),$$

see Consequence of Lemma 13.2. 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 Lemma 13.2 that for both the interaction model and the additive model, the sample means $\overline{Y}_{1\bullet}, \dots, \overline{Y}_{G\bullet}$ are LSE's of estimable parameters $\overline{m}_{1\bullet}, \dots, \overline{m}_{G\bullet}$ 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 $\theta_{g_1,g_2} = \overline{m}_{g_1\bullet} - \overline{m}_{g_2\bullet}$, $g_1 = 1, \dots, G-1$, $g_2 = g_1 + 1, \dots, G$ 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

$$\mathrm{H}_{g_1,g_2}:\; \theta_{g_1,g_2} = \theta^0_{g_1,g_2}, \qquad g_1 < g_2,$$

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 $\overline{m}^{wt}_{1\bullet}, \dots, \overline{m}^{wt}_{G\bullet}$ 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 $\widehat{Y}_{g,h,j} = \overline{Y}_{g,h\bullet}$, $g = 1, \dots, G$, $h = 1, \dots, H$, $j = 1, \dots, n_{g,h}$ (Lemma 13.2). Hence the sample means $\overline{Y}_{1\bullet}, \dots, \overline{Y}_{G\bullet}$, 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 $\overline{m}^{wt}_{1\bullet}, \dots, \overline{m}^{wt}_{G\bullet}$. To show that the sample means $\overline{Y}_{1\bullet}, \dots, \overline{Y}_{G\bullet}$ are the LSE's for the estimable parameters $\overline{m}^{wt}_{1\bullet}, \dots, \overline{m}^{wt}_{G\bullet}$ 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 $\mathsf{MS}_e^\star$ and $\nu_e^\star$ 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 $\mathsf{MS}_e^\star$. 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

$$\overline{Y}_{g_1\bullet} - \overline{Y}_{g_2\bullet} \;\pm\; \mathsf{q}_{G,\nu_e^\star}(1-\alpha)\sqrt{\frac{1}{2}\Bigl(\frac{1}{n_{g_1\bullet}} + \frac{1}{n_{g_2\bullet}}\Bigr)\mathsf{MS}_e^\star}.$$

Calculation of the P-values adjusted for multiple comparison related to the multiple testing problem with elementary hypotheses

$$\mathrm{H}_{g_1,g_2}:\ \theta^{wt}_{g_1,g_2} = \theta^{wt,0}_{g_1,g_2}, \qquad g_1 < g_2,$$

for chosen $\theta^{wt,0}_{g_1,g_2} \in \mathbb{R}$, is based on statistics

$$T_{g_1,g_2}(\theta^{wt,0}_{g_1,g_2}) = \frac{\overline{Y}_{g_1\bullet} - \overline{Y}_{g_2\bullet} - \theta^{wt,0}_{g_1,g_2}}{\sqrt{\frac{1}{2}\Bigl(\frac{1}{n_{g_1\bullet}} + \frac{1}{n_{g_2\bullet}}\Bigr)\mathsf{MS}_e^\star}}, \qquad g_1 < g_2.$$

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

14.4 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.

14.4.1 Max-abs-t distribution

Definition 14.6 Max-abs-t-distribution.

Let $\boldsymbol{T} = \bigl(T_1, \dots, T_m\bigr)^\top \sim \mathsf{mvt}_{m,\nu}\bigl(\boldsymbol{\Sigma}\bigr)$*, where* $\boldsymbol{\Sigma}$ *is a positive* semidefinite *matrix. The distribution of a random variable*

$$H = \max_{j=1,\dots,m} |T_j|$$

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

$$\mathsf{P}\Bigl(\max_{j=1,\dots,m} |T_j| \leq \mathsf{h}_{m,\nu}(p;\,\boldsymbol{\Sigma})\Bigr) = p.$$

- 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 $\boldsymbol{T} \sim \mathsf{mvt}_{m,\nu}\bigl(\boldsymbol{\Sigma}\bigr)$ has a density w.r.t. Lebesgue measure

$$f_T(\boldsymbol{t}) = \frac{\Gamma\bigl(\frac{\nu+m}{2}\bigr)}{\Gamma\bigl(\frac{\nu}{2}\bigr)\,\nu^{\frac{m}{2}}\,\pi^{\frac{m}{2}}}\,\bigl|\boldsymbol{\Sigma}\bigr|^{-\frac{1}{2}}\left\{1 + \frac{\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}}{\nu}\right\}^{-\frac{\nu+m}{2}}, \qquad \boldsymbol{t} \in \mathbb{R}^m.$$

- The distribution function $\mathsf{CDF}_{\mathsf{h},m,\nu}(\cdot;\,\boldsymbol{\Sigma})$ of a random variable $H = \max_{j=1,\dots,m} |T_j|$ is then (for $h > 0$):

$$\mathsf{CDF}_{\mathsf{h},m,\nu}(h;\,\boldsymbol{\Sigma}) = \mathsf{P}\Bigl(\max_{j=1,\dots,m} |T_j| \leq h\Bigr) = \mathsf{P}\Bigl(\forall j = 1,\dots,m \quad |T_j| \leq h\Bigr)$$

$$= \int_{-h}^{h} \cdots \int_{-h}^{h} f_T(t_1,\dots,t_m)\,\mathsf{d}t_1 \cdots \mathsf{d}t_m.$$

- 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 $\boldsymbol{\Sigma}$.

14.4.2 General multiple comparison procedure for a linear model

Assumptions.

In the following, we consider a normal linear model

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

Further, let

$$\mathbb{L}_{m\times k} = \begin{pmatrix} \mathbf{l}_1^\top \\ \vdots \\ \mathbf{l}_m^\top \end{pmatrix}$$

be a matrix and

$$\boldsymbol{\theta} := \mathbb{L}\boldsymbol{\beta} = \big(\mathbf{l}_1^\top\boldsymbol{\beta},\, \dots,\, \mathbf{l}_m^\top\boldsymbol{\beta}\big)^\top = \big(\theta_1,\, \dots,\, \theta_m\big)^\top,$$

$\mathbf{l}_1 \neq \mathbf{0}_k,\, \dots,\, \mathbf{l}_m \neq \mathbf{0}_k$.

Notes.

- The number m of the estimable parameters of interest may be arbitrary, i.e., even greater than 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 $\theta_1,\, \dots,\, \theta_m$ with the multiple testing problem composed of m elementary hypotheses

$$\mathrm{H}_j: \ \ \theta_j = \theta_j^0, \qquad j = 1,\, \dots,\, m,$$

for some $\boldsymbol{\theta}^0 = \big(\theta_1^0,\, \dots,\, \theta_m^0\big)^\top \in \mathbb{R}^m$. The global null hypothesis is as usual $\mathrm{H}_0: \ \boldsymbol{\theta} = \boldsymbol{\theta}^0$.

Notation. In the following, the following (standard) notation will be used:

- $\widehat{\boldsymbol{\beta}} = \big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{X}^\top\boldsymbol{Y}$;
- $\widehat{\boldsymbol{\theta}} = \mathbb{L}\widehat{\boldsymbol{\beta}} = \big(\mathbf{l}_1^\top\widehat{\boldsymbol{\beta}},\, \dots,\, \mathbf{l}_m^\top\widehat{\boldsymbol{\beta}}\big)^\top = \big(\widehat{\theta}_1,\, \dots,\, \widehat{\theta}_m\big)^\top$: LSE of $\boldsymbol{\theta}$;
- $\mathbb{V} = \mathbb{L}\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\mathbb{L}^\top = \big(v_{j,l}\big)_{j,l=1,\dots,m}$;
- $\mathbb{D} = \mathsf{diag}\Bigg(\dfrac{1}{\sqrt{v_{1,1}}},\, \dots,\, \dfrac{1}{\sqrt{v_{m,m}}}\Bigg)$;
- MS_e: the residual mean square of the model with $\nu_e = n - k$ degrees of freedom.

Reminders from Chapter 6

- For $j = 1, \dots, m$, (both conditionally given $\mathbb{X}$ and unconditionally as well):

$$Z_j := \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\sigma^2\, v_{j,j}}} \ \sim \ \mathcal{N}(0,\, 1), \qquad\qquad T_j := \frac{\widehat{\theta}_j - \theta_j}{\sqrt{\mathsf{MS}_e\, v_{j,j}}} \ \sim \ \mathsf{t}_{n-k}.$$

- Further (conditionally given $\mathbb{X}$):

$$\boldsymbol{Z} = \big(Z_1,\, \dots,\, Z_m\big)^\top = \frac{1}{\sqrt{\sigma^2}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \ \sim \ \mathcal{N}_m\big(\mathbf{0}_m,\, \mathbb{D}\mathbb{V}\mathbb{D}\big),$$

$$\boldsymbol{T} = \big(T_1,\, \dots,\, T_m\big)^\top = \frac{1}{\sqrt{\mathsf{MS}_e}}\,\mathbb{D}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \ \sim \ \mathsf{mvt}_{m,\,n-k}\big(\mathbb{D}\mathbb{V}\mathbb{D}\big).$$

Notes.

- Matrices $\mathbb{V}$ and $\mathbb{DVD}$ are not necessarily invertible.
- If $\mathsf{rank}\big(\mathbb{L}\big) = m \leq k$ then both matrices $\mathbb{V}$ and $\mathbb{DVD}$ are invertible and Theorem 6.2 further provides (both conditionally given $\mathbb{X}$ and unconditionally as well) that under $\mathrm{H}_0: \;\; \boldsymbol{\theta} = \boldsymbol{\theta}^0$:
$$Q_0 = \frac{1}{m}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big)^\top \big(\mathsf{MS}_e\,\mathbb{V}\big)^{-1}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big) = \frac{1}{m}\,\boldsymbol{T}^\top\,\big(\mathbb{DVD}\big)^{-1}\boldsymbol{T} \;\;\sim\;\; \mathcal{F}_{m,\,n-k}.$$
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 $\mathrm{H}_0: \; \boldsymbol{\theta} = \boldsymbol{\theta}^0$:
$$Q_0 = \frac{1}{m}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big)^\top \big(\mathsf{MS}_e\,\mathbb{V}\big)^{+}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big) = \frac{1}{m}\,\boldsymbol{T}^\top\,\big(\mathbb{DVD}\big)^{+}\boldsymbol{T} \;\;\sim\;\; \mathcal{F}_{m_0,\,n-k}$$
(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 $\mathrm{H}_0: \; \boldsymbol{\theta} = \boldsymbol{\theta}^0$:

$$\boldsymbol{T}\big(\boldsymbol{\theta}^0\big) := \big(T_1(\theta_1^0),\,\ldots,\,T_m(\theta_m^0)\big)^\top \;\sim\; \mathsf{mvt}_{m,\,n-k}(\mathbb{DVD}).$$

We then have, for $0 < \alpha < 1$:

$$\begin{aligned}
1 - \alpha \;&=\; \mathsf{P}\Big(\max_{j=1,\ldots,m}\big|T_j(\theta_j^0)\big| < \mathsf{h}_{m,\,n-k}(1-\alpha;\,\mathbb{DVD});\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big)\\
&=\; \mathsf{P}\Big(\text{for all } j = 1, \ldots, m \quad \big|T_j(\theta_j^0)\big| < \mathsf{h}_{m,\,n-k}(1-\alpha;\,\mathbb{DVD});\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big)\\
&=\; \mathsf{P}\left(\text{for all } j = 1, \ldots, m \quad \left|\frac{\widehat{\theta}_j - \theta_j^0}{\sqrt{\mathsf{MS}_e\,v_{j,j}}}\right| < \mathsf{h}_{m,\,n-k}(1-\alpha;\,\mathbb{DVD});\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\right)\\
&=\; \mathsf{P}\Big(\text{for all } j = 1, \ldots, m \quad \Big(\theta_j^{HL}(\alpha),\; \theta_j^{HU}(\alpha)\Big) \ni \theta_j^0;\; \boldsymbol{\theta} = \boldsymbol{\theta}^0\Big),
\end{aligned} \tag{14.9}$$

where

$$\begin{aligned}
\theta_j^{HL}(\alpha) \;&= \widehat{\theta}_j - \mathsf{h}_{m,\,n-k}(1-\alpha;\,\mathbb{DVD})\,\sqrt{\mathsf{MS}_e\,v_{j,j}},\\
\theta_j^{HU}(\alpha) \;&= \widehat{\theta}_j + \mathsf{h}_{m,\,n-k}(1-\alpha;\,\mathbb{DVD})\,\sqrt{\mathsf{MS}_e\,v_{j,j}}, \qquad j = 1,\,\ldots,\,m.
\end{aligned} \tag{14.10}$$

Theorem 14.5 Hothorn-Bretz-Westfall MCP for linear hypotheses in a normal linear model.

Random intervals given by (14.10) are simultaneous confidence intervals for parameters $\theta_j = \mathbf{l}_j^\top \boldsymbol{\beta}$, $j = 1, \ldots, m$, with an exact coverage of $1-\alpha$, i.e., for any $\boldsymbol{\theta}^0 = \bigl(\theta_1^0, \ldots, \theta_m^0\bigr)^\top \in \mathbb{R}^m$

$$\mathsf{P}\Bigl(\textit{for all } j = 1, \ldots, m \quad \bigl(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\bigr) \ni \theta_j^0;\ \boldsymbol{\theta} = \boldsymbol{\theta}^0\Bigr) \;=\; 1-\alpha.$$

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

$$p_j^H = 1 \;-\; \mathsf{CDF}_{\mathsf{h},m,n-k}\Bigl(\bigl|t_j^0\bigr|;\ \mathbb{DVD}\Bigr), \qquad j = 1, \ldots, m,$$

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 $\Bigl(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Bigr)$, $j = 1, \ldots, m$, are simultaneous confidence intervals for parameters $\theta_j = \mathbf{l}_j^\top \boldsymbol{\beta}$ with an *exact* coverage of $1-\alpha$ follows from (14.9).

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 $j = 1, \ldots, m$):

$$\Bigl(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Bigr) \not\ni \theta_j^0 \quad\Longleftrightarrow\quad \Bigl|T_j\bigl(\theta_j^0\bigr)\Bigr| \geq \mathsf{h}_{m,\,n-k}(1-\alpha;\ \mathbb{DVD}).$$

It now follows from monotonicity of the quantiles of a continuous max-abs-t-distribution that

$$p_j^H = \inf\Bigl\{\alpha:\ \Bigl(\theta_j^{HL}(\alpha),\ \theta_j^{HU}(\alpha)\Bigr) \not\ni \theta_j^0\Bigr\} = \inf\Bigl\{\alpha:\ \Bigl|T_j\bigl(\theta_j^0\bigr)\Bigr| \geq \mathsf{h}_{m,\,n-k}(1-\alpha;\ \mathbb{DVD})\Bigr\}$$

is attained for p_j^H satisfying

$$\Bigl|T_j\bigl(\theta_j^0\bigr)\Bigr| = \mathsf{h}_{m,\,n-k}(1-p_j^H;\ \mathbb{DVD}).$$

That is, if t_j^0 is a value of the statistic $T_j\bigl(\theta_j^0\bigr)$ attained with given data, we have

$$p_j^H = 1 \;-\; \mathsf{CDF}_{\mathsf{h},m,n-k}\Bigl(\bigl|t_j^0\bigr|;\ \mathbb{DVD}\Bigr).$$

❑

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 $\mathsf{h}_{m,\,n-k}(1-\alpha)$, 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 $\mathsf{h}_{m,\,n-k}(1-\alpha)$ 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.

