Dr. Andreas Groll

Summer Term 2015

Lecture Notes on Different Aspects of Regression Analysis

Department of Mathematics, Workgroup Financial Mathematics, Ludwig-Maximilians-University Munich, Theresienstr. 39, 80333 Munich, Germany,*andreas.groll@math.lmu.de*

Preface

These lecture notes were written in order to support the students of the graduate course "Different Aspects of Regression Analysis" at the Mathematics Department of the Ludwig Maximilian University of Munich in their first approach to regression analysis.

Regression analysis is one of the most used statistical methods for the analysis of empirical problems in economic, social and other sciences. A variety of model classes and inference concepts exists, reaching from the classical linear regression to modern non- and semi-parametric regression. The aim of this course is to give an overview of the most important concepts of regression and to give an impression of its flexibility. Because of the limited time the different regression methods cannot be explained very detailed, but their overall ideas should become clear and potential fields of application are mentioned. For more detailed information it is referred to corresponding specialist literature whenever possible.

Contents

1

Introduction

The aim is to model characteristics of a *response variable* y that is depending on some covariates $x_1, \ldots, x_p$. Most parts of this chapter are based on Fahrmeir et al. (2007). The response variable y is often also denoted as the *dependent variable* and the covariates as *explanatory variables* or *regressors*. All models that are introduced in the following primarily differ in the different types of response variables (continuous, binary, categorial or counting variables) and the different types of covariates (also continuous, binary or categorial).

One essential characteristic of regression problems is that the relationship between the response variable y and the covariates is not given as an exact function $f(x_1, \ldots, x_p)$ of $x_1, \ldots, x_p$, but is overlain by random errors, which are random variables. Consequently, also the response variable y becomes a random variable, whose distribution is depending on the covariates.

Hence, a major objective of regression analysis is the investigation of the influence of the covariates on the mean of the response variable. In other words, we model the (conditional) expectation $E[y|x_1, \ldots, x_p]$ of y in dependency of the covariates. Thus, the expectation is a function of the covariates:

$$E[y|x_1, \ldots, x_p] = f(x_1, \ldots, x_p).$$

Then the response variable can be decomposed into

$$y = E[y|x_1, \ldots, x_p] + \varepsilon = f(x_1, \ldots, x_p) + \varepsilon,$$

where ε denotes the random variation from the mean, which is not explained by the covariates. Often, $f(x_1, \ldots, x_p)$ is denoted as the *systematic component*, the random variation ε is also denoted as *stochastic component* or *error term*. So a major objective of regression analysis is the estimation of the systematic component f from the data $y_i, x_{i1}, \ldots, x_{ip}, i = 1, \ldots, n$, and to separate it from the stochastic component ε.

1.1 Ordinary Linear Regression

The most famous class is the class of *linear regression models*

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,$$

which assumses that the function f is linear, so that

$$E[y|x_1, \ldots, x_p] = f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

holds. For the data we get the following n equations

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

with unknown *regression parameters* or *regression coefficients*, respectively, $\beta_0, \ldots, \beta_p$. Hence, in the linear model each covariate has a linear effect on y and the effects of single covariates aggregate additively. The linear regression model is particularly reasonable, if the response variable y is continuous and (ideally) approximately normally distributed.

Example 1.1.1 (Munich rent levels). Usually the average rent of a flat depends on some explanatory variables such as type, size, quality etc. of the flat, which hence is a regression problem. We use the so-called net rent as the response variable, which is the monthly rent income, after deduction of all operational and incidental costs. Alternatively, the net rent per square meter (sm) could be used as response variable.

Here, part of the data and variables of the 1999 Munich rent levels are used (compare Fahrmeir et al., 2007). More recent rent level data were either not publicly available or less suitable for illustration. The current rent levels for Munich is available at `http://mietspiegel-muenchen.de`. Table 1.1 contains the abbreviations together with a short description for selected covariates that are used later on in our analysis. The data contain information of more than 3000 flats and have been collected in a representative random sample.

In the following only the flats with a construction year of 1966 or later are investigated. The sample is separated into three parts corresponding to the three different qualities of location. Figure 1.1 shows a scatterplot of the flats with normal location for the response variable *rent* and the *size* as the explanatory variable. The scatterplot indicates an approximately linear influence of the size of the flat on the rent:

$$rent_i = \beta_0 + \beta_1 \cdot size_i + \varepsilon_i. \tag{1.1.1}$$

The error terms ε_i can be interpreted as random variations of the straight line $\beta_0 + \beta_1 \cdot size_i$. As systematic differences from zero are already captured by the parameter β_0, it is assumed that $E[\varepsilon_i] = 0$. An alternative formulation of the Model (1.1.1) is

$$E[rent|size] = \beta_0 + \beta_1 \cdot size,$$

which means that the expected rent is a linear function of the size of the flat. △

Variable	Description	mean/frequency in %
rent	net rent per month (in DM)	895.90
rentsm	net rent per month and sm (in DM)	13.87
size	living area in sm	67.37
year	year of construction	1956.31
loc	quality of location estimated by a consultant	
	1=normal location	58.21
	2=good location	39.26
	3=perfect location	2.53
bath	equipment of the bathroom	
	0=standard	93.80
	1=upscale	6.20
kit	equipment of the kitchen	
	0=standard	95.75
	1=upscale	4.25
ch	central heating	
	0=no	10.42
	1=yes	89.58
dis	district of Munich	

Table 1.1: Description of the variables of the Munich rent level data in 1999.

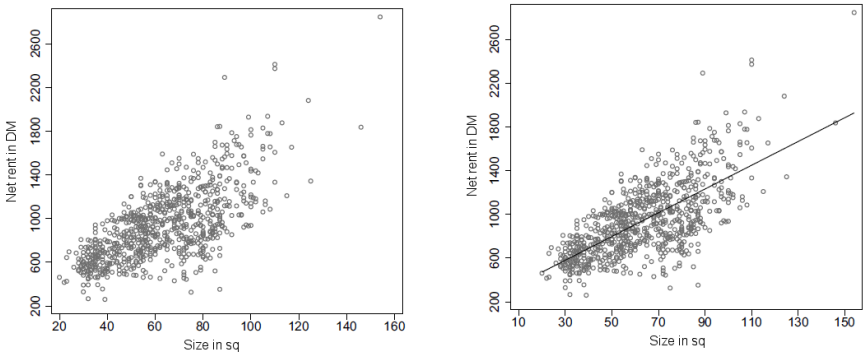

Fig. 1.1: Scatterplot between net rent and size of the flat for flats with a construction year of 1966 or later and normal location (left). Additionally, in the right figure the regression line corresponding to Model (1.1.1) is illustrated.

The example was an application of the *ordinary linear regression model*

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

or in more general form

$$y = f(x) + \varepsilon = E[y|x] + \varepsilon,$$

respectively, where the function $f(x)$ or the expectation $E[y|x]$ are assumed to be linear, $f(x) = E[y|x] = \beta_0 + \beta_1 \cdot x$.

In general, for the *standard model of ordinary linear regression* the following assumptions hold:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1.2}$$

where the error terms ε_i are independent and identically distributed (iid) with

$$E[\varepsilon_i] = 0 \quad \text{and} \quad Var(\epsilon_i) = \sigma^2, \quad \sigma > 0.$$

The property that all error terms have identical variances σ^2 is denoted as *homoscedasticity*. For the construction of confidence intervals and test statistics it is useful, if additionally (at least approximately) the *normal distribution assumption*

$$\varepsilon_i \sim N(0, \sigma^2)$$

holds. Then, also the response variables are (conditionally) normally distributed with

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i, \quad Var(y_i|x_i) = \sigma^2,$$

and are (conditionally) independent for given covariates x_i.

The unknown parameters β_0 and β_1 can be estimated according to the method of least squares (LS-method). The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by minimization of the sum of the squared distances

$$LS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

for given data $(y_i, x_i), i = 1, \ldots, n$. The method of least squares is discussed in more detail in Section 2.2.1. Putting $\hat{\beta}_0, \hat{\beta}_1$ into the linear part of the model, yields the estimated regression line $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. The regression line can be considered as an estimate $\widehat{E[y|x]}$ of the conditional expectation of y, given x, and thus can be used for the prediction of y, which is defined as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Example 1.1.2 (Munich rent levels - ordinary linear regression). We illustrate the ordinary linear regression with the data shown in Figure 1.1 using the corresponding Model (1.1.1). A glance on the data raises doubts that the assumption of identical variances $Var(\varepsilon_i) = Var(y_i|x_i) = \sigma^2$ is justified, because the variability seems to rise with increasing living area of the flat, but this is initially ignored.

Using the LS-method for the Model (1.1.1) one obtains the estimates $\hat{\beta}_0 = 253.95, \hat{\beta}_1 = 10.87$. This yields the estimated linear function

$$\hat{f}(size) = 253.95 + 10.87 \cdot size$$

in Figure 1.1 (on the right). The slope parameter $\hat{\beta}_1 = 10.87$ can be interpreted as follows: if the flat size increases by 1 sm, then the average rent increases by 10.87 DM. △

Standard Model of Ordinary Linear Regression

Data

$(y_i, x_i), i = 1, \ldots, n$, for metric variables y and x.

Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (iid) with

$$E[\varepsilon_i] = 0 \quad \text{and} \quad Var(\epsilon_i) = \sigma^2.$$

The estimated regression line $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ can be considered as an estimate $\widehat{E[y|x]}$ of the conditional expectation of y, given x, and can be used for the prediction of y, which is defined as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Note here that for the use of a linear regression model the *linear relationship in the regression coefficients* β_0 and β_1 is crucial. The covariate x - as well as the response variable y - may be transformed adequately. Common transformations are for example $g(x) = \log(x), g(x) = \sqrt{x}$ or $g(x) = 1/x$.