14.5 Confidence band for the regression function

In this section, we shall assume that data are represented by i.i.d. random vectors $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $i = 1, \ldots, n$, being sampled from a distribution of a generic random vector $\bigl(Y,\, \boldsymbol{Z}^\top\bigr)^\top \in \mathbb{R}^{1+p}$. 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

$$\boldsymbol{Y} \,\big|\, \mathbb{Z} \;\sim\; \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\bigr) \tag{14.11}$$

for some $\boldsymbol{\beta} \in \mathbb{R}^k$, $\sigma^2 > 0$. Remember that it follows from (14.11) that

$$Y_i \;\big|\; \boldsymbol{Z}_i \;\sim\; \mathcal{N}\bigl(\boldsymbol{X}_i^\top\boldsymbol{\beta},\, \sigma^2\bigr),$$

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

$$\mathbb{E}\bigl(Y \,\big|\, \boldsymbol{X} = \boldsymbol{t}(\boldsymbol{z})\bigr) = \mathbb{E}\bigl(Y \,\big|\, \boldsymbol{Z} = \boldsymbol{z}\bigr) = m(\boldsymbol{z}) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}, \qquad \boldsymbol{z} \in \mathbb{R}^p.$$

It will further be assumed that the model matrix $\mathbb{X}$ is of full-rank (almost surely), i.e., $\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k$. 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 6.3

Let $\boldsymbol{z} \in \mathbb{R}^p$ be given. Theorem 6.3 then states that a random interval with the lower and upper bounds given as

$$\boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \,\pm\, \mathrm{t}_{n-k}\Bigl(1 - \frac{\alpha}{2}\Bigr)\sqrt{\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{t}(\boldsymbol{z})},$$

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 $\boldsymbol{\beta}^0 \in \mathbb{R}^k$, $\sigma_0^2 > 0$,

$$\mathsf{P}\Bigl(\boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \,\pm\, \mathrm{t}_{n-k}\Bigl(1 - \frac{\alpha}{2}\Bigr)\sqrt{\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{t}(\boldsymbol{z})} \;\ni\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0;\; \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\Bigr) \;=\; 1 - \alpha.$$

Theorem 14.6 Confidence band for the regression function.

Let $\bigl(Y_i,\, \boldsymbol{Z}_i^\top\bigr)^\top$, $i = 1, \ldots, n$, *be i.i.d. random vectors such that* $\boldsymbol{Y} \,|\, \mathbb{Z} \;\sim\; \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\bigr)$, *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* $\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k$. *Finally, let for all* $\boldsymbol{z} \in \mathbb{R}^p$ $\boldsymbol{t}(\boldsymbol{z}) \neq \mathbf{0}_k$. *Then for any* $\boldsymbol{\beta}^0 \in \mathbb{R}^k$, $\sigma_0^2 > 0$,

$$\begin{aligned}&\mathsf{P}\Bigl(\textit{for all } \boldsymbol{z} \in \mathbb{R}^p \\ &\quad \boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \,\pm\, \sqrt{k\,\mathcal{F}_{k,\,n-k}(1-\alpha)\,\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{t}(\boldsymbol{z})} \;\ni\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0; \quad \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\Bigr) \\ &\hspace{30em} = \; 1 - \alpha.\end{aligned}$$

Note. Requirement $\boldsymbol{t}(\boldsymbol{z}) \neq \boldsymbol{0}_k$ 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. **Proof/calculations were skipped and are not requested for the exam.**

Let (for $0 < \alpha < 1$)

$$\mathcal{K} = \Bigl\{\boldsymbol{\beta} \in \mathbb{R}^k : \ \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathbb{X}^\top\mathbb{X}\bigr) \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr) \ \leq \ k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)\Bigr\}.$$

Section 6.2: $\mathcal{K}$ is a confidence *ellipsoid* for $\boldsymbol{\beta}$ with a coverage of $1-\alpha$, that is, for any $\boldsymbol{\beta}^0 \in \mathbb{R}^k$, $\sigma_0^2 > 0$,

$$\mathsf{P}\bigl(\mathcal{K} \ni \boldsymbol{\beta}^0;\ \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\bigr) = 1 - \alpha.$$

$\mathcal{K}$ 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 $\boldsymbol{\beta}^0 \in \mathbb{R}^k$, $\sigma_0^2 > 0$,

$$1 - \alpha = \mathsf{P}\bigl(\mathcal{K} \ni \boldsymbol{\beta}^0;\ \boldsymbol{\beta} = \boldsymbol{\beta}^0\bigr) = \mathsf{P}\bigl(\text{for all } \boldsymbol{z} \in \mathbb{R}^p \quad L(\boldsymbol{z}) \ \leq\ \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0 \ \leq\ U(\boldsymbol{z});\ \boldsymbol{\beta} = \boldsymbol{\beta}^0,\, \sigma^2 = \sigma_0^2\bigr). \tag{14.12}$$

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 $\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathbb{X}^\top\mathbb{X}\bigr) \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr) \ = \ k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)$.

Method of Lagrange multipliers:

$$\varphi(\boldsymbol{\beta},\, \lambda) = \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta} + \frac{1}{2}\,\lambda\Bigl\{\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathbb{X}^\top\mathbb{X}\bigr) \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr) \ -\ k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)\Bigr\}$$

($\frac{1}{2}$ is only included to simplify subsequent expressions).

Derivatives of φ:

$$\frac{\partial\varphi}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta},\, \lambda) = \boldsymbol{t}(\boldsymbol{z}) \ +\ \lambda\,\mathbb{X}^\top\mathbb{X}\,\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr),$$

$$\frac{\partial\varphi}{\partial\lambda}(\boldsymbol{\beta},\, \lambda) = \frac{1}{2}\Bigl\{\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr)^\top \bigl(\mathbb{X}^\top\mathbb{X}\bigr) \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\bigr) \ -\ k\,\mathsf{MS}_e\,\mathcal{F}_{k,n-k}(1-\alpha)\Bigr\}.$$

With given λ, the first set of equations is solved (with respect to $\boldsymbol{\beta}$) for

$$\boldsymbol{\beta}(\lambda) = \widehat{\boldsymbol{\beta}} \ -\ \frac{1}{\lambda}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\boldsymbol{t}(\boldsymbol{z}).$$

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 (14.12) and realizing that, due to continuity,

$$\begin{aligned}
1-\alpha &= \mathsf{P}\big(\text{for all } \boldsymbol{z}\in\mathbb{R}^p \quad L(\boldsymbol{z}) \;\le\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0 \;\le\; U(\boldsymbol{z});\; \boldsymbol{\beta}=\boldsymbol{\beta}^0\big)\\
&= \mathsf{P}\big(\text{for all } \boldsymbol{z}\in\mathbb{R}^p \quad L(\boldsymbol{z}) \;<\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0 \;<\; U(\boldsymbol{z});\; \boldsymbol{\beta}=\boldsymbol{\beta}^0\big)\\
&= \mathsf{P}\Big(\text{for all } \boldsymbol{z}\in\mathbb{R}^p\\
&\qquad \boldsymbol{t}^\top(\boldsymbol{z})\widehat{\boldsymbol{\beta}} \;\pm\; \sqrt{k\,\mathcal{F}_{k,\,n-k}(1-\alpha)\,\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})\big(\mathbb{X}^\top\mathbb{X}\big)^{-1}\boldsymbol{t}(\boldsymbol{z})} \;\ni\; \boldsymbol{t}^\top(\boldsymbol{z})\boldsymbol{\beta}^0;\; \boldsymbol{\beta}=\boldsymbol{\beta}^0,\,\sigma^2=\sigma_0^2\Big).
\end{aligned}$$

❑

Terminology *(Confidence band for the regression function).*

If the covariates $Z_1, \dots, Z_n \in \mathbb{R}$, confidence intervals according to Theorem (14.6) are often calculated for an (equidistant) sequence of values $z_1, \dots, z_N \in \mathbb{R}$ 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

$$\mathsf{t}_{n-k}\Bigl(1-\frac{\alpha}{2}\Bigr)\ \sqrt{\mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})}$$

$$= \sqrt{\mathcal{F}_{1,n-k}(1-\alpha)\ \ \mathsf{MS}_e\,\boldsymbol{t}^\top(\boldsymbol{z})(\mathbb{X}^\top\mathbb{X})^{-1}\boldsymbol{t}(\boldsymbol{z})},$$

since for any $\nu > 0$, $\mathsf{t}_\nu^2\Bigl(1-\frac{\alpha}{2}\Bigr) = \mathcal{F}_{1,\nu}(1-\alpha)$.

For $k \geq 2$, and any $\nu > 0$,

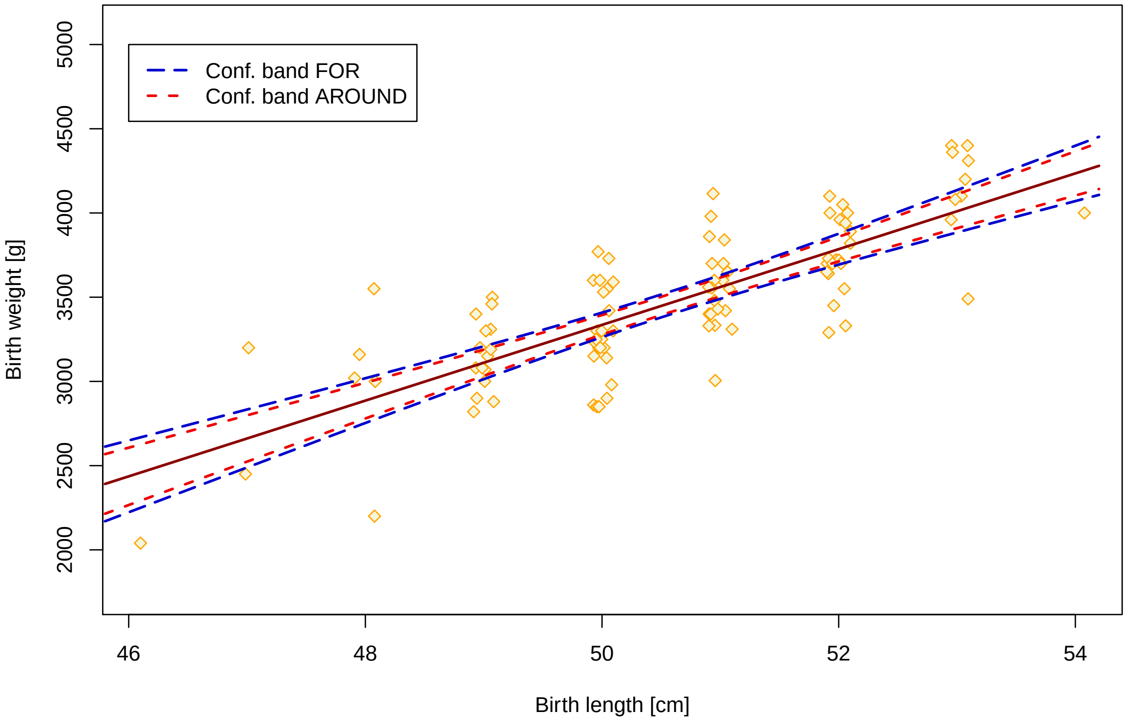
$$k\,\mathcal{F}_{k,\nu}(1-\alpha) > \mathcal{F}_{1,\nu}(1-\alpha)$$

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 $k = 1$.

Illustrations

`Kojeni` **(**$n = 99$**)**

`bweight` $\sim$ `blength`



[6] *pás spolehlivosti pro regresní funkci*

Chapter 15

General Linear Model

We still assume that data are represented by a set of n random vectors $\bigl(Y_i,\,\boldsymbol{X}_i^\top\bigr)^\top$, $\boldsymbol{X}_i = \bigl(X_{i,0},\,\dots,\,X_{i,k-1}\bigr)^\top$, $i = 1,\dots,n$, and use symbols $\boldsymbol{Y}$ for a vector $\bigl(Y_1,\,\dots,\,Y_n\bigr)^\top$ and $\mathbb{X}$ for an $n \times k$ matrix with rows given by the covariate/regressor vectors $\boldsymbol{X}_1,\,\dots,\,\boldsymbol{X}_n$. 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 15.1 General linear model.

The data $\bigl(\boldsymbol{Y},\,\mathbb{X}\bigr)$ satisfy a general linear model[1] *if*

$$\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr) = \mathbb{X}\boldsymbol{\beta}, \qquad \mathsf{var}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr) = \sigma^2\,\mathbb{W}^{-1},$$

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

$$\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\mathbb{W}^{-1}\bigr).$$

- 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 15.1 (Regression based on sample means).

Suppose that data are represented by random vectors $\bigl(\widetilde{Y}_{1,1},\dots,\widetilde{Y}_{1,w_1},\,\boldsymbol{X}_1^\top\bigr)^\top,$
$\dots,$
$\bigl(\widetilde{Y}_{n,1},\dots,\widetilde{Y}_{n,w_n},\,\boldsymbol{X}_n^\top\bigr)^\top$
such that for each $i = 1,\dots,n$, the random variables $\widetilde{Y}_{i,1},\dots,\widetilde{Y}_{i,w_i}$ are uncorrelated *with a common conditional (given $\boldsymbol{X}_i$) variance σ^2.*

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 $Y_1,\,\dots,\,Y_n$, where

$$Y_1 = \frac{1}{w_1}\sum_{j=1}^{w_1}\widetilde{Y}_{1,j}, \quad \dots, \quad Y_n = \frac{1}{w_n}\sum_{j=1}^{w_n}\widetilde{Y}_{n,j}.$$

[1] *obecný lineární model* [2] *zobecněný lineární model*

The covariance matrix (conditional given $\mathbb{X}$) of a random vector $\boldsymbol{Y} = \bigl(Y_1, \dots, Y_n\bigr)^\top$ is then

$$\mathsf{var}\bigl(\boldsymbol{Y} \,\big|\, \mathbb{X}\bigr) = \sigma^2 \underbrace{\begin{pmatrix} \frac{1}{w_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{w_n} \end{pmatrix}}_{\mathbb{W}^{-1}}.$$

Theorem 15.1 Generalized least squares.

Assume a general linear model $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2 \mathbb{W}^{-1}\bigr)$, where $\mathsf{rank}\bigl(\mathbb{X}_{n\times k}\bigr) = k < n$. The following then holds:

(i) A vector

$$\widehat{\boldsymbol{Y}}_G := \mathbb{X}\bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top \mathbb{W}\boldsymbol{Y}$$

is the best linear unbiased estimator (BLUE) of a vector parameter $\boldsymbol{\mu} := \mathbb{E}\bigl(\boldsymbol{Y} \,\big|\, \mathbb{X}\bigr) = \mathbb{X}\boldsymbol{\beta}$, and

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

If further $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\bigr)$ then

$$\widehat{\boldsymbol{Y}}_G \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\, \mathbb{X}\bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\bigr).$$

(ii) Let $\mathbf{l} \in \mathbb{R}^k$, $\mathbf{l} \neq \mathbf{0}_k$ and let

$$\widehat{\boldsymbol{\beta}}_G := \bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top \mathbb{W}\boldsymbol{Y}.$$

Then $\widehat{\theta}_G = \mathbf{l}^\top \widehat{\boldsymbol{\beta}}_G$ is the best linear unbiased estimator (BLUE) of θ with

$$\mathsf{var}\bigl(\widehat{\theta}_G \,\big|\, \mathbb{X}\bigr) = \sigma^2\, \mathbf{l}^\top \bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\mathbf{l}.$$

If further $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\bigr)$ then

$$\widehat{\theta}_G \,\big|\, \mathbb{X} \sim \mathcal{N}\bigl(\theta,\, \sigma^2\, \mathbf{l}^\top \bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\mathbf{l}\bigr).$$

(iii) The vector

$$\widehat{\boldsymbol{\beta}}_G := \bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top \mathbb{W}\boldsymbol{Y}$$

is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ with

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

If additionally $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\bigr)$ then

$$\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X} \sim \mathcal{N}_k\bigl(\boldsymbol{\beta},\, \sigma^2\, \bigl(\mathbb{X}^\top \mathbb{W}\mathbb{X}\bigr)^{-1}\bigr).$$

(iv) The statistic

$$\mathsf{MS}_{e,G} := \frac{\mathsf{SS}_{e,G}}{n-k},$$

where

$$\mathsf{SS}_{e,G} := \Bigl\|\mathbb{W}^{\frac{1}{2}}\,\bigl(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\bigr)\Bigr\|^2 = \bigl(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\bigr)^\top \mathbb{W}\bigl(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\bigr),$$

is the unbiased estimator of the residual variance σ^2.
If additionally $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\bigr)$ then

$$\frac{\mathsf{SS}_{e,G}}{\sigma^2} \sim \chi^2_{n-k},$$

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

$$\mathbb{W} = \mathbb{W}^{\frac{1}{2}}\bigl(\mathbb{W}^{\frac{1}{2}}\bigr)^{\top}, \quad \text{e.g., Cholesky decomposition}$$
$$\mathbb{W}^{-1} = \mathbb{W}^{-\frac{1}{2}}\bigl(\mathbb{W}^{-\frac{1}{2}}\bigr)^{\top}.$$

(i) Let $\boldsymbol{Y}^{\star} = \mathbb{W}^{\frac{1}{2}}\boldsymbol{Y}$.

Then $\mathbb{E}\bigl(\boldsymbol{Y}^{\star}\,\big|\,\mathbb{X}\bigr) = \mathbb{W}^{\frac{1}{2}}\mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr) = \mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta}$,

$$\mathsf{var}\bigl(\boldsymbol{Y}^{\star}\,\big|\,\mathbb{X}\bigr) = \mathbb{W}^{\frac{1}{2}}\underbrace{\mathsf{var}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr)}_{\sigma^2\mathbb{W}^{-1}}\bigl(\mathbb{W}^{\frac{1}{2}}\bigr)^{\top} = \sigma^2\mathbf{I}_n.$$

That is, we have a linear model $\mathsf{M}^{\star}$

$$\mathsf{M}^{\star}:\ \boldsymbol{Y}^{\star}\,\big|\,\mathbb{X} \ \sim\ \bigl(\underbrace{\mathbb{W}^{\frac{1}{2}}\mathbb{X}}_{\mathbb{X}^{\star}}\boldsymbol{\beta},\ \sigma^2\,\mathbf{I}_n\bigr),$$

where $\mathsf{rank}\bigl(\mathbb{X}^{\star}\bigr) = \mathsf{rank}\bigl(\mathbb{W}^{\frac{1}{2}}\mathbb{X}\bigr) = \mathsf{rank}\bigl(\mathbb{X}\bigr) = k$.

The hat matrix for model $\mathsf{M}^{\star}$ is

$$\mathbb{H}^{\star} \ =\ \mathbb{W}^{\frac{1}{2}}\mathbb{X}\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\bigl(\mathbb{W}^{\frac{1}{2}}\bigr)^{\top},$$

where the matrix $\mathbb{X}^{\top}\mathbb{W}\mathbb{X}$ is, given our assumptions, invertible.

The fitted values in model $\mathsf{M}^{\star}$ are then calculated as

$$\widehat{\boldsymbol{Y}}^{\star} \ =\ \mathbb{H}^{\star}\boldsymbol{Y}^{\star} \ =\ \mathbb{W}^{\frac{1}{2}}\mathbb{X}\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\mathbb{W}\boldsymbol{Y}.$$

By Gauss-Markov theorem (Theorem 2.4), the vector $\widehat{\boldsymbol{Y}}^{\star}$ is the best linear unbiased estimator (BLUE) of the vector $\mathbb{E}\bigl(\boldsymbol{Y}^{\star}\,\big|\,\mathbb{X}\bigr) \ =\ \mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta}$ with

$$\mathsf{var}\bigl(\widehat{\boldsymbol{Y}}^{\star}\,\big|\,\mathbb{X}\bigr) \ =\ \sigma^2\,\mathbb{H}^{\star} \ =\ \sigma^2\,\mathbb{W}^{\frac{1}{2}}\mathbb{X}\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\bigl(\mathbb{W}^{\frac{1}{2}}\bigr)^{\top}.$$

By linearity, the vector

$$\widehat{\boldsymbol{Y}}_G \ :=\ \mathbb{W}^{-\frac{1}{2}}\widehat{\boldsymbol{Y}}^{\star} \ =\ \mathbb{X}\,\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\mathbb{W}\boldsymbol{Y}$$

is the BLUE of the vector

$$\mathbb{W}^{-\frac{1}{2}}\mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta} \ =\ \mathbb{X}\boldsymbol{\beta} \ =\ \mathbb{E}\bigl(\boldsymbol{Y}\,\big|\,\mathbb{X}\bigr),$$

and

$$\mathsf{var}\bigl(\widehat{\boldsymbol{Y}}_G\,\big|\,\mathbb{X}\bigr) \ =\ \mathbb{W}^{-\frac{1}{2}}\,\mathsf{var}\bigl(\widehat{\boldsymbol{Y}}^{\star}\,\big|\,\mathbb{X}\bigr)\,\bigl(\mathbb{W}^{-\frac{1}{2}}\bigr)^{\top} \ =\ \sigma^2\,\mathbb{X}\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}.$$

If additionally $\boldsymbol{Y}\,\big|\,\mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\,\sigma^2\mathbb{W}^{-1}\bigr)$, 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

$$\widehat{\boldsymbol{Y}}^{\star}\,\big|\,\mathbb{X} \ \sim\ \mathcal{N}\Bigl(\mathbb{W}^{\frac{1}{2}}\mathbb{X}\boldsymbol{\beta},\ \sigma^2\,\mathbb{W}^{\frac{1}{2}}\mathbb{X}\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\bigl(\mathbb{W}^{\frac{1}{2}}\bigr)^{\top}\Bigr),$$
$$\widehat{\boldsymbol{Y}}_G\,\big|\,\mathbb{X} \ \sim\ \mathcal{N}\Bigl(\mathbb{X}\boldsymbol{\beta}\qquad,\ \sigma^2\,\mathbb{X}\bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\Bigr).$$

(ii) By relationship between the least squares and normal equations, we have that

$$\begin{aligned}
\widehat{\boldsymbol{Y}}^{\star} \ =\ \mathbb{X}^{\star}\widehat{\boldsymbol{\beta}}^{\star} \quad &\Longleftrightarrow\quad \widehat{\boldsymbol{\beta}}^{\star} \text{ solves normal equations in model } \mathsf{M}^{\star}\\
&\Longleftrightarrow\quad \widehat{\boldsymbol{\beta}}^{\star} \text{ solves } \mathbb{X}^{\star\top}\mathbb{X}^{\star}\boldsymbol{b} \ =\ \mathbb{X}^{\star\top}\boldsymbol{Y}^{\star}\\
&\Longleftrightarrow\quad \widehat{\boldsymbol{\beta}}^{\star} \text{ solves } \mathbb{X}^{\top}\mathbb{W}\mathbb{X}\boldsymbol{b} \ =\ \mathbb{X}^{\top}\mathbb{W}\boldsymbol{Y}\\
&\Longleftrightarrow\quad \widehat{\boldsymbol{\beta}}^{\star} \ =\ \bigl(\mathbb{X}^{\top}\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^{\top}\mathbb{W}\boldsymbol{Y}.
\end{aligned}$$

Remember that $\mathbb{X}^\star = \mathbb{W}^{\frac{1}{2}}\mathbb{X}$. Hence,

$$\widehat{\boldsymbol{Y}}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\mathbb{X}\widehat{\boldsymbol{\beta}}^\star \quad \text{if and only if } \widehat{\boldsymbol{\beta}}^\star \;=\; \bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

Further, remember that $\widehat{\boldsymbol{Y}}_G \;=\; \mathbb{W}^{-\frac{1}{2}}\widehat{\boldsymbol{Y}}^\star$. Hence,

$$\widehat{\boldsymbol{Y}}_G \;=\; \mathbb{W}^{-\frac{1}{2}}\mathbb{W}^{\frac{1}{2}}\mathbb{X}\widehat{\boldsymbol{\beta}}^\star \quad \text{if and only if } \widehat{\boldsymbol{\beta}}^\star \;=\; \bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

That is,

$$\widehat{\boldsymbol{Y}}_G \;=\; \mathbb{X}\widehat{\boldsymbol{\beta}}_G \quad \text{if and only if } \widehat{\boldsymbol{\beta}}_G := \widehat{\boldsymbol{\beta}}^\star \;=\; \bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}.$$

Then, by Gauss-Markov theorem (Theorem 2.5),

$$\widehat{\theta}_G \;:=\; \widehat{\theta}^\star \;=\; \mathbf{l}^\top\widehat{\boldsymbol{\beta}}_G$$

is BLUE of the parameter $\theta \;=\; \mathbf{l}^\top\boldsymbol{\beta}$. Furthermore,

$$\mathsf{var}\bigl(\widehat{\theta}_G \,\big|\, \mathbb{X}\bigr) \;=\; \mathsf{var}\bigl(\widehat{\theta}^\star \,\big|\, \mathbb{X}\bigr) \;=\; \sigma^2\,\mathbf{l}^\top\bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbf{l}.$$

If additionally, $\boldsymbol{Y} \,\big|\, \mathbb{X} \sim \mathcal{N}_n\bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\mathbb{W}^{-1}\bigr)$ then by properties of a normal distribution (only linear transformations are involved to calculate $\widehat{\theta}_G$ from $\boldsymbol{Y}$), we have

$$\widehat{\theta}_G \,\big|\, \mathbb{X} \;\sim\; \mathcal{N}\bigl(\theta,\, \sigma^2\,\mathbf{l}^\top\bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbf{l}\bigr).$$

(iii) Suppose that for an $m \times k$ matrix, the parameter $\boldsymbol{\theta} = \mathbb{L}\boldsymbol{\beta}$ is a vector parameter being given as non-trivial linear combinations of regression coefficients of the model $\mathsf{M}\colon\; \boldsymbol{Y} \,\big|\, \mathbb{X} \;\sim\; \bigl(\mathbb{X}\boldsymbol{\beta},\, \sigma^2\,\mathbb{W}^{-1}\bigr)$. By analogous steps as in (ii), we show that

$$\widehat{\boldsymbol{\theta}}_G \;:=\; \mathbb{L}\widehat{\boldsymbol{\beta}}_G, \quad \widehat{\boldsymbol{\beta}}_G \;=\; \bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$$

is BLUE of $\boldsymbol{\theta}$. Furthermore,

$$\mathsf{var}\bigl(\widehat{\boldsymbol{\theta}}_G \,\big|\, \mathbb{X}\bigr) \;=\; \sigma^2\,\mathbb{L}\bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{L}^\top,$$

and under assumption of normality,

$$\widehat{\boldsymbol{\theta}}_G \,\big|\, \mathbb{X} \;\sim\; \mathcal{N}_m\bigl(\boldsymbol{\theta},\, \sigma^2\,\mathbb{L}\bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{L}^\top\bigr).$$

Now, if $\mathsf{rank}\bigl(\mathbb{X}\bigr) = k$, the matrix $\mathbb{X}^\top\mathbb{W}\mathbb{X}$ is invertible. Moreover, by taking $\mathbb{L} = \mathbf{I}_k$ we obtain that the BLUE of the vector $\boldsymbol{\beta}$ is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_G \;&:=\; \bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y},\\ \mathsf{var}\bigl(\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X}\bigr) \;&=\; \sigma^2\,\bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1},\end{aligned}$$

and under assumption of normality,

$$\widehat{\boldsymbol{\beta}}_G \,\big|\, \mathbb{X} \;\sim\; \mathcal{N}_k\bigl(\boldsymbol{\beta},\, \sigma^2\,\bigl(\mathbb{X}^\top\mathbb{W}\mathbb{X}\bigr)^{-1}\bigr).$$

(iv) Let us first calculate the residual sum of squares of the model $\mathsf{M}^\star:\; \boldsymbol{Y}^\star \,\big|\, \mathbb{X} \;\sim\; \bigl(\mathbb{X}^\star\boldsymbol{\beta},\; \sigma^2\,\mathbf{I}_n\bigr)$, where $\boldsymbol{Y}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\boldsymbol{Y}$, $\mathbb{X}^\star \;=\; \mathbb{W}^{\frac{1}{2}}\mathbb{X}$, $\mathsf{rank}\bigl(\mathbb{X}^\star\bigr) \;=\; \mathsf{rank}\bigl(\mathbb{X}\bigr) \;=\; k$. We have (remember further that $\widehat{\boldsymbol{Y}}^\star = \mathbb{W}^{\frac{1}{2}}\widehat{\boldsymbol{Y}}_G$)

$$\begin{aligned}\mathsf{SS}_e^\star \;&=\; \bigl(\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\bigr)^\top\bigl(\boldsymbol{Y}^\star - \widehat{\boldsymbol{Y}}^\star\bigr) \;=\; \bigl(\mathbb{W}^{\frac{1}{2}}\boldsymbol{Y} - \mathbb{W}^{\frac{1}{2}}\widehat{\boldsymbol{Y}}_G\bigr)^\top\bigl(\mathbb{W}^{\frac{1}{2}}\boldsymbol{Y} - \mathbb{W}^{\frac{1}{2}}\widehat{\boldsymbol{Y}}_G\bigr)\\ &=\; \bigl(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\bigr)^\top\mathbb{W}\bigl(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_G\bigr) \;=:\; \mathsf{SS}_{e,G}.\end{aligned}$$

By Lemma 2.7, we have

$$\mathbb{E}\bigl(\mathsf{SS}_{e,G}\bigr) \;=\; \mathbb{E}\bigl(\mathsf{SS}_e^\star\bigr) \;=\; (n-k)\,\sigma^2 \;=\; \mathbb{E}\bigl(\mathsf{SS}_e^\star \,\big|\, \mathbb{X}\bigr) \;=\; \mathbb{E}\bigl(\mathsf{SS}_{e,G} \,\big|\, \mathbb{X}\bigr).$$

That is,

$$\mathsf{MS}_{e,G} := \frac{\mathsf{SS}_{e,G}}{n-k}$$

is the unbiased estimator of the residual variance σ^2.