1.2 Multiple Linear Regression

The standard model of ordinary linear regression from Equation (1.1.2) is a special case ($p = 1$) of the *multiple linear regression model*

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

with p regressors or covariates, respectively, $x_1, \ldots, x_p$. Here, x_{ij} denotes the j-th covariate of observation i, with $i = 1, \ldots, n$. The covariates can be metric, binary or multicategorical (after suitable encoding). Similar to the ordinary linear regression model new variables can be extracted from the original ones by transformation. Also for the error terms the same assumptions are required. If the assumption of normally distributed error terms holds, then the response variables, given the covariates, are again independent and normally distributed:

$$y_i | x_{i1}, \ldots, x_{ip} \sim N(\mu_i, \sigma^2),$$

with $\mu_i = E[y_i | x_{i1}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}$.

Standard Model of Multiple Linear Regression

Data

$(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n$, for a metric variable y and metric or binary encoded categorical regressors $x_1, \ldots, x_p$.

Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n.$$

The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (iid) with

$$E[\varepsilon_i] = 0 \quad \text{and} \quad Var(\epsilon_i) = \sigma^2.$$

The estimated linear function

$$\hat{f}(x_1, \ldots, x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

can be considered as an estimate $\hat{E}[y|x_1, \ldots, x_p]$ of the conditional expectation of y, given $x_1, \ldots, x_p$, and can be used for the prediction of y, which is defined as $\hat{y}$.

The following examples illustrate how flexible the multiple linear regression model is, using suitable transformation and encoding of covariates.

Example 1.2.1 (Munich rent levels - Rents in normal and good location). We now incorporate the flats with good location and mark the data points in the scatterplot in Figure 1.2 accordingly. In addition to the regression line for flats with normal location a separately estimated regression line for flats with good location is plotted. Alternatively, one can analyze both location types jointly in a single model resulting in two regression lines that are parallel shifted. The corresponding regression model has the form

$$rent_i = \beta_0 + \beta_1 size_i + \beta_2 gloc_i + \varepsilon_i. \tag{1.2.1}$$

Here, *gloc* is a binary *indicator variable*

$$gloc_i = \begin{cases} 1 & \text{if the } i\text{-th flat has good location} \\ 0 & \text{if the } i\text{-th flat has normal location.} \end{cases}$$

Using the LS-method we obtain the estimated average rent

$$\widehat{rent} = 219.74 + 11.40 \cdot size + 111.66 \cdot gloc.$$

Due to the 1/0-encoding of the location, an equivalent representation is

$$\widehat{rent} = \begin{cases} 331.40 + 11.40 \cdot size & \text{for good location} \\ 219.74 + 11.40 \cdot size & \text{for normal location.} \end{cases}$$

Both parallel lines are shown in Figure 1.3. The regression coefficients can be interpreted as follows:

- Both in good and normal location an increase of the living area of 1 sm results in an increase of the average rent of 11.40 DM.
- For flats with the same size the average rent of flats with good location exceeds the rent of corresponding flats with normal location by 111.66 DM.

△

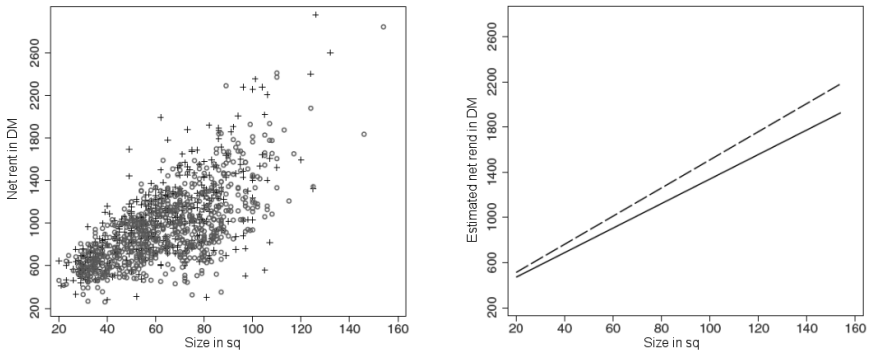

Fig. 1.2: Left: scatterplot between net rent and size for flats with normal (circles) and good (plus) location. Right: separately estimated regression lines for flats with normal (solid line) and good (dashed line) location.

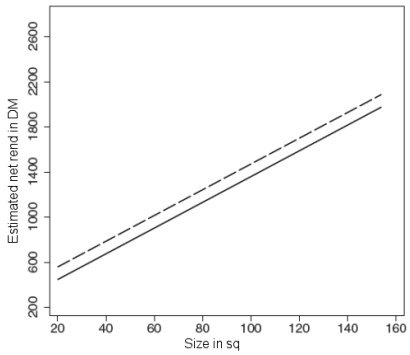

Fig. 1.3: Estimated regression lines for the Model (1.2.1) for flats with normal (solid line) and good (dashed line) location.

Example 1.2.2 (Munich rent levels - Non-linear influence of the flat size). We now use the variable *rentsm* (net rent per sm) as response variable and transform the size of the flat into $x = \frac{1}{size}$. The corresponding model is

$$rentsm_i = \beta_0 + \beta_1 \cdot \frac{1}{size_i} + \beta_2 \cdot gloc_i + \varepsilon_i. \tag{1.2.2}$$

The estimated model for average rent per sm is

$$\widehat{rentsm} = 10.74 + 262.70 \cdot \frac{1}{size_i} + 1.75 \cdot gloc.$$

Both curves for the average rent per sm

$$\widehat{rentsm} = \begin{cases} 12.49 + 262.70 \cdot \frac{1}{size_i} & \text{for good location} \\ 10.74 + 262.70 \cdot \frac{1}{size_i} & \text{for normal location.} \end{cases}$$

are illustrated in Figure 1.4. The slope parameter $\hat{\beta}_1 = 262.70$ can be interpreted as follows: if the flat size increases by one sm to $size + 1$, then the average rent is reduced to

$$\widehat{rentsm} = 10.74 + 262.70 \cdot \frac{1}{size_i + 1} + 1.75 \cdot gloc.$$

△

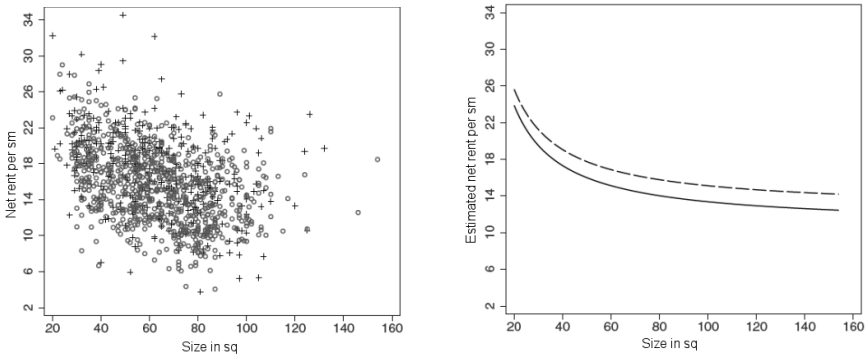

Fig. 1.4: Left: Scatterplot between net rent per sm and size for flats with normal (circles) and good (plus) location. Right: Estimated regression curves for flats with normal (solid line) and good (dashed line) location for the Model (1.2.2).

Example 1.2.3 (Munich rent levels - Interaction between flat size and location). To incorporate an interaction between the size of the flat and its location into the Model (1.2.1), we define an interaction variable *inter* by multiplication of the covariates *size* and *gloc* with values

$$inter_i = size_i \cdot gloc_i.$$

Then

$$inter_i = \begin{cases} size_i & \text{if the } i\text{-th flat has good location} \\ 0 & \text{if the } i\text{-th flat has normal location,} \end{cases}$$

and we extend the Model (1.2.1) by incorporating apart from the *main effects* of *size* and *gloc* also the *interaction effect* of the variable $inter = size \cdot gloc$:

$$rent_i = \beta_0 + \beta_1 size_i + \beta_2 gloc_i + \beta_3 inter_i + \varepsilon_i. \tag{1.2.3}$$

Due to the definition of *gloc* and *inter* we get

$$rent_i = \begin{cases} \beta_0 + \beta_1 size_i + \varepsilon_i & \text{for normal location} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) size_i + \varepsilon_i & \text{for good location.} \end{cases}$$

For $\beta_3 = 0$ no interaction effect is present and one obtains the Model (1.2.1) with parallel lines, i.e. the same slopes β_1. For $\beta_3 \neq 0$ the effect of the flat size, i.e. the slope of the straight line for flats with good location, is changed by the value β_3 compared to flats with normal location.

The LS-estimation is not done separately for both location types as in Figure 1.2 (on the right), but jointly for the data of both location types using the Model (1.2.3). We obtain

$$\hat{\beta}_0 = 253.95, \quad \hat{\beta}_1 = 10.87, \quad \hat{\beta}_2 = 10.15, \quad \hat{\beta}_3 = 1.60,$$

and both regression lines for flats with good and normal location are illustrated in Figure 1.5. At this point, it could be interesting to check, if the modeling of an interaction effect is necessary. This can be done by testing the hypothesis

$$H_0 : \beta_3 = 0 \quad \text{versus} \quad H_1 : \beta_3 \neq 0.$$

How such tests can be constructed will be illustrated during the course. △

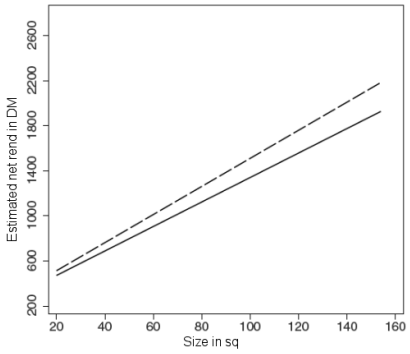

Fig. 1.5: Estimated regression lines for flats with normal (solid line) and good (dashed line) location based on the Interaction-Model (1.2.3).

2

Linear Regression Models

This section deals with the conventional linear regression model $y = \beta_0 + \beta_1 x_1 + \dots \beta_p x_p + \varepsilon$ with iid error terms. In the first major part the model is introduced and the most important properties and asymptotics of the LS-estimators are derived in Section 2.2.1 and 2.2.2. Next, the properties of residuals and predictions are investigated. Another big issue of this chapter is the conventional test- and estimation theory of the linear model, which is described in Section 2.2.5.

2.1 Repetition

This section contains a short insertion recapitulating some important properties of multivariate random variables. For some definitions and properties of multivariate normally distributed random variables, consult Appendix D.1.

Multivariate Random Variables

- Let $\mathbf{x}$ be a vector of p (univariate) random variables, i.e. $\mathbf{x} = (x_1, \dots, x_p)^\intercal$. Let $E[x_i] = \mu_i$ be the expected value of x_i, $i = 1, \dots, p$. Then we get

$$E[\mathbf{x}] = \boldsymbol{\mu}, \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\intercal,$$

  which is the vector of expectations.
- A closed representation of the variation parameters (variances and covariances) of all p random variables would be desirable. We have

$$\begin{aligned} &\text{Variance:} \quad Var(x_i) = E[\{x_i - E(x_i)\}^2] = E[\{x_i - E(x_i)\}\{x_i - E(x_i)\}] \\ &\text{Covariance:} \; Cov(x_i, x_j) = E[\{x_i - E(x_i)\}\{x_j - E(x_j)\}] \end{aligned}$$

  In general, there are p variances and $p(p-1)$ covariances, altogether $p + p(p-1) = p^2$ parameters that contain information concerning the

variation. These are summarized in the $(p \times p)$-covariance matrix (also called variance-covariance matrix):

$$\boldsymbol{\Sigma} := Cov(\mathbf{x}) = E[\{\mathbf{x} - E(\mathbf{x})\}\{\mathbf{x} - E(\mathbf{x})\}^\intercal]$$

Example $p = 2$:

$$\begin{aligned}\boldsymbol{\Sigma} &= E\left[\begin{pmatrix} x_1 - E[x_1] \\ x_2 - E[x_2] \end{pmatrix} \left(x_1 - E[x_1],\, x_2 - E[x_2] \right)\right] \\ &= E\left[\begin{pmatrix} \{x_1 - E[x_1]\}\{x_1 - E[x_1]\} & \{x_1 - E[x_1]\}\{x_2 - E[x_2]\} \\ \{x_2 - E[x_2]\}\{x_1 - E[x_1]\} & \{x_2 - E[x_2]\}\{x_2 - E[x_2])\} \end{pmatrix}\right] \\ &= \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) \\ Cov(x_2, x_1) & Var(x_2) \end{pmatrix}.\end{aligned}$$

- Properties of $\boldsymbol{\Sigma}$:
  (i) quadratic
  (ii) symmetric
  (iii) positive-semidefinite (recall: a matrix $\mathbf{A}$ is positive-semidefinite $\Leftrightarrow$ $\mathbf{x}^\intercal \mathbf{A} \mathbf{x} \geq 0,\ \forall\, \mathbf{x} \neq \mathbf{0}$)

Multivariate Normal Distribution

For more details, see Appendix D.1. The general case:

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Remark 2.1.1. For independent random variables $\boldsymbol{\Sigma}$ is a diagonal matrix.

Figures 2.1 to 2.4 show the density functions of two-dimensional normal distributions.

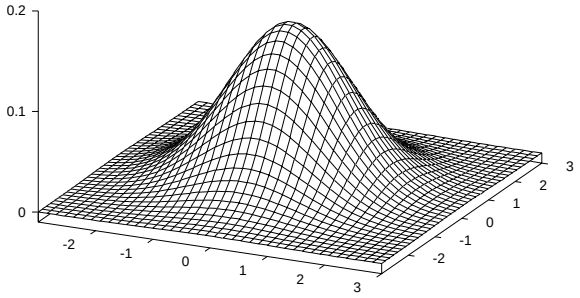

Fig. 2.1: Density function of a two-dimensional normal distribution for uncorrelated factors, $\rho = 0$, with $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1.0$

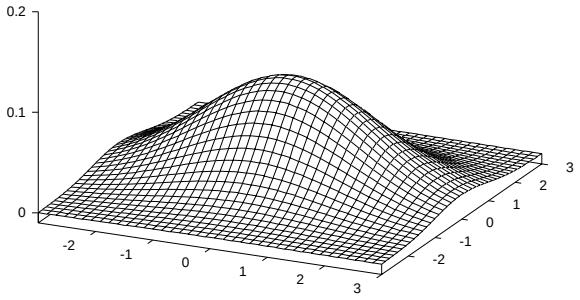

Fig. 2.2: Density function of a two-dimensional normal distribution for uncorrelated factors, $\rho = 0$, with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1.5$, $\sigma_2 = 1.0$

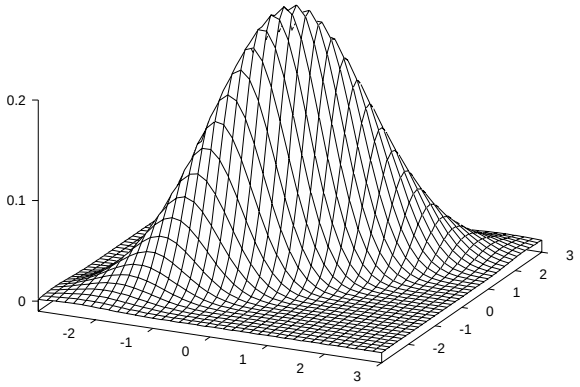

Fig. 2.3: Density function of a two-dimensional normal distribution, $\rho = 0.8$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1.0$

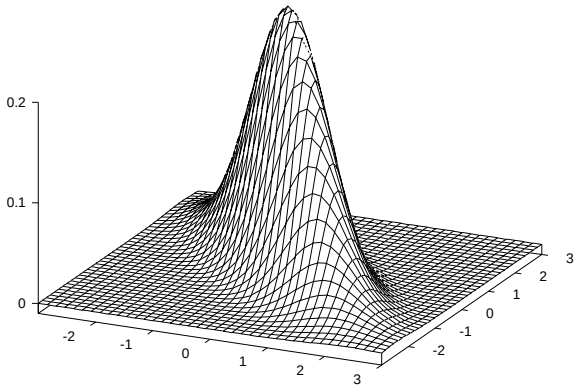

Fig. 2.4: Density function of a two-dimensional normal distribution, $\rho = -0.8$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1.0$

2.2 The Ordinary Multiple Linear Regression Model

Let data be given by y_i and $x_{i1}, \ldots, x_{ip}$. We collect the covariates and the unknown parameters in the $(p+1)$-dimensional vectors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\intercal$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\intercal$. Hence, for each observation we get the following equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2.1}$$

By definition of suitable vectors and matrices we get a compact form of our model in matrix notation. With

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

we can write the Model (2.2.1) in the simpler form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

- $\mathbf{y}$: response variable
- $\mathbf{X}$: design matrix or matrix of regressors, respectively
- $\boldsymbol{\varepsilon}$: error term
- $\boldsymbol{\beta}$: unknown vector of regression parameters

Assumptions:

- $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n)$, i.e. no systematic error, the error terms are uncorrelated and all have the same variance (homoscedasticity)
- $\mathbf{X}$ deterministic

Remark 2.2.1. Usually $p + 1 \leq n$ holds.

The Ordinary Multiple Linear Regression Model

The model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is called ordinary (multiple) linear regression model, if the following assumptions hold:

1. $E[\boldsymbol{\varepsilon}] = \mathbf{0}$.
2. $Cov(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal] = \sigma^2 \mathbf{I}_n$.
3. The design matrix $\mathbf{X}$ has full column rank, i.e. $rk(\mathbf{X}) = p + 1$.

The model is called ordinary normal regression model, if additionally the following assumption holds:

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

2.2.1 LS-Estimation

Principle of the LS-estimation: minimize the sum of squared errors

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2 = \sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon} \tag{2.2.2}$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\epsilon_i^2 = \boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.2.3}$$

Alternatively two other approaches are supposable:

(i) Minimize the sum of errors $SE(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}) = \sum_{i=1}^{n}\varepsilon_i \Longrightarrow$ Problem: positive and negative errors can eliminate each other and thus the solution of the minimization problem is usually not unique.

(ii) Minimize the sum of absolute errors $AE(\boldsymbol{\beta}) = \sum_{i=1}^{n}|y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}| = \sum_{i=1}^{n}|\varepsilon_i|$ $\Longrightarrow$ There is no analytical solving method for the computation of the solution and solving methods are more demanding (e.g. simplex-based methods or iteratively re-weighted least squares are used).

Lemma 2.2.2. *Let $\boldsymbol{B}$ be an $n \times (p+1)$ matrix. Then the matrix $\boldsymbol{B}^\intercal\boldsymbol{B}$ is symmetric and positive semi-definite. It is positive definite, if $\boldsymbol{B}$ has full column rank. Then, besides $\boldsymbol{B}^\intercal\boldsymbol{B}$, also $\boldsymbol{B}\boldsymbol{B}^\intercal$ is positive semi-definite.*

Theorem 2.2.3. *The LS-estimator of the unknown parameters $\boldsymbol{\beta}$ is*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y},$$

if $\mathbf{X}$ has full column rank $p+1$.

Proof. First, we show that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ holds.

According to the LS-approach from Equation (2.2.2) one has to minimize the following function

$$LS(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^\intercal\mathbf{y} - 2\boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{y} + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta}.$$

A necessary condition for a minimum is that the gradient is equal to a vector full of zeros. With the derivation rules for vectors and matrices from Proposition A.0.1 (see Appendix A) we obtain

$$\begin{aligned}
\frac{\partial LS(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= -2\mathbf{X}^\intercal\mathbf{y} + 2\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \stackrel{!}{=} \mathbf{0} \\
&\Leftrightarrow \quad \mathbf{X}^\intercal\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\intercal\mathbf{y} \\
&\overset{rk(\mathbf{X})=p+1}{\Leftrightarrow} \hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}.
\end{aligned}$$

A sufficient condition for a minimum is that the Hesse-matrix (the matrix of the second partial derivatives) has to be positive-semidefinite. In our case we obtain

$$\frac{\partial^2 LS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal} = 2\mathbf{X}^\intercal\mathbf{X} \geq 0, \quad \text{as } \mathbf{X}^\intercal\mathbf{X} \text{ is positive-semidefinite,}$$

so $\hat{\boldsymbol{\beta}}$ is in fact a minimum. □

On the basis of the LS-estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ for $\boldsymbol{\beta}$ we are able to estimate the (conditional) expectation of $\mathbf{y}$ by

$$\widehat{E[\mathbf{y}]} = \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Inserting the formula of the LS-estimator yields

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{H}\mathbf{y},$$

where the $n \times n$ matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal \tag{2.2.4}$$

is called *prediction-matrix* or *hat-matrix*. The following proposition summarizes its properties.