Furthermore, if normality is assumed, Theorem 6.2 applied to model $\mathsf{M}^\star$ provides that

$$\frac{\mathsf{SS}_e^\star}{\sigma^2} \sim \chi^2_{n-k}.$$

Since $\mathsf{SS}_e^\star = \mathsf{SS}_{e,G}$, we have directly

$$\frac{\mathsf{SS}_{e,G}}{\sigma^2} \sim \chi^2_{n-k}.$$

Finally, Theorem 6.2 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. 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 $\widehat{\boldsymbol{Y}}_G = \mathbb{X}\left(\mathbb{X}^\top\mathbb{W}\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{W}\boldsymbol{Y}$ 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 $\mathsf{MS}_{e,G} = \dfrac{\mathsf{SS}_{e,G}}{n-k}$ 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.

Note. The most common use of the generalized least squares is the situation described in Example 15.1, 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

$$\mathbb{X}^\top\mathbb{W}\boldsymbol{Y} = \sum_{i=1}^n w_i Y_i \boldsymbol{X}_i, \qquad \mathbb{X}^\top\mathbb{W}\mathbb{X} = \sum_{i=1}^n w_i \boldsymbol{X}_i \boldsymbol{X}_i^\top,$$

$$\mathsf{SS}_{e,G} = \sum_{i=1}^n w_i\left(Y_i - \widehat{Y}_{G,i}\right)^2.$$

The method of the generalized least squares is then usually referred to as the method of *the weighted least squares (WLS).*[7]

[3] *zobecněné vyrovnané hodnoty* [4] *zobecněný reziduální součet čtverců* [5] *zobecněný střední čtverec* [6] *odhad metodou zobecněných nejmenších čtverců* [7] *vážené nejmenší čtverce*

Illustrations

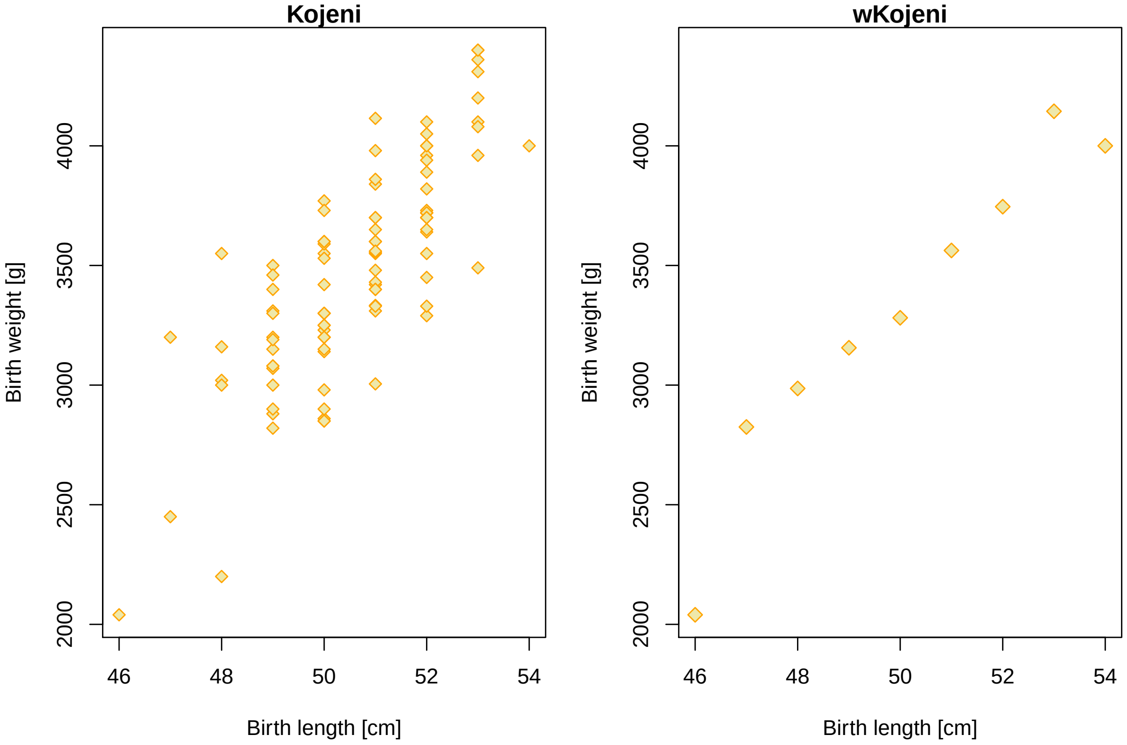
Kojeni

- Data on $n = 99$ newborn children.
- Y: birth weight (`bweight`).
- X: birth length (`blength`)
  - Only (nine) discrete values 46, 47, $\ldots$, 54 [cm] appear in data due to rounding.

wKojeni

- $n = 9$.
- Y: average birth weight of all children from data `Kojeni` with the same birth length.

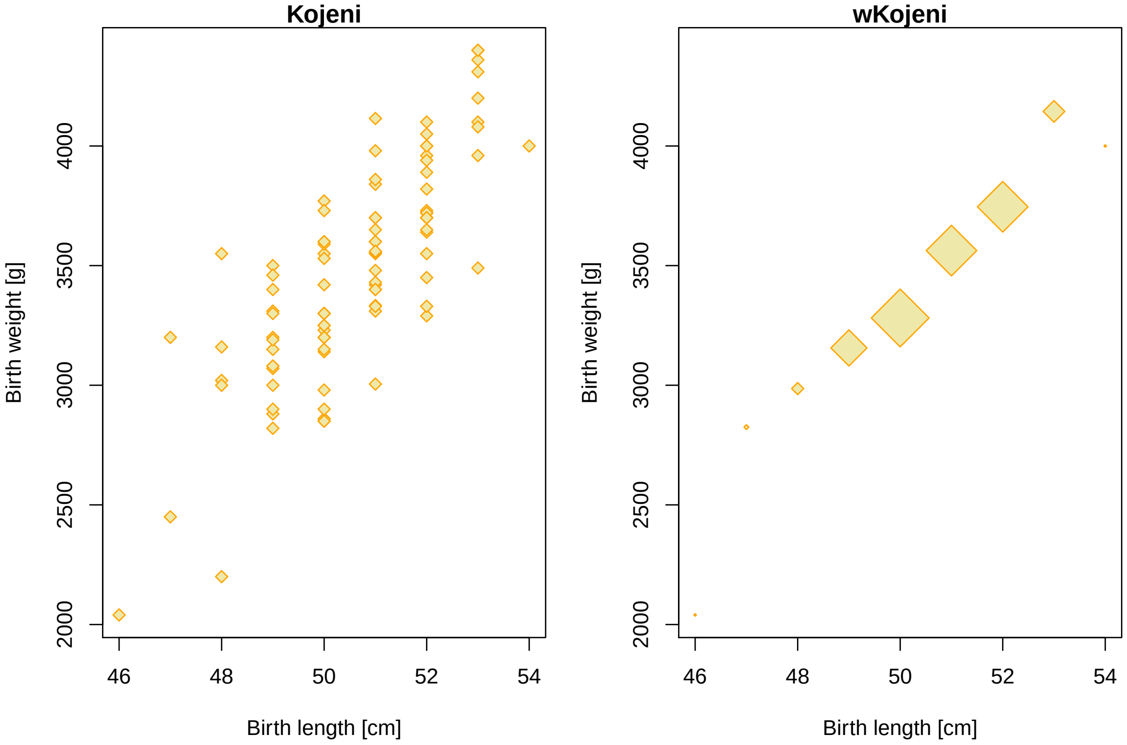
Data `Kojeni` and `wKojeni`



Kojeni

wKojeni

Data `Kojeni` and `wKojeni`



Kojeni

wKojeni

Illustrations

Data Kojeni

bweight ∼ blength

Ordinary least squares using complete data Kojeni

```
m1 <- lm(bweight ~ blength, data = Kojeni)
summary(m1)
confint(m1)
```

```
### summary(m1):
Call:
lm(formula = bweight ~ blength, data = Kojeni)

Residuals:
    Min      1Q  Median      3Q     Max
-685.93 -152.83  -30.76  196.83  664.07

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7905.80     895.45  -8.829 4.52e-14 ***
blength       224.83      17.69  12.709  < 2e-16 ***
 ---

Residual standard error: 271.7 on 97 degrees of freedom
Multiple R-squared:  0.6248,        Adjusted R-squared:  0.6209
F-statistic: 161.5 on 1 and 97 DF,  p-value: < 2.2e-16

### confint(m1):
                 2.5 %     97.5 %                      2.5 %     97.5 %
(Intercept) -9683.0226 -6128.5847          blength  189.7184   259.9372
```

Data wKojeni

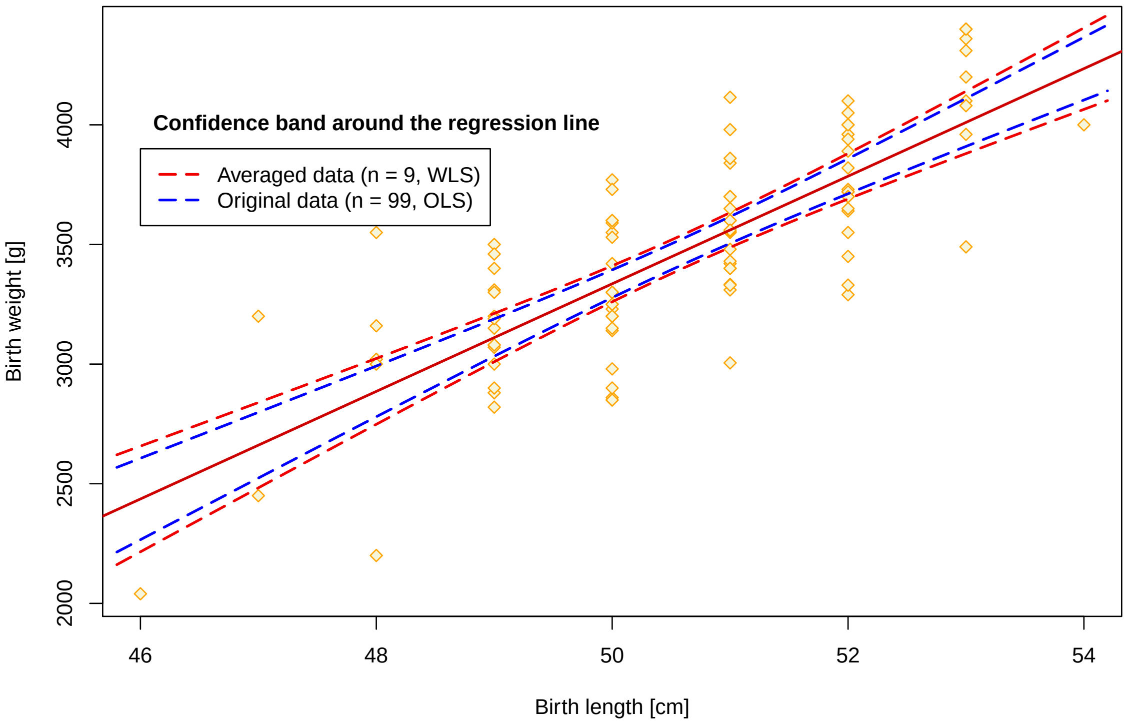
bweight ∼ blength

Weighted least squares using averaged data wKojeni

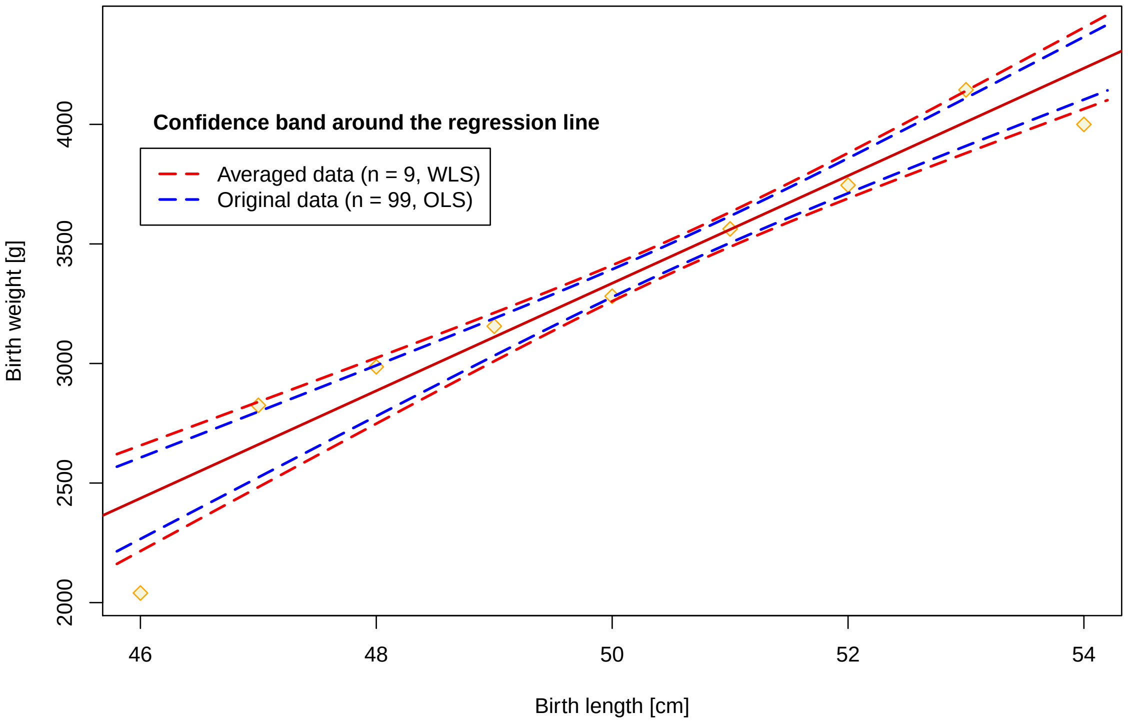
```
wm1 <- lm(bweight ~ blength, weights = w, data = wKojeni)
summary(wm1)
confint(wm1)
```

```
### summary(wm1):
Call:
lm(formula = bweight ~ blength, data = wKojeni, weights = w)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-396.28 -234.90   10.75  223.76  403.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7905.80     975.42  -8.105 8.39e-05 ***
blength       224.83      19.27  11.667 7.68e-06 ***
---

Residual standard error: 295.9 on 7 degrees of freedom
Multiple R-squared:  0.9511,        Adjusted R-squared:  0.9441
F-statistic: 136.1 on 1 and 7 DF,  p-value: 7.676e-06

### confint(wm1):
                  2.5 %     97.5 %                       2.5 %     97.5 %
(Intercept) -10212.3079 -5599.2995          blength   179.2623   270.3934
```

Data `Kojeni` and `wKojeni`



Data `Kojeni` and `wKojeni`



Illustrations

Data wKojeni replicated

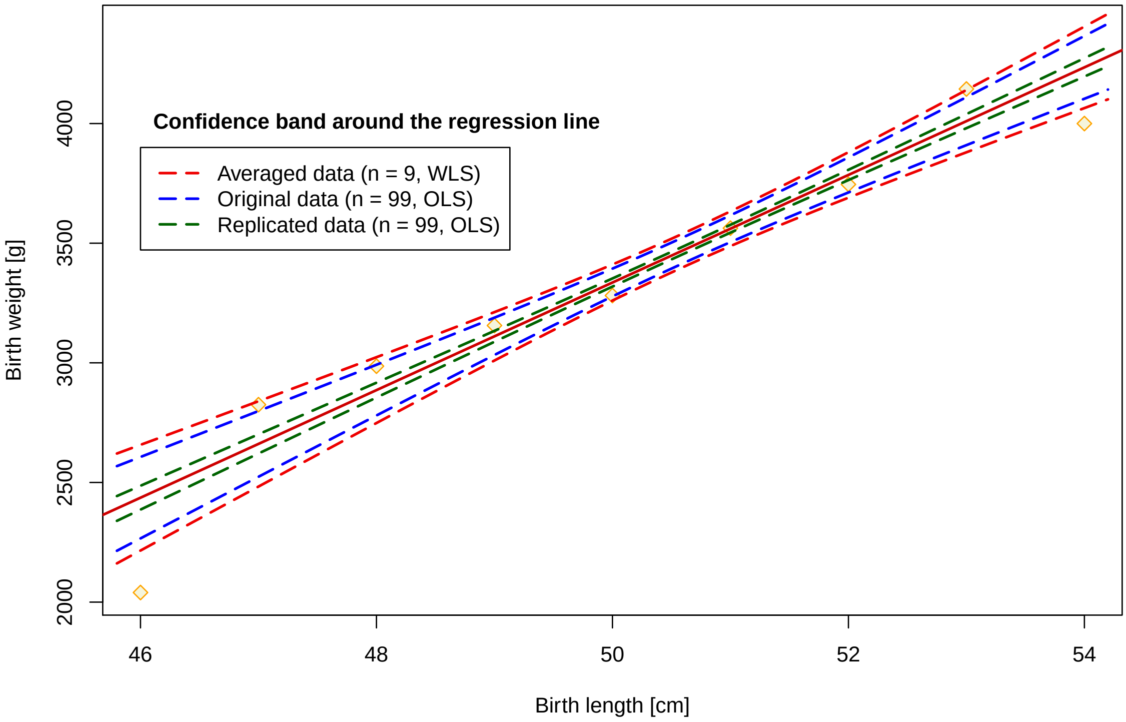
bweight ∼ blength

Ordinary least squares for data replicated from wKojeni

```
replKojeni <- data.frame(bweight = rep(wKojeni[, "bweight"], wKojeni[, "w"]),
                         blength = rep(wKojeni[, "blength"], wKojeni[, "w"]))
m1repl <- lm(bweight ~ blength, data = replKojeni)
summary(m1repl)
confint(m1repl)
```

```
### summary(m1repl):
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7905.804    262.033  -30.17   <2e-16 ***
blength       224.828      5.177   43.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.5 on 97 degrees of freedom
Multiple R-squared:  0.9511,       Adjusted R-squared:  0.9506
F-statistic:  1886 on 1 and 97 DF,  p-value: < 2.2e-16

### confint(m1repl):
                  2.5 %     97.5 %                  2.5 %     97.5 %
(Intercept)  -8425.8658 -7385.7416     blength    214.5539   235.1018
```

Data Kojeni and wKojeni



Chapter 16

Asymptotic Properties of the LSE and Sandwich Estimator

16.1 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:

$$\begin{aligned}
\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big) &= \mathbb{E}\big(Y - \boldsymbol{X}^\top\boldsymbol{\beta} \,\big|\, \boldsymbol{X}\big) && = 0,\\
\mathsf{var}\big(\varepsilon \,\big|\, \boldsymbol{X}\big) &= \mathsf{var}\big(Y - \boldsymbol{X}^\top\boldsymbol{\beta} \,\big|\, \boldsymbol{X}\big) = \mathsf{var}\big(Y \,\big|\, \boldsymbol{X}\big) && =: \sigma^2(\boldsymbol{X}),\\
\mathbb{E}\big(\varepsilon\big) &= \mathbb{E}\Big(\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) = \mathbb{E}\big(0\big) && = 0,\\
\mathsf{var}\big(\varepsilon\big) &= \mathsf{var}\Big(\mathbb{E}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) + \mathbb{E}\Big(\mathsf{var}\big(\varepsilon \,\big|\, \boldsymbol{X}\big)\Big) = \mathsf{var}\big(0\big) + \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\big\} = \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\big\}.
\end{aligned}$$

Assumption (A1).

Let the covariate random vector $\boldsymbol{X} = \bigl(X_0,\,\dots,\,X_{k-1}\bigr)^\top$ satisfy

(i) $\mathbb{E}\bigl|X_j\,X_l\bigr| < \infty, \quad j,\,l = 0,\,\dots,\,k-1;$

(ii) $\mathbb{E}\bigl(\boldsymbol{X}\boldsymbol{X}^\top\bigr) = \mathbb{W}$, where $\mathbb{W}$ is a positive definite matrix.

Notation (Covariates second and first mixed moments).

Let $\mathbb{W} = \bigl(w_{j,l}\bigr)_{j,l=0,\dots,k-1}$. We have,

$$\begin{aligned} w_j^2 := w_{j,j} &= \mathbb{E}\bigl(X_j^2\bigr), \qquad && j = 0,\dots,k-1,\\ w_{j,l} &= \mathbb{E}\bigl(X_j\,X_l\bigr), && j \neq l. \end{aligned}$$

Let

$$\mathbb{V} := \mathbb{W}^{-1} = \bigl(v_{j,l}\bigr)_{j,l=0,\dots,k-1}.$$

Notation (Data of size n).

For $n \geq 1$:

$$\boldsymbol{Y}_n := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{X}_n := \begin{pmatrix} \boldsymbol{X}_1^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix}, \qquad \begin{aligned} \mathbb{W}_n &:= \mathbb{X}_n^\top \mathbb{X}_n = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top, \\ \mathbb{V}_n &:= \bigl(\mathbb{X}_n^\top \mathbb{X}_n\bigr)^{-1} \text{ (if it exists).} \end{aligned}$$

Lemma 16.1 Consistent estimator of the second and first mixed moments of the covariates.

Let assumpions (A0) and (A1) hold. Then

$$\begin{aligned} \frac{1}{n}\,\mathbb{W}_n &\xrightarrow{a.s.} \mathbb{W} \qquad \textit{as}\;\; n \to \infty,\\ n\,\mathbb{V}_n &\xrightarrow{a.s.} \mathbb{V} \qquad \textit{as}\;\; n \to \infty. \end{aligned}$$

Proof. The statement of Lemma follows from applying, for each $j = 0,\dots,k-1$ and $l = 0,\dots,k-1$, the strong law of large numbers for i.i.d. random variables (Theorem C.2) to a sequence

$$Z_{i,j,l} = X_{i,j}\,X_{i,l}, \qquad i = 1,2,\dots.$$

❑

LSE based on data of size n

Since $\frac{1}{n}\,\mathbb{X}_n^\top\mathbb{X}_n \;\overset{\text{a.s.}}{\longrightarrow}\; \mathbb{W} > 0$ then

$$\mathsf{P}\bigl(\text{there exists } n_0 > k \text{ such that for all } n \geq n_0 \quad \mathsf{rank}\bigl(\mathbb{X}_n\bigr) = k\bigr) = 1$$

and we define (for $n \geq n_0$)

$$\widehat{\boldsymbol{\beta}}_n \;=\; \bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbb{X}_n^\top\boldsymbol{Y}_n = \Bigl(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top\Bigr)^{-1}\Bigl(\sum_{i=1}^{n}\boldsymbol{X}_iY_i\Bigr),$$

$$\mathsf{MS}_{e,n} \;=\; \frac{1}{n-k}\,\bigl\|\boldsymbol{Y}_n - \mathbb{X}_n\widehat{\boldsymbol{\beta}}_n\bigr\|^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n)^2,$$

which are the LSE of $\boldsymbol{\beta}$ and the residual mean square based on the assumed linear model for data of size n.

$$\mathsf{M}_n:\ \boldsymbol{Y}_n \,\big|\, \mathbb{X}_n \sim \bigl(\mathbb{X}_n\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\bigr).$$

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 $\mathbf{l} = \bigl(l_0,\, l_1,\, \ldots,\, l_{k-1}\bigr)^\top \neq \mathbf{0}_k$ 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 $\mathsf{M}_n:\ \boldsymbol{Y}_n \,\big|\, \mathbb{X}_n \sim \bigl(\mathbb{X}_n\boldsymbol{\beta},\, \sigma^2\,\mathbf{I}_n\bigr)$ is correct);

(ii) *heteroscedastic* errors where $\mathsf{var}\bigl(\varepsilon \,\big|\, \boldsymbol{X}\bigr)$ 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 $\mathbb{E}\big\{\sigma^2(\boldsymbol{X})\big\} < \infty$ and also for each $j, l = 0, \dots, k-1$,

$$\mathbb{E}\big\{\sigma^2(\boldsymbol{X}) X_j\, X_l\big\} < \infty.$$

Notes.

- Condition (A2 heteroscedastic) states that the matrix

  $$\mathbb{W}^\star := \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\big\}$$

  is a real matrix (with all elements being finite).
- If (A0) and (A1) are assumed then

  (A2 homoscedastic) $\Longrightarrow$ (A2 heteroscedastic).

  Hence everything that will be proved under (A2 heteroscedastic) holds also under (A2 homoscedastic).
- Under assumptions (A0) and (A2 homoscedastic), we have

  $$\mathbb{E}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \boldsymbol{X}_i^\top \boldsymbol{\beta}, \qquad \mathsf{var}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big) = \sigma^2, \qquad i = 1, 2, \dots,$$

  and for each $n > 1$, $Y_1, \dots, Y_n$ 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

  $$\mathbb{E}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \boldsymbol{X}_i^\top \boldsymbol{\beta}, \qquad \mathsf{var}\big(Y_i \,\big|\, \boldsymbol{X}_i\big) = \mathsf{var}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big) = \sigma^2(\boldsymbol{X}_i), \qquad i = 1, 2, \dots,$$

  and for each $n > 1$, $Y_1, \dots, Y_n$ are, given $\mathbb{X}_n$, independent with

  $$\mathbb{E}\big(\boldsymbol{Y}_n \,\big|\, \mathbb{X}_n\big) = \mathbb{X}_n\boldsymbol{\beta}, \qquad \mathsf{var}\big(\boldsymbol{Y}_n \,\big|\, \mathbb{X}_n\big) = \begin{pmatrix} \sigma^2(\boldsymbol{X}_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2(\boldsymbol{X}_n) \end{pmatrix}.$$

16.2 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 16.2 Strong consistency of LSE.

Let assumptions (A0), (A1) and (A2 heteroscedastic*) hold.*
Then

$$\begin{aligned} & \widehat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta} & & \text{as } n \to \infty, \\ \mathbf{1}^\top \widehat{\boldsymbol{\beta}}_n = \; & \widehat{\theta}_n \xrightarrow{a.s.} \theta & = \mathbf{1}^\top \boldsymbol{\beta} \quad & \text{as } n \to \infty, \\ \mathbb{L}\widehat{\boldsymbol{\beta}}_n = \; & \widehat{\boldsymbol{\xi}}_n \xrightarrow{a.s.} \boldsymbol{\xi} & = \mathbb{L}\boldsymbol{\beta} \quad & \text{as } n \to \infty. \end{aligned}$$

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 16.1.

Further

$$\begin{aligned} \boldsymbol{B}_n = \frac{1}{n}\mathbb{X}_n^\top \boldsymbol{Y}_n &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\big(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta} + \boldsymbol{X}_i^\top\boldsymbol{\beta}\big) \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i}_{\boldsymbol{C}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\beta}}_{\boldsymbol{D}_n}. \end{aligned}$$

(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 ($j = 0, \dots, k-1$) 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 $X_{i,j}\varepsilon_i$, $i = 1, 2, \dots$ are i.i.d. by (A0).
- By Cauchy-Schwarz inequality: $\mathbb{E}\big|X_{i,j}\varepsilon_i\big| \le \sqrt{\mathbb{E}X_{i,j}^2\,\mathbb{E}\varepsilon_i^2} < \infty$, because $\mathbb{E}X_{i,j}^2 < \infty$ by (A1) and $\mathbb{E}\varepsilon_i^2 = \mathsf{var}\varepsilon_i = \mathbb{E}\big\{\sigma^2(\boldsymbol{X})\big\} < \infty$ by (A2).

- $$\begin{aligned} \mathbb{E}\big(X_{i,j}\varepsilon_i\big) &= \mathbb{E}\Big(\mathbb{E}\big(X_{i,j}\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big) \\ &= \mathbb{E}\Big(X_{i,j}\mathbb{E}\big(\varepsilon_i \,\big|\, \boldsymbol{X}_i\big)\Big) \\ &= \mathbb{E}\Big(X_{i,j}\, 0\Big) \\ &= 0. \end{aligned}$$

(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 16.1.

In summary: $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{A}_n\left(\boldsymbol{C}_n + \boldsymbol{D}_n\right)$, where

$$\begin{aligned} \boldsymbol{A}_n &\xrightarrow{\text{a.s.}} \mathbb{W}^{-1}, \\ \boldsymbol{C}_n &\xrightarrow{\text{a.s.}} \mathbf{0}_k, \\ \boldsymbol{D}_n &\xrightarrow{\text{a.s.}} \mathbb{W}\,\boldsymbol{\beta}. \end{aligned}$$

Hence

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{\text{a.s.}} \mathbb{W}^{-1}\,\mathbb{W}\,\boldsymbol{\beta} = \boldsymbol{\beta}.$$

❏

Theorem 16.3 Strong consistency of the mean squared error.

Let assumptions (A0), (A1), (A2 homoscedastic*) hold.*

Then

$$\mathsf{MS}_{e,n} \xrightarrow{a.s.} \sigma^2 \quad \text{as } n \to \infty.$$

Proof.