Proposition 2.2.4. *The hat-matrix $\boldsymbol{H} = (h_{ij})_{1 \leq i,j \leq n}$ has the following properties:*

(i) $\mathbf{H}$ *is symmetric.*
(ii) $\mathbf{H}$ *is idempotent (Definition: a quadratic matrix $\boldsymbol{A}$ is idempotent, if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{A}$ holds).*
(iii) $rk(\boldsymbol{H}) = tr(\boldsymbol{H}) = p + 1$. *Here, $tr(\cdot)$ denotes the trace of a matrix.*
(iv) $0 \leq h_{ii} \leq 1, \forall i = 1, \ldots, n$.
(v) the matrix $\boldsymbol{I}_n - \boldsymbol{H}$ is also symmetric and idempotent with $rk(\boldsymbol{I}_n - \boldsymbol{H}) = n - p - 1$.

Proof. → see exercises. □

Remark 2.2.5. One can even show that $\frac{1}{n} \leq h_{ii} \leq \frac{1}{r}, \forall i = 1, \ldots, n$, where r denotes the number of rows in $\mathbf{X}$ that are identical, see for example Hoaglin and Welsch (1978). Hence, if all rows are distinct, one has $\frac{1}{n} \leq h_{ii} \leq 1, \forall i = 1, \ldots, n$.

Next, we derive an estimator for σ^2. From a heuristic perspective, as $E[\varepsilon_i] = 0$, it seems plausible to use the empirical variance as an estimate: $\hat{\sigma}^2 = \widehat{Var(\varepsilon_i)} = \frac{1}{n}\boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon}$. Problem: the vector $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ of the true residuals is unknown (because the true coefficient vector $\boldsymbol{\beta}$ is unknown as well). Solution: we use

the vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ of estimated residuals. This estimate is also obtained by the following strategy.

It seems likely to estimate the variance σ^2 using maximum-likelihood (ML) estimation technique. Under the assumption of normally distributed error terms, i.e. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, we get $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ (using Proposition D.0.1 from Appendix D) and obtain the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right). \tag{2.2.5}$$

Taking the logarithm yields the log-likelihood

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.2.6}$$

Theorem 2.2.6. *The ML-estimator of the unknown parameter σ^2 is $\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}}{n}$, with $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.*

Proof. For the computation of the ML-estimator for σ^2, we have to maximize the likelihood or the log-likelihood from Equations (2.2.5) and (2.2.6), respectively. Setting the partial derivative of the log-likelihood (2.2.6) with respect to σ^2 equal to zero yields

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

Inserting the LS-estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ into the last equation yields

$$\begin{aligned} & -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ = \; & -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \hat{\mathbf{y}})^\intercal(\mathbf{y} - \hat{\mathbf{y}}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}} = 0 \end{aligned}$$

and $\sigma^2 \neq 0$, we hence obtain $\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}}{n}$. □

Yet, note that this estimator for σ^2 is only rarely used, because it is biased. This is shown in the following proposition.

Proposition 2.2.7. *For the ML-estimator $\hat{\sigma}^2_{ML}$ of σ^2 it holds that*

$$E[\hat{\sigma}^2_{ML}] = \frac{n - p - 1}{n}\sigma^2$$

.

Proof.

$$\begin{aligned} E[\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}] &= E[(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})] \\ &= E[(\mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y})^\intercal(\mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y})] \end{aligned}$$

$$
\begin{aligned}
&= E[(\mathbf{y} - \mathbf{H}\mathbf{y})^\intercal(\mathbf{y} - \mathbf{H}\mathbf{y})] \\
&= E[\mathbf{y}^\intercal(\mathbf{I}_n - \mathbf{H})^\intercal(\mathbf{I}_n - \mathbf{H})\mathbf{y}] \\
&= E[\mathbf{y}^\intercal(\mathbf{I}_n - \mathbf{H})\mathbf{y}] \\
&\overset{(*)}{=} tr((\mathbf{I}_n - \mathbf{H})\sigma^2\mathbf{I}_n) + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal(\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} \\
&= \sigma^2(n - p - 1) + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\mathbf{X}\boldsymbol{\beta} \\
&= \sigma^2(n - p - 1) + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \\
&= \sigma^2(n - p - 1) + \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \\
&= \sigma^2(n - p - 1).
\end{aligned}
$$

In $(*)$ calculation rule 6 from Theorem D.0.2 for expectation vectors and covariance matrices has been used (see Appendix D). □

Hence, immediately an unbiased estimator $\hat{\sigma}^2$ for σ^2 can be constructed by:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}}{n - p - 1}. \tag{2.2.7}$$

There also exists an alternative representation for this estimator, which is shown next.

Proposition 2.2.8. *The adjusted estimator from* (2.2.7) *of the unknown parameter* σ^2 *can also be written as*

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}}{n - p - 1} = \frac{\mathbf{y}^\intercal\mathbf{y} - \hat{\boldsymbol{\beta}}^\intercal\mathbf{X}^\intercal\mathbf{y}}{n - p - 1},$$

with $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

Proof.

$$
\begin{aligned}
\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= (\mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y})^\intercal(\mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}) \\
&= \mathbf{y}^\intercal(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)^\intercal(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\mathbf{y} \\
&= \mathbf{y}^\intercal(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\mathbf{y} \\
&= \mathbf{y}^\intercal\mathbf{y} - \underbrace{\mathbf{y}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}}_{\hat{\boldsymbol{\beta}}^\intercal}\mathbf{X}^\intercal\mathbf{y}.
\end{aligned}
$$

□

Remark 2.2.9. It can be shown that the estimator (2.2.7) maximizes the *marginal likelihood*

$$L(\sigma^2) = \int L(\boldsymbol{\beta}, \sigma^2)d\boldsymbol{\beta}$$

and is thus called a restricted maximum-likelihood (REML) estimator.

Proposition 2.2.10. *The LS-estimator* $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ *is equivalent to the ML-estimator based on maximization of the log-likelihood* (2.2.6)*.*

Proof. This follows immediately, because maximization of the term $-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is equivalent to minimization of $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, which is exactly the objective function in the LS-criterion (2.2.3).

Nevertheless, for the sake of completeness, we compute the ML-estimator for $\boldsymbol{\beta}$ by maximizing the likelihood or the log-likelihood from Equations (2.2.5) and (2.2.6), respectively. Setting the partial derivative of the log-likelihood (2.2.6) with respect to $\boldsymbol{\beta}$ equal to zero yields

$$\frac{\partial l(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}}) = -\frac{1}{\sigma^2}(\mathbf{X}^\intercal\mathbf{y} - \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta}) = 0.$$

Similar to the proof of Theorem 2.2.3 it follows:

$$\begin{aligned} & -\mathbf{X}^\intercal\mathbf{y} + \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \stackrel{!}{=} \mathbf{0} \\ \Leftrightarrow \quad & \mathbf{X}^\intercal\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\intercal\mathbf{y} \\ \overset{rk(\mathbf{X})=p+1}{\Leftrightarrow} \quad & \hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}. \end{aligned}$$

□

Parameter estimators in the multiple linear regression model

Estimator for $\boldsymbol{\beta}$

In the ordinary linear model the estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$$

minimizes the LS-criterion

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2.$$

Under the assumption of normally distributed error terms the LS-estimator is equivalent to the ML-estimator for $\boldsymbol{\beta}$.

Estimator for σ^2

The estimate

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}$$

is unbiased and can be characterized as REML-estimator for σ^2.

Proposition 2.2.11. *For the LS-estimator* $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ *and the REML-estimator* $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}$ *the following properties hold:*

(i) $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$.
(ii) $E[\hat{\sigma}^2] = \sigma^2$

Proof. (i) The LS-estimator can be represented as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\\ &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\\ &= \boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\varepsilon}\\ \text{(and also } \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\varepsilon} \text{ holds).}\end{aligned}$$

Now, we are able to derive the vector of expectations and the covariance matrix of the LS-estimator:

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal E[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

and

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}])^\intercal]\\ &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal]\\ &= E[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}]\\ &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal]\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\\ &= \sigma^2\underbrace{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}}_{=\mathbf{I}_{p+1}}(\mathbf{X}^\intercal\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.\end{aligned}$$

Altogether we obtain

$$\hat{\boldsymbol{\beta}} \sim (\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}).$$

(ii) The proof follows directly from Proposition 2.2.7. An alternative proof is based on a linear representation of $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\varepsilon}$. We have

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\mathbf{y}\\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\\ &= \underbrace{\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\underbrace{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}}_{=\mathbf{I}_{p+1}}\boldsymbol{\beta}}_{=0} + (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\boldsymbol{\varepsilon}\\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\boldsymbol{\varepsilon} \quad \text{(linear function of } \boldsymbol{\varepsilon})\\ &= \mathbf{M}\boldsymbol{\varepsilon},\end{aligned}$$

where $\mathbf{M} := (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)$ is a (deterministic) symmetric and idempotent matrix, see Proposition 2.2.4 (v). Hence, we can write

$$\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\intercal \mathbf{M}^\intercal \mathbf{M} \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\intercal \mathbf{M} \boldsymbol{\varepsilon},$$

and obtain a quadratic form in $\boldsymbol{\varepsilon}$, with other words a scalar.
With the help of the trace operator tr we obtain

$$\begin{aligned}
E[\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}] &= E[\boldsymbol{\varepsilon}^\intercal \mathbf{M} \boldsymbol{\varepsilon}] \\
&= E[tr(\boldsymbol{\varepsilon}^\intercal \mathbf{M} \boldsymbol{\varepsilon})] \quad (\boldsymbol{\varepsilon}^\intercal \mathbf{M} \boldsymbol{\varepsilon} \text{ is a scalar!}) \\
&= E[tr(\mathbf{M} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\intercal)] \quad (\text{use } tr(\mathbf{AB}) = tr(\mathbf{BA})) \\
&= tr(\mathbf{M} E[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\intercal]) \\
&= tr(\mathbf{M} \sigma^2 \mathbf{I}_n) \\
&= \sigma^2 tr(\mathbf{M}) \\
&= \sigma^2 tr(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) \\
&= \sigma^2 [tr(\mathbf{I}_n) - tr(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal)] \quad (\text{use } tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{B}) + tr(\mathbf{A})) \\
&= \sigma^2 (n - p - 1)
\end{aligned}$$

and hence $E[\hat{\sigma}^2] = E\left[\frac{\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}}{n-p-1}\right] = \sigma^2$.