We have

$$\mathsf{MS}_{e,n} = \frac{1}{n-k}\,\mathsf{SS}_{e,n} = \frac{n}{n-k}\,\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}\bigr)^2.$$

Since $\lim\limits_{n\to\infty}\frac{n}{n-k} = 1$, it is sufficient to show that

$$\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n\bigr)^2 \xrightarrow{\text{a.s.}} \sigma^2 \qquad \text{as } n \to \infty.$$

We have

$$\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n\bigr)^2 = \frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta} + \boldsymbol{X}_i^\top\boldsymbol{\beta} - \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_n\bigr)^2$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\bigr)^2}_{\boldsymbol{A}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Bigl\{\boldsymbol{X}_i^\top\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)\Bigr\}^2}_{\boldsymbol{B}_n} + \underbrace{\frac{2}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\bigr)\,\boldsymbol{X}_i^\top\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)}_{\boldsymbol{C}_n}.$$

(a) $\boldsymbol{A}_n = \frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\bigr)^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \xrightarrow{\text{a.s.}} \sigma^2$ 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).
- $\mathbb{E}\bigl(\varepsilon_i\bigr) = 0$

  $\Longrightarrow \mathbb{E}\bigl(\varepsilon_i^2\bigr) = \mathsf{var}\bigl(\varepsilon_i\bigr) = \mathbb{E}\bigl\{\sigma^2(\boldsymbol{X}_i)\bigr\} = \mathbb{E}\bigl(\sigma^2\bigr) = \sigma^2$ by assumption (A2 homoscedastic).
- $\mathbb{E}\bigl|\varepsilon_i^2\bigr| = \mathbb{E}\bigl(\varepsilon_i^2\bigr) = \sigma^2 < \infty$ by assumption (A2 homoscedastic).

(b) $\boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n}\Bigl\{\boldsymbol{X}_i^\top\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)\Bigr\}^2 \xrightarrow{\text{a.s.}} 0$, which is seen as follows.

$$\begin{aligned}
\boldsymbol{B}_n &= \frac{1}{n}\sum_{i=1}^{n}\Bigl\{\boldsymbol{X}_i^\top\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)\Bigr\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)^\top \boldsymbol{X}_i\boldsymbol{X}_i^\top\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr) \\
&= \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)^\top\Bigl(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top\Bigr)\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr) \\
&= \bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr)^\top\Bigl(\frac{1}{n}\mathbb{X}_n^\top\mathbb{X}_n\Bigr)\bigl(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\bigr).
\end{aligned}$$

Now $\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ due to Theorem 16.2.
$\dfrac{1}{n}\mathbb{X}_n^\top \mathbb{X}_n \xrightarrow{\text{a.s.}} \mathbb{W}$ due to Lemma 16.1.

Hence

$$\boldsymbol{B}_n \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top\, \mathbb{W}\, \mathbf{0}_k = 0.$$

(c) $\boldsymbol{C}_n = \dfrac{2}{n}\displaystyle\sum_{i=1}^n \big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\big)\, \boldsymbol{X}_i^\top \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} 0$, which is justified by the following.

$$\begin{aligned}\boldsymbol{C}_n &= \frac{2}{n}\sum_{i=1}^n \big(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\big)\, \boldsymbol{X}_i^\top \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)\\ &= 2\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i^\top\Big)\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big).\end{aligned}$$

Now $\dfrac{1}{n}\displaystyle\sum_{i=1}^n \varepsilon_i \boldsymbol{X}_i^\top \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top$ as was shown in the proof of Theorem 16.2.
$\big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ due to Theorem 16.2.

Hence

$$\boldsymbol{C}_n \xrightarrow{\text{a.s.}} \mathbf{0}_k^\top\, \mathbf{0}_k = 0.$$

In summary: $\mathsf{MS}_{e,n} = \dfrac{n}{n-k}\big(\boldsymbol{A}_n + \boldsymbol{B}_n + \boldsymbol{C}_n\big)$, 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

$$\mathsf{MS}_{e,n} \xrightarrow{\text{a.s.}} 1\,(\sigma^2 + 0 + 0) = \sigma^2.$$

❑

16.3 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. $\mathbb{V} = \Bigl\{\mathbb{E}\bigl(\boldsymbol{X}\boldsymbol{X}^\top\bigr)\Bigr\}^{-1}$.

Theorem 16.4 Asymptotic normality of LSE in homoscedastic case.

Let assumptions (A0), (A1), (A2 homoscedastic*) hold. Further, let* $\mathbb{E}\bigl|\varepsilon^2\, X_j\, X_l\bigr| < \infty$ *for each* $j,\ l = 0,\,\dots,k-1$. *Then*

$$\sqrt{n}\bigl(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}_k\bigl(\mathbf{0}_k,\ \sigma^2\,\mathbb{V}\bigr) \qquad \textit{as } n \to \infty,$$

$$\sqrt{n}\bigl(\widehat{\theta}_n - \theta\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}_1\bigl(0,\ \sigma^2\,\mathbf{l}^\top\mathbb{V}\,\mathbf{l}\bigr) \qquad \textit{as } n \to \infty,$$

$$\sqrt{n}\bigl(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}_m\bigl(\mathbf{0}_m,\ \sigma^2\,\mathbb{L}\,\mathbb{V}\,\mathbb{L}^\top\bigr) \qquad \textit{as } n \to \infty.$$

Proof. Will be provided jointly with Theorem 16.5. ❑

16.3.1 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 6):

$$T_n := \frac{\widehat{\theta}_n - \theta}{\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbf{l}}}, \tag{16.1}$$

$$Q_n := \frac{1}{m}\,\frac{\bigl(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\bigr)^\top\Bigl\{\mathbb{L}\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbb{L}^\top\Bigr\}^{-1}\bigl(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\bigr)}{\mathsf{MS}_{e,n}}. \tag{16.2}$$

Reminder.

- $\mathbb{V}_n = \bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}$.
- Under assumptions (A0) and (A1): $n\,\mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$ as $n \to \infty$.

Consequence of Theorem 16.4: Asymptotic distribution of t- and F-statistics.

Under assumptions of Theorem 16.4:

$$\begin{aligned} T_n &\xrightarrow{\mathcal{D}} \mathcal{N}_1(0,\,1) && \textit{as } n \to \infty, \\ m\,Q_n &\xrightarrow{\mathcal{D}} \chi^2_m && \textit{as } n \to \infty. \end{aligned}$$

Proof. It follows directly from Lemma 16.1, Theorem 16.4 and Cramér-Slutsky theorem (Theorem C.7) as follows.

$$T_n = \frac{\mathbf{l}^\top\widehat{\boldsymbol\beta}_n - \mathbf{l}^\top\boldsymbol\beta}{\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}}} = \underbrace{\frac{\sqrt{n}(\mathbf{l}^\top\widehat{\boldsymbol\beta}_n - \mathbf{l}^\top\boldsymbol\beta)}{\sqrt{\sigma^2\,\mathbf{l}^\top\mathbb{V}\mathbf{l}}}}_{\xrightarrow{\mathcal{D}}\ \mathcal{N}(0,\,1)}\ \underbrace{\sqrt{\frac{\sigma^2\,\mathbf{l}^\top\mathbb{V}\mathbf{l}}{\mathsf{MS}_{e,n}\mathbf{l}^\top\Big\{n\,\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\Big\}\mathbf{l}}}}_{\xrightarrow{\mathsf{P}}\ 1}.$$

$$\begin{aligned} m\,Q_m &= \big(\mathbb{L}\widehat{\boldsymbol\beta}_n - \mathbb{L}\boldsymbol\beta\big)^\top\Big\{\mathsf{MS}_{e,n}\mathbb{L}\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbb{L}^\top\Big\}^{-1}\big(\mathbb{L}\widehat{\boldsymbol\beta}_n - \mathbb{L}\boldsymbol\beta\big) \\ &= \underbrace{\sqrt{n}\big(\mathbb{L}\widehat{\boldsymbol\beta}_n - \mathbb{L}\boldsymbol\beta\big)^\top}_{\xrightarrow{\mathcal{D}}\ \mathcal{N}_m\left(\mathbf{0}_m,\ \sigma^2\,\mathbb{L}\mathbb{V}\mathbb{L}^\top\right)}\ \Big\{\underbrace{\mathsf{MS}_{e,n}\mathbb{L}\,n\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbb{L}^\top}_{\xrightarrow{\mathsf{P}}\ \sigma^2\mathbb{L}\mathbb{V}\mathbb{L}^\top}\Big\}^{-1}\ \underbrace{\big(\mathbb{L}\widehat{\boldsymbol\beta}_n - \mathbb{L}\boldsymbol\beta\big)\sqrt{n}}_{\xrightarrow{\mathcal{D}}\ \mathcal{N}_m\left(\mathbf{0}_m,\ \sigma^2\,\mathbb{L}\mathbb{V}\mathbb{L}^\top\right)}\ . \end{aligned}$$

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 $\boldsymbol{Y}_n \,\big|\, \mathbb{X}_n \sim \mathcal{N}_n\big(\mathbb{X}_n\boldsymbol\beta,\,\sigma^2\mathbf{I}_n\big)$ then Theorem 6.2 (LSE under the normality) provides

$$\begin{aligned} T_n &\sim \mathsf{t}_{n-k}, \\ Q_n &\sim \mathcal{F}_{m,\,n-k}. \end{aligned}$$

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:

$$\begin{aligned} &T_\nu \sim \mathsf{t}_\nu && \text{then} \quad T_\nu \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,1) && \text{as } \nu \to \infty, \\ &Q_\nu \sim \mathcal{F}_{m,\,\nu} && \text{then} \quad m\,Q_\nu \xrightarrow{\mathcal{D}} \chi^2_m && \text{as } \nu \to \infty. \end{aligned} \tag{16.3}$$

This, together with Consequence of Theorem 16.4 then justify asymptotic validity of a classical inference based on statistics T_n (Eq. 16.1) and Q_n (Eq. 16.2), 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 16.4.

That is, for example, both intervals

(i) $\mathcal{I}_n^{\mathcal{N}} := \Big(\widehat\theta_n - u(1-\alpha/2)\,\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}},\ \ \widehat\theta_n + u(1-\alpha/2)\,\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\big(\mathbb{X}_n^\top\mathbb{X}_n\big)^{-1}\mathbf{l}}\Big);$

(ii) $\mathcal{I}_n^{\mathsf{t}} := \Bigl(\widehat{\theta}_n - \mathsf{t}_{n-k}(1-\alpha/2)\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbf{l}},\quad \widehat{\theta}_n + \mathsf{t}_{n-k}(1-\alpha/2)\sqrt{\mathsf{MS}_{e,n}\,\mathbf{l}^\top\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbf{l}}\Bigr)$,

satisfy, for any $\theta^0 \in \mathbb{R}$ (even without normality of the error terms)

$$\mathsf{P}\bigl(\mathcal{I}_n^{\mathcal{N}} \ni \theta^0;\ \theta = \theta^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\bigl(\mathcal{I}_n^{\mathsf{t}} \ni \theta^0;\ \theta = \theta^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

Analogously, due to a general asymptotic property of the F-distribution (Eq. 16.3), 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) $\mathcal{K}_n^{\chi} := \Bigl\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr)^\top \Bigl\{\mathsf{MS}_{e,n}\mathbb{L}\,\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbb{L}^\top\Bigr\}^{-1}\bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr) \ <\ \chi^2_m(1-\alpha)\Bigr\}$;

(ii) $\mathcal{K}_n^{\mathcal{F}} := \Bigl\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr)^\top \Bigl\{\mathsf{MS}_{e,n}\mathbb{L}\,\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}\mathbb{L}^\top\Bigr\}^{-1}\bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr) \ <\ m\,\mathcal{F}_{m,n-k}(1-\alpha)\Bigr\}$,

we have for any $\boldsymbol{\xi}^0 \in \mathbb{R}^m$ (under assumptions of Theorems 16.4):

$$\mathsf{P}\bigl(\mathcal{K}_n^{\chi} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\bigl(\mathcal{K}_n^{\mathcal{F}} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

16.4 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.

- $\mathbb{V} = \Bigl\{\mathbb{E}\bigl(\boldsymbol{X}\boldsymbol{X}^\top\bigr)\Bigr\}^{-1}$.
- $\mathbb{W}^\star = \mathbb{E}\bigl\{\sigma^2(\boldsymbol{X})\,\boldsymbol{X}\boldsymbol{X}^\top\bigr\}$.

Theorem 16.5 Asymptotic normality of LSE in heteroscedastic case.

Let assumptions (A0), (A1), (A2 heteroscedastic*) hold. Further, let* $\mathbb{E}\bigl|\varepsilon^2\,X_j\,X_l\bigr| < \infty$ *for each* $j,\,l = 0,\,\dots,k-1$. *Then*

$$\begin{aligned}
\sqrt{n}\bigl(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\bigr) &\xrightarrow{\mathcal{D}} \mathcal{N}_k\bigl(\mathbf{0}_k,\ \mathbb{V}\mathbb{W}^\star\mathbb{V}\bigr) && \textit{as } n\to\infty,\\
\sqrt{n}\bigl(\widehat{\theta}_n - \theta\bigr) &\xrightarrow{\mathcal{D}} \mathcal{N}_1\bigl(0,\ \mathbf{l}^\top\mathbb{V}\mathbb{W}^\star\mathbb{V}\,\mathbf{l}\bigr) && \textit{as } n\to\infty,\\
\sqrt{n}\bigl(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\bigr) &\xrightarrow{\mathcal{D}} \mathcal{N}_m\bigl(\mathbf{0}_m,\ \mathbb{L}\,\mathbb{V}\mathbb{W}^\star\mathbb{V}\,\mathbb{L}^\top\bigr) && \textit{as } n\to\infty.
\end{aligned}$$

Proof. We will jointly prove also Theorem 16.4.

We have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_n &= \underbrace{\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}}_{\mathbb{V}_n}\,\mathbb{X}_n^\top\boldsymbol{Y}_n\\
&= \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i Y_i\\
&= \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\bigl(\boldsymbol{X}_i^\top\boldsymbol{\beta} + \varepsilon_i\bigr)\\
&= \mathbb{V}_n\underbrace{\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\right)}_{\mathbb{V}_n^{-1}}\boldsymbol{\beta} + \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\\
&= \boldsymbol{\beta} + \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i.
\end{aligned}$$

That is,

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \mathbb{V}_n\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i = n\,\mathbb{V}_n\,\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i. \tag{16.4}$$

By Lemma 16.1, $n\,\mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$ which implies

$$n\,\mathbb{V}_n \xrightarrow{\mathsf{P}} \mathbb{V} \qquad \text{as } n\to\infty. \tag{16.5}$$

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 $\boldsymbol{X}_i\,\varepsilon_i$, $i = 1,\dots,n$. The mean of the distribution of those random vectors is

$$\mathbb{E}\big(\boldsymbol{X}\varepsilon\big) \;=\; \mathbf{0}_k \qquad \text{(was shown in the proof of Theorem 16.2).}$$

The covariance matrix is equal to $\mathsf{var}\big(\boldsymbol{X}\varepsilon\big)$. All elements of this covariance matrix are finite due to assumption $\mathbb{E}\big|\varepsilon^2\,X_j\,X_l\big| < \infty$ for each $j,\, l = 0,\,\dots, k-1$.

Then

$$\begin{aligned}\mathsf{var}\big(\boldsymbol{X}\varepsilon\big) &= \mathbb{E}\Big(\mathsf{var}\big(\boldsymbol{X}\varepsilon\,\big|\,\boldsymbol{X}\big)\Big) + \mathsf{var}\Big(\mathbb{E}\big(\boldsymbol{X}\varepsilon\,\big|\,\boldsymbol{X}\big)\Big)\\ &= \mathbb{E}\Big(\boldsymbol{X}\,\underbrace{\mathsf{var}\big(\varepsilon\,\big|\,\boldsymbol{X}\big)}_{\sigma^2(\boldsymbol{X})}\,\boldsymbol{X}^\top\Big) + \mathsf{var}\big(\boldsymbol{X}\,\underbrace{\mathbb{E}\big(\varepsilon\,\big|\,\boldsymbol{X}\big)}_{0}\big)\\ &= \mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big).\end{aligned}$$

Depending, on whether (A2 homoscedastic) or (A2 heteroscedastic) is assumed, we have

$$\mathsf{var}\big(\boldsymbol{X}\varepsilon\big) \;=\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big) \;=\; \begin{cases}\sigma^2\,\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big) = \sigma^2\mathbb{W}, & \text{(A2 homoscedastic)},\\ \mathbb{W}^\star, & \text{(A2 heteroscedastic)}.\end{cases} \tag{16.6}$$

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 (16.4) and (16.5), 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 (16.6) 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)

$$\mathbb{V}\,\mathbb{E}\big(\sigma^2(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top\big)\,\mathbb{V}^\top = \mathbb{V}\,\mathbb{W}^\star\,\mathbb{V}$$

and hence

$$\sqrt{n}\,\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}_k\Big(\mathbf{0}_k,\; \mathbb{V}\,\mathbb{W}^\star\,\mathbb{V}\Big) \qquad \text{as } n\to\infty.$$

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

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

- Hat matrix: $\mathbb{H}_n = \mathbb{X}_n \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \mathbb{X}_n^\top$;
- 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.

- $\mathbb{V}_n = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} = \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1}$.
- Under assumptions (A0) and (A1): $n \, \mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$ as $n \to \infty$.

Theorem 16.6 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

$$n \, \mathbb{V}_n \, \mathbb{W}_n^\star \, \mathbb{V}_n \ \xrightarrow{\text{a.s.}} \ \mathbb{V} \, \mathbb{W}^\star \, \mathbb{V} \qquad \text{as } n \to \infty,$$

where for $n = 1, \, 2, \, \ldots$,

$$\mathbb{W}_n^\star = \sum_{i=1}^{n} U_{n,i}^2 \, \boldsymbol{X}_i \boldsymbol{X}_i^\top = \mathbb{X}_n^\top \boldsymbol{\Omega}_n \mathbb{X}_n,$$

$$\boldsymbol{\Omega}_n = \mathsf{diag}\big(\omega_{n,1}, \, \ldots, \, \omega_{n,n}\big), \qquad \omega_{n,i} = U_{n,i}^2, \quad i = 1, \ldots, n.$$

Proof.

First, remind that

$$\mathbb{V} \, \mathbb{W}^\star \, \mathbb{V} \ = \ \Big\{\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\Big\}^{-1} \mathbb{E}\big(\sigma^2(\boldsymbol{X}) \, \boldsymbol{X}\boldsymbol{X}^\top\big) \, \Big\{\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\Big\}^{-1},$$

and we know from Lemma 16.1 that

$$n \, \mathbb{V}_n \ = \ n \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \ \xrightarrow{\text{a.s.}} \ \Big\{\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\Big\}^{-1} = \mathbb{V} \qquad \text{as } n \to \infty.$$

Hence, if we show that

$$\frac{1}{n} \, \mathbb{W}_n^\star \ = \ \frac{1}{n} \sum_{i=1}^{n} U_{n,i}^2 \, \boldsymbol{X}_i \, \boldsymbol{X}_i^\top \ \xrightarrow{\text{a.s.}} \ \mathbb{E}\big(\sigma^2(\boldsymbol{X}) \, \boldsymbol{X}\boldsymbol{X}^\top\big) = \mathbb{W}^\star \qquad \text{as } n \to \infty,$$

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 $j,\, l = 0, \dots, k-1$

$$\begin{aligned}\mathbb{E}\big(\varepsilon^2\, X_j\, X_l\big) &= \mathbb{E}\Big(\mathbb{E}\big(\varepsilon^2\, X_j\, X_l \,\big|\, \boldsymbol{X}\big)\Big)\\ &= \mathbb{E}\Big(X_j\, X_l\, \mathbb{E}\big(\varepsilon^2 \,\big|\, \boldsymbol{X}\big)\Big)\\ &= \mathbb{E}\big(\sigma^2(\boldsymbol{X})\, X_j\, X_l\big).\end{aligned}$$

For each $j,\, l = 0, \dots, k-1$,

$$\mathbb{E}\big|\varepsilon^2\, X_j\, X_l\big| < \infty$$

by assumptions of Theorem. By assumption (A0), $\varepsilon_i\, X_{i,j}\, X_{i,l}$, $i = 1, 2, \dots$, 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,

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\, \boldsymbol{X}_i \boldsymbol{X}_i^\top \;\xrightarrow{\text{a.s.}}\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\, \boldsymbol{X}\boldsymbol{X}^\top\big) \;=\; \mathbb{W}^\star \qquad \text{as } n \to \infty. \tag{16.7}$$

In the following, we show that (unobservable) squared error terms ε_i^2 in (16.7) 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 $\mathbb{W}^\star$ as in (16.7).

We have

$$\begin{aligned}\underbrace{\frac{1}{n}\sum_{i=1}^{n} U_{n,i}^2\, \boldsymbol{X}_i \boldsymbol{X}_i^\top}_{\mathbb{W}_n^\star} &= \frac{1}{n}\sum_{i=1}^{n} \big(Y_i \,-\, \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n\big)^2\, \boldsymbol{X}_i\, \boldsymbol{X}_i^\top\\ &= \frac{1}{n}\sum_{i=1}^{n} \big(\underbrace{Y_i \,-\, \boldsymbol{X}_i^\top \boldsymbol{\beta}}_{\varepsilon_i} + \boldsymbol{X}_i^\top \boldsymbol{\beta} \,-\, \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n\big)^2\, \boldsymbol{X}_i\, \boldsymbol{X}_i^\top\\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\, \boldsymbol{X}_i \boldsymbol{X}_i^\top}_{\mathbb{A}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)\, \boldsymbol{X}_i \boldsymbol{X}_i^\top}_{\mathbb{B}_n}\\ &\quad + \underbrace{\frac{2}{n}\sum_{i=1}^{n} \big(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n\big)^\top \boldsymbol{X}_i \varepsilon_i\, \boldsymbol{X}_i \boldsymbol{X}_i^\top}_{\mathbb{C}_n}.\end{aligned}$$

(a) $\mathbb{A}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \varepsilon_i^2\, \boldsymbol{X}_i \boldsymbol{X}_i^\top \;\xrightarrow{\text{a.s.}}\; \mathbb{E}\big(\sigma^2(\boldsymbol{X})\, \boldsymbol{X}\boldsymbol{X}^\top\big) \;=\; \mathbb{W}^\star$ due to (16.7).

(b) To work with $\mathbb{B}_n = \dfrac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top \boldsymbol{X}_i\boldsymbol{X}_i^\top\,\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)\,\boldsymbol{X}_i\boldsymbol{X}_i^\top$, 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 16.2: $\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ 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) $\mathbb{C}_n \;=\; \dfrac{2}{n}\sum_{i=1}^{n}\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big)^\top \boldsymbol{X}_i\varepsilon_i\,\boldsymbol{X}_i\boldsymbol{X}_i^\top$ 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 16.2: $\big(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n\big) \xrightarrow{\text{a.s.}} \mathbf{0}_k$ as $n\to\infty$.
- Due to assumption (A0) and assumption $\mathbb{E}\big|\varepsilon\,X_s\,X_j\,X_l\big| < \infty$ for any $s,\,j,\,l = 0,\ldots,k-1$, by Theorem C.2 (SLLN, i.i.d.), for any $j,\,l = 0,\ldots,k-1$:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\,X_{i,j}X_{i,l} \;\xrightarrow{\text{a.s.}}\; \mathbb{E}\big(\boldsymbol{X}\varepsilon\,X_jX_l\big).$$

- Hence, for any $j,\,l = 0,\ldots,k-1$, $C_n(j,\,l) \xrightarrow{\text{a.s.}} 2\,\mathbf{0}_k^\top\,\mathbb{E}\big(\boldsymbol{X}\varepsilon\,X_jX_l\big) = 0$ and finally,

$$\mathbb{C}_n \;\xrightarrow{\text{a.s.}}\; \mathbf{0}_{k\times k} \qquad \text{as } n\to\infty.$$

In summary:

$$\begin{aligned} n\,\mathbb{V}_n\mathbb{W}_n^\star\mathbb{V}_n \;&=\; n\,\mathbb{V}_n\,\left(\frac{1}{n}\mathbb{W}_n^\star\right)\,n\,\mathbb{V}_n \\ &=\; n\,\mathbb{V}_n\,\big(\mathbb{A}_n \,+\, \mathbb{B}_n \,+\, \mathbb{C}_n\big)\,n\,\mathbb{V}_n, \end{aligned}$$

where $n\,\mathbb{V}_n \xrightarrow{\text{a.s.}} \mathbb{V}$,

$$\mathbb{A}_n \xrightarrow{\text{a.s.}} \mathbb{W}^{\star},$$

$$\mathbb{B}_n \xrightarrow{\text{a.s.}} \mathbf{0}_{k\times k},$$

$$\mathbb{C}_n \xrightarrow{\text{a.s.}} \mathbf{0}_{k\times k}.$$

Hence

$$n\,\mathbb{V}_n\mathbb{W}_n^{\star}\mathbb{V}_n \quad \xrightarrow{\text{a.s.}} \quad \mathbb{V}\left(\mathbb{W}^{\star} \; + \; \mathbf{0}_{k\times k} \; + \; \mathbf{0}_{k\times k}\right)\mathbb{V} = \mathbb{V}\mathbb{W}^{\star}\mathbb{V} \qquad \text{as } n \to \infty.$$

❑

Terminology (Heteroscedasticity consistent (sandwich) estimator of the covariance matrix).

Matrix

$$\mathbb{V}_n\,\mathbb{W}_n^{\star}\,\mathbb{V}_n = \left(\mathbb{X}_n^{\top}\mathbb{X}_n\right)^{-1}\mathbb{X}_n^{\top}\,\boldsymbol{\Omega}_n\,\mathbb{X}_n\left(\mathbb{X}_n^{\top}\mathbb{X}_n\right)^{-1} \tag{16.8}$$

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 (16.8) 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 16.6 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 $\boldsymbol{\Omega}_n$ can, for a chosen sequence ν_n, such that $\frac{n}{\nu_n} \to 1$ as $n \to \infty$ and a suitable sequence $\boldsymbol{\delta}_n = \left(\delta_{n,1},\,\ldots,\,\delta_{n,n}\right)$, $n = 1,\,2,\,\ldots$, be replaced by a matrix
$$\boldsymbol{\Omega}_n^{HC} := \mathsf{diag}\left(\omega_{n,1},\,\ldots,\,\omega_{n,n}\right),$$
$$\omega_{n,i} = \frac{n}{\nu_n}\,\frac{U_{n,i}^2}{m_{n,i}^{\delta_{n,i}}}, \qquad\qquad i = 1,\ldots,n.$$

- 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 16.6.

  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: $\nu_n = n$, $\delta_{n,i} = \min\bigl\{4,\, n\, h_{n,i}/k\bigr\}$, 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

$$\delta_{n,i} = \min\Bigl\{4,\, \frac{h_{n,i}}{\overline{h}_n}\Bigr\}, \qquad \overline{h}_n = \frac{1}{n}\sum_{i=1}^{n} h_{n,i}.$$

- 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.

16.4.1 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 $n \geq n_0 > k$,

$$\mathbb{V}_n^{HC} := \bigl(\mathbb{X}_n^\top \mathbb{X}_n\bigr)^{-1} \mathbb{X}_n^\top\, \boldsymbol{\Omega}_n^{HC}\, \mathbb{X}_n\, \bigl(\mathbb{X}_n^\top \mathbb{X}_n\bigr)^{-1}.$$

Finally, let the statistics T_n^{HC} and Q_n^{HC} be defined as

$$T_n^{HC} := \frac{\widehat{\theta}_n - \theta}{\sqrt{\mathbf{l}^\top \mathbb{V}_n^{HC}\, \mathbf{l}}},$$

$$Q_n^{HC} := \frac{1}{m}\bigl(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\bigr)^\top \bigl(\mathbb{L}\mathbb{V}_n^{HC}\mathbb{L}^\top\bigr)^{-1}\bigl(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}\bigr).$$

Note that the statistics T_n^{HC} and Q_n^{HC}, respectively, are the usual statistics T_n (Eq. 16.1) and Q_n (16.2), respectively, in which the term $\mathsf{MS}_{e,n}\,\bigl(\mathbb{X}_n^\top\mathbb{X}_n\bigr)^{-1}$ is replaced by the sandwich estimator $\mathbb{V}_n^{HC}$.

Consequence of Theorems 16.5 and 16.6: Heteroscedasticity consistent asymptotic inference.