□

In the following we show another very important property of the LS-estimator, namely that it is BLUE (best linear unbiased estimator). In general, a linear estimator $\hat{\boldsymbol{\beta}}^L$ has the form

$$\hat{\boldsymbol{\beta}}^L = \mathbf{b} + \mathbf{A}\mathbf{y},$$

where $\mathbf{b}$ is a $(p+1) \times 1$ vector and $\mathbf{A}$ is a matrix of dimension $(p+1) \times n$. This means that the components β_j of $\boldsymbol{\beta}$ are estimated by a linear combination of observations y_i of the response variable,

$$\hat{\beta}_j^L = b_j + a_{j1} y_1 + \ldots + a_{jn} y_n, \quad j = 0, \ldots, p.$$

Obviously the LS-estimator $\hat{\boldsymbol{\beta}}$ belongs to the class of linear estimators with the special case $\mathbf{b} = \mathbf{0}$ and $\mathbf{A} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal$. In general, for any linear estimator $\hat{\boldsymbol{\beta}}^L$ we get

$$E[\hat{\boldsymbol{\beta}}^L] = \mathbf{b} + \mathbf{A}\mathbf{X}\boldsymbol{\beta}, \qquad Cov(\hat{\boldsymbol{\beta}}^L) = \sigma^2 \mathbf{A}\mathbf{A}^\intercal.$$

Hence, linear estimators are not necessarily unbiased. The property that the LS-estimator has minimal variance is ensured by the famous Gauss-Markov-Theorem, which will be proofed next.

Theorem 2.2.12 (Gauss-Markov). *The LS-estimator is BLUE. This means that the LS-estimator has minimal variance among all linear and unbiased estimators* $\hat{\boldsymbol{\beta}}^L$*,*

$$Var(\hat{\beta}_j) \leq Var(\hat{\beta}_j^L), \quad j = 0, \ldots, p.$$

Furthermore, for an arbitrary linear combination $\boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}$ *it holds that*

$$Var(\boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}) \leq Var(\boldsymbol{c}^\intercal \hat{\boldsymbol{\beta}}^L).$$

Proof. For an unbiased linear estimator, $E[\hat{\boldsymbol{\beta}}^L] = \mathbf{b} + \mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta}$ must hold *for all* $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. For the special case $\boldsymbol{\beta} = \mathbf{0}$ it follows $\mathbf{b} = \mathbf{0}$ as a necessary condition so that $\hat{\boldsymbol{\beta}}^L$ is unbiased. Transformation of the condition $\mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta}$ yields $(\mathbf{AX} - \mathbf{I}_{p+1})\boldsymbol{\beta} = \mathbf{0}$, which then leads to the condition $\mathbf{AX} = \mathbf{I}_{p+1}$. Because $rk(\mathbf{AX}) = \min(rk(\mathbf{X}), rk(\mathbf{A})) = rk(\mathbf{I}_{p+1}) = p+1$, also $rk(\mathbf{A}) = p+1$ must hold.

Let the matrix $\mathbf{A}$ w.l.o.g. be of the form $\mathbf{A} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal + \mathbf{B}$. Inserting this into the unbiasedness-condition $\mathbf{I}_{p+1} = \mathbf{AX}$ yields

$$\mathbf{I}_{p+1} = \mathbf{AX} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X} + \mathbf{BX} = \mathbf{I}_{p+1} + \mathbf{BX},$$

and hence $\mathbf{BX} = \mathbf{0}$. Using this, for the covariance matrix of $\hat{\boldsymbol{\beta}}^L$ we get:

$$\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}^L) &= \sigma^2\mathbf{A}\mathbf{A}^\intercal \\
&= \sigma^2\{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal + \mathbf{B}\}\{\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1} + \mathbf{B}^\intercal\} \\
&= \sigma^2\left\{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{B}^\intercal + \right. \\
&\quad \left. \mathbf{BX}(\mathbf{X}^\intercal\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}^\intercal\right\} \\
&= \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1} + \sigma^2\mathbf{B}\mathbf{B}^\intercal \\
&= Cov(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{B}\mathbf{B}^\intercal
\end{aligned}$$

From Lemma 2.2.2 we know that $\mathbf{BB}^\intercal$ is positive semi-definite, so that we get

$$Cov(\hat{\boldsymbol{\beta}}^L) - Cov(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{B}\mathbf{B}^\intercal \geq \mathbf{0}. \tag{2.2.8}$$

We are now able to derive the minimization properties of the LS-estimator from this general proposition. The variances of $\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}^L$ and $\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}$ are given by

$$Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}^L) = \mathbf{c}^\intercal Cov(\hat{\boldsymbol{\beta}}^L)\mathbf{c} \quad \text{and} \quad Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}) = \mathbf{c}^\intercal Cov(\hat{\boldsymbol{\beta}})\mathbf{c}.$$

As the difference of the covariance matrices in (2.2.8) is positive semi-definite it follows directly by the definition

$$\mathbf{c}^\intercal Cov(\hat{\boldsymbol{\beta}}^L)\mathbf{c} - \mathbf{c}^\intercal Cov(\hat{\boldsymbol{\beta}})\mathbf{c} \geq 0 \quad \forall\mathbf{c},$$

and hence the proposition

$$Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}^L) \geq Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}})$$

is proofed. As the vector $\mathbf{c}$ is arbitrary, we are able to choose for each $j = 0, \ldots, p$ the corresponding vector $\mathbf{c} = (0, \ldots, 1, \ldots, 0)^\intercal$ with a single one at position $j + 1$ and obtain:

$$Var(\hat{\beta}_j^L) = Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}^L) \geq Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}) = Var(\hat{\beta}_j).$$

□

Another quantity that is important in linear regression analysis is the *coefficient of determination*. It reflects the proportion of variability in a data set that is accounted for by the statistical model and can be derived with the aid of the geometric properties of the LS-estimator.

Definition 2.2.13. *The coefficient of determination is defined by*

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.2.9}$$

and measures the proportion of variability in y that is accounted for by the statistical model from the overall variation in y.

Lemma 2.2.14. *The method of least squares yields the following geometrical results:*

(i) The fitted values $\hat{\boldsymbol{y}}$ are orthogonal to the residuals $\hat{\boldsymbol{\varepsilon}}$, i.e. $\hat{\boldsymbol{y}}^\intercal\hat{\boldsymbol{\varepsilon}} = 0$.

(ii) The columns of $\mathbf{X}$ are orthogonal to the residuals $\hat{\boldsymbol{\varepsilon}}$, i.e. $\mathbf{X}^\intercal\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$.

(iii) The residuals are zero on average, i.e.

$$\sum_{i=1}^{n}\hat{\varepsilon}_i = 0 \quad \textit{and} \quad \bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = 0.$$

(iv) The mean of the estimated values $\hat{y}_i$ is equal to the mean of the observed values y_i, i.e.

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}.$$

Proof. The first two properties can easily be derived using the hat-matrix $\mathbf{H}$ from (2.2.4):

(i)

$$\begin{aligned}\hat{\mathbf{y}}^\intercal\hat{\boldsymbol{\varepsilon}} &= \mathbf{y}^\intercal\mathbf{H}(\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{y}^\intercal\mathbf{H}\mathbf{y} - \mathbf{y}^\intercal\mathbf{H}\mathbf{H}\mathbf{y}\\ &= \mathbf{y}^\intercal\mathbf{H}\mathbf{y} - \mathbf{y}^\intercal\mathbf{H}\mathbf{y} = 0.\end{aligned}$$

(ii)

$$\begin{aligned}\mathbf{X}^\intercal\hat{\boldsymbol{\varepsilon}} &= \mathbf{X}^\intercal(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}^\intercal(\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{X}^\intercal\mathbf{y} - \mathbf{X}^\intercal\mathbf{H}\mathbf{y}\\ &= \mathbf{X}^\intercal\mathbf{y} - \mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{0}.\end{aligned}$$

(iii) Denote the first column of the design matrix by $\mathbf{x}^0$. As $\mathbf{x}^0$ is the $\mathbf{1}$-vector, using the orthogonality from (*ii*) we get:

$$0 = (\mathbf{x}^0)^\intercal\hat{\boldsymbol{\varepsilon}} = \mathbf{1}^\intercal\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n}\hat{\varepsilon}_i.$$

(iv) Using property (*iii*), we have

$$\sum_{i=1}^{n}\hat{y}_i = \sum_{i=1}^{n}(y_i - \hat{\varepsilon}_i) = \sum_{i=1}^{n}y_i - \sum_{i=1}^{n}\hat{\varepsilon}_i = \sum_{i=1}^{n}y_i.$$

□

Lemma 2.2.15. *The following decomposition holds:*

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2. \tag{2.2.10}$$

Proof. First, we define the $n\times n$-matrix

$$\mathbf{C}=\mathbf{I}_n-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal.$$

Obviously, $\mathbf{C}$ is idempotent and symmetric and has several remarkable properties. If we multiplicate $\mathbf{C}$ by an arbitrary vector $\mathbf{a}\in\mathbb{R}^n$, we obtain

$$\mathbf{C}\mathbf{a}=\begin{pmatrix}a_1-\bar{a}\\ \vdots\\ a_n-\bar{a}\end{pmatrix}. \tag{2.2.11}$$

Hence, multiplication with $\mathbf{C}$ has the effect of centering the vector $\mathbf{a}$. For the quadratic form $\mathbf{a}^\intercal\mathbf{C}\mathbf{a}$ we obtain

$$\mathbf{a}^\intercal\mathbf{C}\mathbf{a}=\sum_{i=1}^{n}(a_i-\bar{a})^2. \tag{2.2.12}$$

First of all, we multiplicate the equality $\mathbf{y}=\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}}$ by $\mathbf{C}$ from the left and get

$$\mathbf{C}\mathbf{y}=\mathbf{C}\hat{\mathbf{y}}+\mathbf{C}\hat{\boldsymbol{\varepsilon}}.$$

On the basis of (2.2.11) and property (iii) from Lemma 2.2.14 it holds that $\mathbf{C}\hat{\boldsymbol{\varepsilon}}=\hat{\boldsymbol{\varepsilon}}$ and it follows that

$$\mathbf{C}\mathbf{y}=\mathbf{C}\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}}\quad\text{or}\quad\mathbf{y}^\intercal\mathbf{C}=\hat{\mathbf{y}}^\intercal\mathbf{C}+\hat{\boldsymbol{\varepsilon}}^\intercal,$$

respectively. Using this result, we obtain

$$\begin{aligned}\mathbf{y}^\intercal\mathbf{C}\mathbf{C}\mathbf{y}&=(\hat{\mathbf{y}}^\intercal\mathbf{C}+\hat{\boldsymbol{\varepsilon}}^\intercal)(\mathbf{C}\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}})\\&=\hat{\mathbf{y}}^\intercal\mathbf{C}\mathbf{C}\hat{\mathbf{y}}+\hat{\mathbf{y}}^\intercal\mathbf{C}\hat{\boldsymbol{\varepsilon}}+\hat{\boldsymbol{\varepsilon}}^\intercal\mathbf{C}\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}\\&=\hat{\mathbf{y}}^\intercal\mathbf{C}\hat{\mathbf{y}}+\hat{\mathbf{y}}^\intercal\hat{\boldsymbol{\varepsilon}}+\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}.\end{aligned} \tag{2.2.13}$$

Using (2.2.12), for the left side of (2.2.13) we get

$$\mathbf{y}^\intercal\mathbf{C}\mathbf{C}\mathbf{y}=\mathbf{y}^\intercal\mathbf{C}\mathbf{y}=\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Regarding the property (iv) from Lemma 2.2.14, namely that $\bar{\hat{y}}=\bar{y}$, we obtain

$$\hat{\mathbf{y}}^\intercal\mathbf{C}\hat{\mathbf{y}}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

From (i) in Lemma 2.2.14 we know that $\hat{\mathbf{y}}^\intercal \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}^\intercal \hat{\mathbf{y}} = 0$ and we finally obtain

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

□

Proposition 2.2.16. *The coefficient of determination R^2 can be transformed into*

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{y} - n\bar{y}^2}{\mathbf{y}^\intercal \mathbf{y} - n\bar{y}^2}$$

Proof. The first equality follows immediately from the decomposition from Lemma 2.2.15.

For the second equality, we regard the nominator and denominator from (2.2.9) separately. Both nominator and denominator in Definition 2.2.13 can be easily transformed:

$$\begin{aligned}
\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \hat{\mathbf{y}}^\intercal \hat{\mathbf{y}} - n\bar{y}^2 \\
&= \hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{X} \hat{\boldsymbol{\beta}} - n\bar{y}^2 \\
&= \hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} - n\bar{y}^2 \\
&= \hat{\boldsymbol{\beta}}^\intercal \mathbf{X}^\intercal \mathbf{y} - n\bar{y}^2 \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \mathbf{y}^\intercal \mathbf{y} - n\bar{y}^2
\end{aligned}$$

□

Remark 2.2.17. The following further properties of R^2 hold (without proof):

(i) $0 \leq R^2 \leq 1$
(ii) R^2 increases automatically with p

Definition 2.2.18. *The corrected coefficient of determination $\bar{R}^2$ is defined by*

$$\bar{R}^2 := 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2).$$

Properties of the LS-estimator

Without assumption of a particular distribution

1. *Expectation*: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, i.e. the LS-estimator is unbiased.
2. *Covariance matrix*: $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$, in particular

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2},$$

   holds, where R_j^2 is the coefficient of determination of a regression between x_j as response variable and the remaining covariates (for more details, consult Wooldridge, 2006). An estimate of the covariance matrix is given by

$$\widehat{Cov(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\mathbf{X}^\intercal\mathbf{X})^{-1} = \frac{\hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}}{n-p-1}(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

3. *Gauss-Markov-Theorem:* The LS-estimator has minimal variance among all linear and unbiased estimators $\hat{\boldsymbol{\beta}}^L$, i.e.

$$Var(\hat{\beta}_j) \leq Var(\hat{\beta}_j^L), \quad j = 0, \ldots, p.$$

   Furthermore, for an arbitrary linear combination $\mathbf{c}^\intercal\boldsymbol{\beta}$ it holds that

$$Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}) \leq Var(\mathbf{c}^\intercal\hat{\boldsymbol{\beta}}^L).$$

Under assumption of normal distribution

1. Distribution of response:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

2. Distribution of LS-estimator:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}).$$

3. Distribution of weighted distance:

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\intercal(\mathbf{X}^\intercal\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{p+1}.$$

   (The proof of this result follows immediately from Theorem D.1.6 in Appendix D.)

2.2.2 Asymptotic Properties of the LS-Estimator

The asymptotic properties of the LS-estimator form the basis for the tests and confidence intervals that are introduced in Section 2.2.5. In order that they are valid exactly, the assumption of the normal distribution is necessary. However, some of the assertions remain still valid without the assumption of the normal distribution asymptotically or approximately, if the sample size n goes to infinity or is at least sufficiently big. For clarification we subscript our matrix model,

$$\mathbf{y}_n = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad E[\boldsymbol{\varepsilon}_n] = \mathbf{0}, \quad Cov(\boldsymbol{\varepsilon}_n) = \sigma^2\mathbf{I}_n,$$

as well as the LS-estimator $\hat{\boldsymbol{\beta}}_n$ and the variance estimator $\hat{\sigma}_n^2$ by the sample size n. Beside the basic assumptions 1-3 on page 14, for the validity of asymptotic assertions another assumption on the limiting behaviour of the design matrix $\mathbf{X}_n$ is necessary:

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}_n^\intercal\mathbf{X}_n = \mathbf{V}, \quad \mathbf{V} \text{ positive definite.} \tag{2.2.14}$$

In this case we get the following properties:

Asymptotic Properties of the LS-estimator

1. The LS-estimator $\hat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$ as well as the ML- and REML-estimators $\hat{\sigma}_n^2$ for σ^2 are consistent.
2. The LS-estimator $\hat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$ is asymptotically normally distributed:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{V}^{-1}),$$

   i.e. the difference $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$, normalized by $\sqrt{n}$, converges in distribution to the normal distribution on the right hand side.

For consistency see Definition B.2.4 and B.2.5 in Appendix B, for convergence in distribution see Definition C.0.1 in Appendix C.

Hence, for sufficiently large n it follows that $\hat{\boldsymbol{\beta}}_n$ is approximately normally distributed with

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} N(\boldsymbol{\beta}, \sigma^2\mathbf{V}^{-1}/n).$$

If one replaces σ^2 by its consistent estimate $\hat{\sigma}_n^2$ and $\mathbf{V}$ by the approximation $\mathbf{V} \overset{a}{\approx} 1/n\mathbf{X}_n^\intercal\mathbf{X}_n$, one obtains

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} N(\boldsymbol{\beta}, \hat{\sigma}_n^2(\mathbf{X}_n^\intercal\mathbf{X}_n)^{-1}),$$

which means that, even without the assumption of the normal distribution for $\boldsymbol{\varepsilon}$, the LS-estimator has asymptotically the same normal distribution as

if the assumption of the normal distribution for $\boldsymbol{\varepsilon}$ holds, provided that assumption (2.2.14) holds. This is particularly true, if the observed covariates $\mathbf{x}_i, i = 1, \ldots, n$ are themselves realisations of iid random covariates $\mathbf{x} = (\mathbf{1}, x_1, \ldots, x_p)^\intercal$, i.e. if $(y_i, \mathbf{x}_i)$ result from a random sample of $(y, \mathbf{x})$. This requirement is met in many empirical studies, such as for example in our application on the Munich rent levels. Here, the law of large numbers tells us

$$\frac{1}{n}\mathbf{X}_n^\intercal \mathbf{X}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\intercal \to E[\mathbf{x}\mathbf{x}^\intercal] =: \mathbf{V}.$$

2.2.3 Properties of the Residuals

In this section we investigate the statistical properties of the residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}$.

Proposition 2.2.19. *(i) Similar to the error terms, also the residuals have expectation zero.*
(ii) In contrast to the error terms, the residuals are not *uncorrelated.*

Proof. (i) and (ii):

With the hat matrix $\mathbf{H}$ from (2.2.4) the residuals can be expressed as

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y},$$

and we obtain the expectation

$$E[\hat{\boldsymbol{\varepsilon}}] = E[\mathbf{y}] - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

and the covariance matrix

$$Cov(\hat{\boldsymbol{\varepsilon}}) = Cov((\mathbf{I}_n - \mathbf{H})\mathbf{y}) = (\mathbf{I}_n - \mathbf{H})\sigma^2 \mathbf{I}_n (\mathbf{I}_n - \mathbf{H})^\intercal = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

For the derivation of the covariance matrix we used calculation rule 5 from Theorem D.0.2 (Appendix D) and that the matrix $\mathbf{I}_n - \mathbf{H}$ is symmetric and idempotent. In particular, for the variances of the residuals we obtain

$$Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}),$$

where h_{ii} is the i-th diagonal element of the hat matrix $\mathbf{H}$. □

Remark 2.2.20. In contrast to the error terms, the residuals have *heteroscedastic variances*. Due to $0 \leq h_{ii} \leq 1$ from Proposition 2.2.4 (iv) the variance of the i-th residual is the smaller, the closer h_{ii} is to one.