Under assumptions of Theorem 16.5 and 16.6:

$$\begin{aligned} T_n^{HC} &\xrightarrow{\mathcal{D}} \mathcal{N}_1(0,\,1) && \textit{as } n \to \infty,\\ m\, Q_n^{HC} &\xrightarrow{\mathcal{D}} \chi_m^2 && \textit{as } n \to \infty. \end{aligned}$$

Proof. **Proof/calculations are available in the handnotes.** ❑

Due to a general asymptotic property of the Student t-distribution (Eq. 16.3), 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 16.5 and 16.6, both intervals

(i) $\mathcal{I}_n^{\mathcal{N}} := \Bigl(\widehat{\theta}_n \,-\, u(1-\alpha/2)\,\sqrt{\mathbf{l}^\top\,\mathbb{V}_n^{HC}\,\mathbf{l}},\quad \widehat{\theta}_n \,+\, u(1-\alpha/2)\,\sqrt{\mathbf{l}^\top\,\mathbb{V}_n^{HC}\,\mathbf{l}}\Bigr);$

(ii) $\mathcal{I}_n^{\mathsf{t}} := \Bigl(\widehat{\theta}_n \,-\, \mathsf{t}_{n-k}(1-\alpha/2)\,\sqrt{\mathbf{l}^\top\,\mathbb{V}_n^{HC}\,\mathbf{l}},\quad \widehat{\theta}_n \,+\, \mathsf{t}_{n-k}(1-\alpha/2)\,\sqrt{\mathbf{l}^\top\,\mathbb{V}_n^{HC}\,\mathbf{l}}\Bigr),$

satisfy, for any $\theta^0 \in \mathbb{R}$:

$$\mathsf{P}\bigl(\mathcal{I}_n^{\mathcal{N}} \ni \theta^0;\ \theta = \theta^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\bigl(\mathcal{I}_n^{\mathsf{t}} \ni \theta^0;\ \theta = \theta^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

Analogously, due to a general asymptotic property of the F-distribution (Eq. 16.3), 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

(i) $\mathcal{K}_n^{\chi} := \Bigl\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr)^\top \bigl(\mathbb{L}\,\mathbb{V}_n^{HC}\mathbb{L}^\top\bigr)^{-1}\bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr) \,<\, \chi^2_m(1-\alpha)\Bigr\};$

(ii) $\mathcal{K}_n^{\mathcal{F}} := \Bigl\{\boldsymbol{\xi} \in \mathbb{R}^m :\ \bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr)^\top \bigl(\mathbb{L}\,\mathbb{V}_n^{HC}\mathbb{L}^\top\bigr)^{-1}\bigl(\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\bigr) \,<\, m\,\mathcal{F}_{m,n-k}(1-\alpha)\Bigr\},$

we have for any $\boldsymbol{\xi}^0 \in \mathbb{R}^m$ (under assumptions of Theorems 16.5 and 16.6):

$$\mathsf{P}\bigl(\mathcal{K}_n^{\chi} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty,$$

$$\mathsf{P}\bigl(\mathcal{K}_n^{\mathcal{F}} \ni \boldsymbol{\xi}^0;\ \boldsymbol{\xi} = \boldsymbol{\xi}^0\bigr) \longrightarrow 1-\alpha \qquad \text{as } n \to \infty.$$

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) $\big(\mathbb{A}\mathbb{A}^{+}\big)^{\top} = \mathbb{A}\mathbb{A}^{+}$;

(iv) $\big(\mathbb{A}^{+}\mathbb{A}\big)^{\top} = \mathbb{A}^{+}\mathbb{A}$.

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*

$$\mathbb{D}=\mathbb{A}\otimes\mathbb{C}=\begin{pmatrix}a_{1,1}\mathbb{C} & \dots & a_{1,s}\mathbb{C}\\ \vdots & \vdots & \vdots\\ a_{r,1}\mathbb{C} & \dots & a_{r,s}\mathbb{C}\end{pmatrix}=\bigl(a_{i,j}\mathbb{C}\bigr)_{i=1,\dots,m,j=1,\dots,n}.$$

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) $\bigl(\mathbb{A}\otimes\mathbb{C}\bigr)^{-1}=\mathbb{A}^{-1}\,\otimes\,\mathbb{C}^{-1}$, *if the inversions exist.*

(vii) $\bigl(\mathbb{A}\otimes\mathbb{C}\bigr)^{-}=\mathbb{A}^{-}\,\otimes\,\mathbb{C}^{-}$, *for arbitrary pseudoinversions.*

(viii) $\bigl(\mathbb{A}\otimes\mathbb{C}\bigr)^{\top}=\mathbb{A}^\top\otimes\mathbb{C}^\top$.

(ix) $\bigl(\mathbb{A},\;\mathbb{C}\bigr)\,\otimes\,\mathbb{D}=\bigl(\mathbb{A}\otimes\mathbb{D},\;\;\mathbb{C}\otimes\mathbb{D}\bigr)$.

(x) *Upon a suitable reordering of the columns, matrices* $\bigl(\mathbb{A}\otimes\mathbb{C},\;\;\mathbb{A}\otimes\mathbb{D}\bigr)$ *and* $\mathbb{A}\,\otimes\,\bigl(\mathbb{C},\;\;\mathbb{D}\bigr)$ *are the same.*

(xi) $\mathsf{rank}\bigl(\mathbb{A}\otimes\mathbb{C}\bigr)=\mathsf{rank}\bigl(\mathbb{A}\bigr)\,\mathsf{rank}\bigl(\mathbb{C}\bigr)$.

Proof. See Rao (1973, Section 1b.8). ❑

Definition A.4 Elementwise product of two vectors.

Let $\boldsymbol{a} = \bigl(a_1, \dots, a_p\bigr)^\top \in \mathbb{R}^p$, $\boldsymbol{c} = \bigl(c_1, \dots, c_p\bigr)^\top \in \mathbb{R}^p$. *Their* elementwise product[1] *is a vector* $\bigl(a_1\, c_1,\, \dots,\, a_p\, c_p\bigr)^\top$ *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

$$\mathbb{A}_{n\times p} = \bigl(\boldsymbol{a}^1,\, \dots,\, \boldsymbol{a}^p\bigr) \quad \textit{and}\ \mathbb{C}_{n\times q} = \bigl(\boldsymbol{c}^1,\, \dots,\, \boldsymbol{c}^q\bigr)$$

be real matrices. Their columnwise product[2] $\mathbb{A} : \mathbb{C}$ *is a matrix* $\mathbb{D}_{n\times p\cdot q}$ *such that*

$$\mathbb{D} = \mathbb{A} : \mathbb{C} = \bigl(\boldsymbol{a}^1 : \boldsymbol{c}^1,\, \dots,\, \boldsymbol{a}^p : \boldsymbol{c}^1,\, \dots,\, \boldsymbol{a}^1 : \boldsymbol{c}^q,\, \dots,\, \boldsymbol{a}^p : \boldsymbol{c}^q\bigr).$$

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

$$\mathbb{A} : \mathbb{C} = \begin{pmatrix} \boldsymbol{a}_1^\top \otimes \boldsymbol{c}_1^\top \\ \vdots \\ \boldsymbol{a}_n^\top \otimes \boldsymbol{c}_n^\top \end{pmatrix} = \bigl(\boldsymbol{a}^1 : \boldsymbol{c}^1,\, \dots,\, \boldsymbol{a}^1 : \boldsymbol{c}^q,\, \dots,\, \boldsymbol{a}^p : \boldsymbol{c}^1,\, \dots,\, \boldsymbol{a}^p : \boldsymbol{c}^q\bigr),$$

  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

$$T = \frac{U + \lambda}{\sqrt{\dfrac{V}{\nu}}}$$

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:

$$\mathbb{E}(T) = \begin{cases} \lambda\sqrt{\dfrac{\nu}{2}}\,\dfrac{\Gamma\bigl(\frac{\nu-1}{2}\bigr)}{\Gamma\bigl(\frac{\nu}{2}\bigr)}, & \text{if } \nu > 1, \\ \text{does not exist,} & \text{if } \nu \le 1. \end{cases}$$

$$\mathsf{var}(T) = \begin{cases} \dfrac{\nu(1+\lambda^2)}{\nu-2} - \dfrac{\nu\lambda^2}{2}\left\{\dfrac{\Gamma\bigl(\frac{\nu-1}{2}\bigr)}{\Gamma\bigl(\frac{\nu}{2}\bigr)}\right\}^2, & \text{if } \nu > 2, \\ \text{does not exist,} & \text{if } \nu \le 2. \end{cases}$$

[1] *necentrální Studentovo t-rozdělení* [2] *stupně volnosti* [3] *parametr necentrality*

Definition B.2 Non-central χ^2 distribution.

Let $U_1, \dots, U_k$ be independent *random variables. Let further $U_i \sim \mathcal{N}(\mu_i,\, 1)$, $i = 1, \dots, k$, for some $\mu_1, \dots, \mu_k \in \mathbb{R}$. That is $\boldsymbol{U} = \bigl(U_1, \dots, U_k\bigr)^\top \sim \mathcal{N}_k\bigl(\boldsymbol{\mu},\, \mathbf{I}_k\bigr)$, where $\boldsymbol{\mu} = \bigl(\mu_1, \dots, \mu_k\bigr)^\top$. 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

$$X \sim \chi_k^2(\lambda).$$

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 $\mu_1, \dots, \mu_k$.
- As an exercise for the use of a convolution theorem, we can derive a density of the $\chi_k^2(\lambda)$ distribution which is
$$f(x) = \begin{cases} \dfrac{\mathrm{e}^{-\frac{x+\lambda}{2}}\, x^{\frac{k-2}{2}}}{2^{\frac{k}{2}}\,\Gamma\bigl(\frac{k-1}{2}\bigr)\,\Gamma\bigl(\frac{1}{2}\bigr)} \displaystyle\sum_{j=0}^{\infty} \frac{\lambda^j\, x^j}{(2j)!}\, B\Bigl(\frac{k-1}{2},\, \frac{1}{2} + j\Bigr), & x > 0, \\ 0, & x \leq 0. \end{cases}$$
- 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.
- $\chi_\nu^2(0) \equiv \chi_\nu^2$.
- 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*

$$Q = \frac{\dfrac{X}{\nu_1}}{\dfrac{Y}{\nu_2}}$$

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 $\boldsymbol{U} \sim \mathcal{N}_p(\boldsymbol{0}_p,\, \boldsymbol{\Sigma})$, 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

$$f(\boldsymbol{t}) = \frac{\Gamma\bigl(\frac{\nu+p}{2}\bigr)}{\Gamma\bigl(\frac{\nu}{2}\bigr)\,\nu^{\frac{p}{2}}\,\pi^{\frac{p}{2}}}\,\bigl|\boldsymbol{\Sigma}\bigr|^{-\frac{1}{2}}\left\{1 + \frac{\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}}{\nu}\right\}^{-\frac{\nu+p}{2}}, \qquad \boldsymbol{t} \in \mathbb{R}^p.$$

- Expectation and a covariance matrix of $\boldsymbol{T} \sim \mathsf{mvt}_{p,\nu}(\boldsymbol{\Sigma})$ are

$$\mathbb{E}(\boldsymbol{T}) = \begin{cases} \boldsymbol{0}_p, & \text{if } \nu > 1,\\[2ex] \text{does not exist}, & \text{if } \nu \le 1.\end{cases}$$

$$\mathsf{var}(\boldsymbol{T}) = \begin{cases} \dfrac{\nu}{\nu-2}\,\boldsymbol{\Sigma}, & \text{if } \nu > 2,\\[2ex] \text{does not exist}, & \text{if } \nu \le 2.\end{cases}$$

Lemma B.1 Marginals of the multivariate Student t-distribution.

Let $\boldsymbol{T} = \bigl(T_1,\dots,T_p\bigr)^\top \sim \mathsf{mvt}_{p,\nu}(\boldsymbol{\Sigma})$, where the scale matrix $\boldsymbol{\Sigma}$ has positive diagonal elements $\sigma_1^2 > 0, \dots, \sigma_p^2 > 0$. Then

$$\frac{T_j}{\sigma_j} \sim \mathsf{t}_\nu, \qquad j = 1,\dots,p.$$

Proof.

- From definition of the multivariate t-distribution, $\boldsymbol{T}$ can be written as $\boldsymbol{T} = \boldsymbol{U}\sqrt{\dfrac{\nu}{V}}$, where $\boldsymbol{U} = \bigl(U_1,\dots,U_p\bigr)^\top \sim \mathcal{N}_p(\boldsymbol{0}_p,\, \boldsymbol{\Sigma})$ and $V \sim \chi^2_\nu$ are independent.
- Then for all $j = 1,\dots,p$:

$$\frac{T_j}{\sigma_j} = \frac{U_j}{\sigma_j}\sqrt{\frac{\nu}{V}} = \frac{Z_j}{\sqrt{\frac{V}{\nu}}},$$

where $Z_j \sim \mathcal{N}(0,\,1)$ is independent of $V \sim \chi^2_\nu$.

❑

[6] *vícerozměrné Studentovo t-rozdělení* [7] *měřítková matice*

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: $\boldsymbol{\phi} = \big(\phi_1, \dots, \phi_{n-1}\big)^\top$.
- 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 $Z_1, Z_2, \dots$ *be a sequence of* independent not necessarily identically distributed (i.n.n.i.d.) *random variables. Let* $\mathbb{E}(Z_i) = \mu_i$, $\mathsf{var}(Z_i) = \sigma_i^2$, $i = 1, 2, \dots$. *Let*

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty.$$

Then

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu_i) \overset{\text{a.s.}}{\longrightarrow} 0 \qquad \text{as } n \to \infty.$$

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 $Z_1, Z_2, \dots$ *be a sequence of* independent identically distributed (i.i.d.) *random variables.*
Then

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \overset{\text{a.s.}}{\longrightarrow} \mu \qquad \text{as } n \to \infty$$

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 $Z_1,\,Z_2,\,\dots$ be a sequence of i.n.n.i.d. random variables with

$$\mathbb{E}\bigl(Z_i\bigr) = \mu_i, \quad \infty > \mathsf{var}\bigl(Z_i\bigr) = \sigma_i^2 > 0, \qquad i = 1,\,2,\,\dots$$

Let for some $\delta > 0$

$$\frac{\sum_{i=1}^n \mathbb{E}\bigl|Z_i - \mu_i\bigr|^{2+\delta}}{\Bigl(\sum_{i=1}^n \sigma_i^2\Bigr)^{\frac{2+\delta}{2}}} \longrightarrow 0 \quad \textit{as } n \to \infty.$$

Then

$$\frac{\sum_{i=1}^n (Z_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,1) \quad \textit{as } n \to \infty.$$

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 $Z_1,\,Z_2,\,\dots$ be a sequence of i.i.d. random variables with

$$\mathbb{E}\bigl(Z_i\bigr) = \mu, \quad \infty > \mathsf{var}\bigl(Z_i\bigr) = \sigma^2 > 0, \qquad i = 1,\,2,\,\dots.$$

Let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$.

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i - \mu}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,1) \quad \textit{as } n \to \infty,$$

$$\sqrt{n}\,\bigl(\overline{Z}_n - \mu\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,\sigma^2) \qquad \textit{as } n \to \infty.$$

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 $\boldsymbol{Z}_1,\,\boldsymbol{Z}_2,\,\dots$ be a sequence of i.i.d. p-dimensional random vectors with

$$\mathbb{E}\bigl(\boldsymbol{Z}_i\bigr) = \boldsymbol{\mu}, \qquad \mathsf{var}\bigl(\boldsymbol{Z}_i\bigr) = \boldsymbol{\Sigma}, \qquad i = 1,\,2,\,\dots,$$

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

$$\boldsymbol{Z}_n \xrightarrow{\mathcal{D}} \boldsymbol{Z} \qquad \textit{as} \quad n \to \infty,$$

where $\boldsymbol{Z}$ be a random vector. Let $S_1,\,S_2,\,\ldots$ be a sequence of random variables such that

$$S_n \xrightarrow{\mathsf{P}} S \qquad \textit{as} \quad n \to \infty,$$

where $S \in \mathbb{R}$ is a real constant.

Then

(i) $S_n\,\boldsymbol{Z}_n \xrightarrow{\mathcal{D}} S\,\boldsymbol{Z} \qquad$ *as* $n \to \infty$.

(ii) $\dfrac{1}{S_n}\,\boldsymbol{Z}_n \xrightarrow{\mathcal{D}} \dfrac{1}{S}\,\boldsymbol{Z} \qquad$ *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 (2020). *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.

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.