If we additionally assume that the error terms are normally distributed, we are able to derive the distribution of the residuals and obtain

$$\hat{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H})). \tag{2.2.15}$$

As $rk(\mathbf{H}) = p+1 \leq n$, this distribution is a singular normal distribution, compare Section D.1.1 from Appendix D. Using (2.2.15), the following proposition can be shown:

Proposition 2.2.21. *Beside the usual assumptions, additionally assume that the error terms are normally distributed. Then the following properties hold:*

(i) The distribution of the squared sum of residuals is given by:

$$\frac{\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}}{\sigma^2} = (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

(ii) The squared sum of residuals $\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}$ and the LS-estimator $\hat{\boldsymbol{\beta}}$ are independent.

Proof. $\rightarrow$ see exercises. □

Properties of the residuals

Without assumption of a particular distribution

1. *Expectation*: $E[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}$, i.e. the residuals should be zero on average.
2. *Variances*: It holds

   $$Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}),$$

   i.e. the residuals have heteroscedastic variances (in contrast to the error terms ε_i).
3. *Covariance matrix*:

   $$Cov(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal),$$

   i.e. the residuals are correlated (in contrast to the error terms ε_i).

Under assumption of normal distribution

1. Distribution of residuals:

   $$\hat{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H})).$$

2. Distribution of squared sum of residuals:

   $$\frac{\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}}{\sigma^2} = (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

3. *Independence*: the squared sum of residuals $\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}$ and the LS-estimator $\hat{\boldsymbol{\beta}}$ are independent.

2.2.4 Prediction

Let $\mathbf{y}_0$ denote the T_0-dimensional response vector that is to be predicted and $\mathbf{X}_0$ the corresponding matrix of covariates. Then we have

$$\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$$

with corresponding estimate:

$$\hat{\mathbf{y}}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}}$$

For the prediction error, which is defined as $\mathbf{y}_0 - \hat{\mathbf{y}}_0$, the following properties hold:

Proposition 2.2.22. *(i)* $E[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = \mathbf{0}$ *(The expected prediction error is zero)*
(ii) $E[(\hat{\mathbf{y}}_0 - \mathbf{y}_0)(\hat{\mathbf{y}}_0 - \mathbf{y}_0)^\intercal] = \sigma^2(\mathbf{X}_0(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}_0^\intercal + \mathbf{I}_{T_0})$ *(Prediction error covariance matrix)*

Proof. (i) Remember that $\hat{\boldsymbol{\beta}} \sim (\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1})$ holds and hence, using Proposition D.0.1 from Appendix D,

$$\hat{\mathbf{y}}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}} \sim (\mathbf{X}_0\boldsymbol{\beta}, \sigma^2\mathbf{X}_0(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}_0^\intercal).$$

The true value is given by $\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$. For the prediction error $\hat{\mathbf{y}}_0 - \mathbf{y}_0$ one obtains

$$\begin{aligned} E[\hat{\mathbf{y}}_0 - \mathbf{y}_0] &= E[\mathbf{X}_0\hat{\boldsymbol{\beta}} - \mathbf{X}_0\boldsymbol{\beta} - \boldsymbol{\varepsilon}_0] \\ &= E[\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0] \\ &= \mathbf{X}_0 E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] - E[\boldsymbol{\varepsilon}_0] = \mathbf{0}. \end{aligned}$$

(ii) For the prediction error variance one obtains

$$\begin{aligned} E[(\hat{\mathbf{y}}_0 - \mathbf{y}_0)^\intercal(\hat{\mathbf{y}}_0 - \mathbf{y}_0)] &= E[(\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0)(\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0)^\intercal] \\ &= \mathbf{X}_0 E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal]\mathbf{X}_0^\intercal + E[\boldsymbol{\varepsilon}_0\boldsymbol{\varepsilon}_0^\intercal] \\ &\quad -\mathbf{X}_0 E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\varepsilon}_0^\intercal] - E[\boldsymbol{\varepsilon}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal]\mathbf{X}_0^\intercal \\ &= \sigma^2(\mathbf{X}_0(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}_0^\intercal + \mathbf{I}_{T_0}). \end{aligned}$$

In the last equality we have used the fact that $\boldsymbol{\varepsilon}_0$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ are independent. The reason for this is that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ includes only the first $1, \ldots, n$ observations, whereas $\boldsymbol{\varepsilon}_0$ corresponds to new observations.

□

2.2.5 Hypotheses Testing and Confidence Intervals

Before we are able to introduce suitable statistical tests and corresponding confidence intervals, we need some preparations:

- The basis for hypotheses testing and the derivation of confidence intervals is the additional assumption of a normal distribution for the error terms: $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I}_n)$. With Proposition D.0.1 from Appendix D it follows that

$$\begin{aligned} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} &\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} &\sim N_{p+1}((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-1}) \\ &= N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}). \end{aligned}$$

  Analogously, for the i-th component of the $\hat{\boldsymbol{\beta}}$-vector we get

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}_{i+1,i+1})$$

  or

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}_{i+1,i+1}}} \sim N(0,1) \quad (\Longrightarrow \text{Gauss-test}),$$

  respectively.
- Consider the following relation with row vector $\mathbf{R}_1 := (r_0, r_1, \ldots, r_p)$:

$$\mathbf{R}_1 \boldsymbol{\beta} = r.$$

  Hence, $\mathbf{R}_1 \boldsymbol{\beta}$ is a linear combination of the regression parameters. Using Proposition D.0.1 from Appendix D again, one obtains

$$\mathbf{R}_1 \hat{\boldsymbol{\beta}} \sim N(\underbrace{\mathbf{R}_1 \boldsymbol{\beta}}_{(1\times 1)}, \underbrace{\sigma^2 \mathbf{R}_1 (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{R}_1^\intercal}_{(1\times 1)}) \quad \Longleftrightarrow \quad \frac{\mathbf{R}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma \sqrt{\mathbf{R}_1 (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{R}_1^\intercal}} \sim N(0,1).$$

- Problem: σ^2 is unknown. Therefore, we have to use $\hat{\sigma}^2$. But as $\hat{\sigma}^2$ is a random variable (it depends on the estimated residuals, which themselves are dependent on the LS-estimator $\hat{\boldsymbol{\beta}}$, which in turn depends on the random variable $\mathbf{y}$!), the question of the corresponding distribution arises.
- Approach: we have seen in the proof of Proposition 2.2.11 that

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\intercal \hat{\boldsymbol{\varepsilon}}}{n-p-1} = \frac{\boldsymbol{\varepsilon}^\intercal \mathbf{M} \boldsymbol{\varepsilon}}{n-p-1}$$

  holds, with $\mathbf{M} := (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal)$. Our new model assumption $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ can also be written as

$$\frac{\boldsymbol{\varepsilon}}{\sigma} \sim N_n(\mathbf{0}, \mathbf{I}_n) \Leftrightarrow \frac{\epsilon_i}{\sigma} \sim N(0,1) \quad \text{for all } i = 1, \ldots, n.$$

Remember (Lemma B.1.6, Appendix B) that the square of a standard normally distributed random variable is χ^2-distributed and the sum of n squared independent standard normally distributed random variables is χ^2-distributed with n degrees of freedom. So we obtain

$$\frac{\epsilon_i^2}{\sigma^2} \sim \chi_1^2 \quad \Longrightarrow \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 = \frac{\boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi_n^2.$$

The following condition holds (without proof; see Theorem D.1.6 from Appendix D): if $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\mathbf{M}$ is symmetric and idempotent with $rk(\mathbf{M}) = n - p - 1$, then

$$\frac{\boldsymbol{\varepsilon}^\intercal\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi_{n-p-1}^2$$

holds. Hence, we get

$$\hat{\sigma}^2 = \frac{\sigma^2}{n-p-1}\frac{\boldsymbol{\varepsilon}^\intercal\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2}$$

and

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p-1}\chi_{n-p-1}^2 \quad \text{as well as} \quad \frac{\hat{\sigma}^2(n-p-1)}{\sigma^2} \sim \chi_{n-p-1}^2.$$

For the t-distribution from Definition B.1.7 (Appendix B) the following important result can be shown (without proof):

Lemma 2.2.23.
Let $Z \sim N(0,1)$ and $X \sim \chi_k^2$ be independent random variables. Then the random variable

$$T := \frac{Z}{\sqrt{\frac{X}{k}}}$$

is t-distributed with k degrees of freedom.

Putting all results together and using the definition of the t-distribution from Lemma 2.2.23, we are now able to define a suitable test-statistic with known distribution:

$$\frac{\mathbf{R}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{R}_1^\intercal}} = \frac{\frac{\mathbf{R}_1(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma\sqrt{\mathbf{R}_1(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{R}_1^\intercal}}}{\sqrt{\frac{\hat{\sigma}^2(n-p-1)}{\sigma^2(n-p-1)}}} \sim t_{n-p-1}. \tag{2.2.16}$$

Based on this test-statistic (and with suitable choice of row vector $\mathbf{R}_1$), we can test a variety of hypotheses. The following special cases are most common:

1. *Test for significance of one single covariate:*

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

2. *Test for equality:*

$$H_0 : \beta_j - \beta_l = 0 \quad \text{vs.} \quad H_1 : \beta_j - \beta_l \neq 0$$

In general, the following hypotheses can be tested:

(i) $H_0 : \mathbf{R}_1\boldsymbol{\beta} = r \quad \text{vs.} \quad H_1 : \mathbf{R}_1\boldsymbol{\beta} \neq r,$
(ii) $H_0 : \mathbf{R}_1\boldsymbol{\beta} \geq r \quad \text{vs.} \quad H_1 : \mathbf{R}_1\boldsymbol{\beta} < r,$
(iii) $H_0 : \mathbf{R}_1\boldsymbol{\beta} \leq r \quad \text{vs.} \quad H_1 : \mathbf{R}_1\boldsymbol{\beta} > r.$

Under H_0,

$$\mathbf{R}_1\hat{\boldsymbol{\beta}} \overset{H_0}{\sim} N(r, \sigma^2\mathbf{R}_1(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{R}_1^\intercal)$$

holds. For unknown σ^2, a reasonable test-statistic is

$$T = \frac{\mathbf{R}_1\hat{\boldsymbol{\beta}} - r}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{R}_1^\intercal}} \sim t_{n-p-1}.$$

The corresponding rejection areas are

(i) $|T| > t_{1-\alpha/2,n-p-1},$
(ii) $T < -t_{1-\alpha,n-p-1},$
(iii) $T > t_{1-\alpha,n-p-1}.$

Based on (2.2.16), also $(1-\alpha)$-confidence intervals for $\mathbf{R}_1\boldsymbol{\beta}$ are obtained:

$$\mathbf{R}_1\hat{\boldsymbol{\beta}} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{R}_1^\intercal}.$$

Note that for the special case $\mathbf{R}_1 = (0, \ldots, 0, 1, 0, \ldots, 0)$ with only a single one at the $(i+1)$-th position, we obtain $\mathbf{R}_1\boldsymbol{\beta} = \beta_i$ and the test-statistic from (2.2.16) has the simpler form

$$\frac{\mathbf{R}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{R}_1}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{a_{i+1,i+1}}},$$

with $a_{i+1,i+1} = (\mathbf{X}^\intercal\mathbf{X})^{-1}_{i+1,i+1}$ denoting the $(i+1)$-th diagonal element.

In a similar way, we can derive an $(1-\alpha)$-confidence interval for the prediction y_0. Let y_0 denote the (unknown) prediction value of the response variable and let $\mathbf{x}_0$ be the vector of (known) realizations of the covariates at the time of prediction. Under the assumption of normally distributed error terms and with Proposition D.0.1 from Appendix D we obtain

$$\hat{y}_0 = \mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0^\intercal\boldsymbol{\beta}, \sigma^2(\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0)) \quad \text{and} \quad y_0 \sim N(\mathbf{x}_0^\intercal\boldsymbol{\beta}, \sigma^2).$$

Putting both results together and using the fact that $\hat{y}_0$ and y_0 are independent, this yields

$$\hat{y}_0 - y_0 \sim N(0, \sigma^2(\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0 + 1)) \quad \Longleftrightarrow \quad \frac{\hat{y}_0 - y_0}{\sigma\sqrt{\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0 + 1}} \sim N(0,1).$$

If the variance of the prediction error is estimated by means of $\hat{\sigma}^2$, one obtains

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0 + 1}} = \frac{\frac{\hat{y}_0 - y_0}{\sigma\sqrt{\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0+1}}}{\sqrt{\frac{\hat{\sigma}^2(n-p-1)}{\sigma^2(n-p-1)}}} \sim t_{n-p-1}.$$

The corresponding confidence interval for y_0 is:

$$\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0 + 1}.$$

Another common test situation, which is not already covered by the test-statistic (2.2.16), is the test of a subvector $\boldsymbol{\beta}_s = (\beta_{i_1}, \ldots, \beta_{i_r})^\intercal$, with $i_j \in \{0, \ldots, p\}$ for $j = 1, \ldots, r$, $r \leq p + 1$, and corresponding hypotheses

$$H_0 : \boldsymbol{\beta}_s = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_s \neq \mathbf{0}\,.$$

For this test situation we need a more general formulation. Note that all test-situations regarded so far are special cases of tests for general *linear hypotheses* of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad \text{vs.} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}. \tag{2.2.17}$$

Here, $\mathbf{C}$ denotes a $r \times (p + 1)$ matrix with $rk(\mathbf{C}) = r \leq p + 1$. A suitable test-statistic for this more general test problem can be derived as follows:

1. Compute the sum of squared errors $\text{SSE} = \hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}}$ of the full model
2. Compute the sum of squared errors $\text{SSE}_{H_0} = \hat{\boldsymbol{\varepsilon}}_{H_0}^\intercal\hat{\boldsymbol{\varepsilon}}_{H_0}$ of the model corresponding to the null hypothesis, i.e. when the restriction $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ is fulfilled.
3. Base the test-statistic on the relative difference between the sum of squared errors corresponding to the restricted and to the full model, i.e.

$$\frac{\Delta\text{SSE}}{\text{SSE}} = \frac{\text{SSE}_{H_0} - \text{SSE}}{\text{SSE}}.$$

In order to guarantee that the distribution of the test-statistic can be easily derived, one uses the test-statistic

$$F = \frac{\frac{1}{r}\Delta\text{SSE}}{\frac{1}{n-p-1}\text{SSE}} = \frac{n-p-1}{r}\frac{\Delta\text{SSE}}{\text{SSE}}, \tag{2.2.18}$$

with r denoting the number of restrictions, i.e. the number of rows in $\mathbf{C}$.

Proposition 2.2.24. *The following properties are essential for the derivation of the distribution of the test-statistic F from* (2.2.18)*:*

(i) Calculation of the LS-estimator under H_0:
The restricted LS-estimator $\hat{\boldsymbol{\beta}}^R$ under H_0, i.e. under $\boldsymbol{C\beta} = \boldsymbol{d}$, yields:

$$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - (\mathbf{X}^\intercal\mathbf{X})^{-1}\boldsymbol{C}^\intercal(\boldsymbol{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\boldsymbol{C}^\intercal)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

(ii) Determination of the difference of the squared sum of residuals:

$$\Delta SSE = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})^{\intercal}(\boldsymbol{C}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{C}^{\intercal})^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}).$$

(iii) Stochastic properties of the difference of the squared sum of residuals:

a) Whether the restriction holds or not, we have:

$$E[\Delta SSE] = r\sigma^2 + (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d})^{\intercal}(\boldsymbol{C}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{C}^{\intercal})^{-1}(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{d}).$$

b) Under H_0 we have: $\frac{\Delta SSE}{\sigma^2} \sim \chi^2_r$.
c) ΔSSE and SSE are independent.

Proof. (i) We want to minimize the squared sum of residuals under the side condition $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$. We can apply the *Lagrange approach*:

$$\begin{aligned} LSR(\boldsymbol{\beta};\boldsymbol{\lambda}) &= LS(\boldsymbol{\beta}) - 2\boldsymbol{\lambda}^{\intercal}(\mathbf{C}\boldsymbol{\beta} - \mathbf{d}) \\ &= \mathbf{y}^{\intercal}\mathbf{y} - 2\mathbf{y}^{\intercal}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\lambda}^{\intercal}\mathbf{C}\boldsymbol{\beta} + 2\boldsymbol{\lambda}^{\intercal}\mathbf{d}. \end{aligned}$$

Here $\boldsymbol{\lambda}$ is a column vector of Lagrange-multiplicators of dimension r. Using the derivation rules from Proposition A.0.1 (Appendix A), we obtain:

$$\begin{aligned} \frac{\partial LSR(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^{\intercal}\mathbf{y} + 2\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{C}^{\intercal}\boldsymbol{\lambda} \\ \frac{\partial LSR(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= -2\mathbf{C}\boldsymbol{\beta} + 2\mathbf{d}. \end{aligned}$$

Setting both equations equal to zero yields:

$$\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^{\intercal}\mathbf{y} = \mathbf{C}^{\intercal}\boldsymbol{\lambda} \tag{2.2.19}$$

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{d}. \tag{2.2.20}$$

First, we solve Equation (2.2.19) with respect to $\boldsymbol{\lambda}$, then insert Equation (2.2.20) and finally solve the result with respect to $\boldsymbol{\beta}$. Multiplication of Equation (2.2.19) with $(\mathbf{X}^{\intercal}\mathbf{X})^{-1}$ from the left yields:

$$\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{C}^{\intercal}\boldsymbol{\lambda}.$$

Multiplication of this equation with $\mathbf{C}$ from the left yields:

$$\mathbf{C}\boldsymbol{\beta} - \mathbf{C}\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{C}^{\intercal}\boldsymbol{\lambda}.$$

Inserting the second equation $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ (2.2.20) from above, we obtain

$$\mathbf{d} - \mathbf{C}\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{C}^{\intercal}\boldsymbol{\lambda}$$

and hence

$$\boldsymbol{\lambda} = (\mathbf{C}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{C}^{\intercal})^{-1}(\mathbf{d} - \mathbf{C}\hat{\boldsymbol{\beta}}).$$

Here we used that the matrix $\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal$ is positive definite, and hence invertible, due to Proposition A.0.2 (Appendix A).

Inserting $\boldsymbol{\lambda}$ into Equation (2.2.19) yields

$$\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\intercal\mathbf{y} = \mathbf{C}^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}(\mathbf{d} - \mathbf{C}\hat{\boldsymbol{\beta}})$$

and finally, multiplication with $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ from the left yields the LS-estimator

$$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

(ii) First, for the restricted LS-estimator $\hat{\boldsymbol{\beta}}^R$ we write

$$\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}} - \Delta_{H_0},$$

with

$$\Delta_{H_0} := (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

Consequently, under the restriction (i.e. under H_0), for the estimated values $\hat{\mathbf{y}}_{H_0}$ we obtain

$$\hat{\mathbf{y}}_{H_0} = \mathbf{X}\hat{\boldsymbol{\beta}}^R = \mathbf{X}(\hat{\boldsymbol{\beta}} - \Delta_{H_0}) = \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\Delta_{H_0} = \hat{\mathbf{y}} - \mathbf{X}\Delta_{H_0},$$

and for the residuals $\hat{\boldsymbol{\varepsilon}}_{H_0}$ under H_0 we get

$$\hat{\boldsymbol{\varepsilon}}_{H_0} = \mathbf{y} - \hat{\mathbf{y}}_{H_0} = \mathbf{y} - \hat{\mathbf{y}} + \mathbf{X}\Delta_{H_0} = \hat{\boldsymbol{\varepsilon}} + \mathbf{X}\Delta_{H_0}.$$

Using this, under H_0 the squared sum of residuals can be calculated as

$$\begin{aligned}
\mathrm{SSE}_{H_0} &= \hat{\boldsymbol{\varepsilon}}_{H_0}^\intercal\hat{\boldsymbol{\varepsilon}}_{H_0} \\
&= (\hat{\boldsymbol{\varepsilon}} + \mathbf{X}\Delta_{H_0})^\intercal(\hat{\boldsymbol{\varepsilon}} + \mathbf{X}\Delta_{H_0}) \\
&= \hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}} + \hat{\boldsymbol{\varepsilon}}^\intercal\mathbf{X}\Delta_{H_0} + \Delta_{H_0}^\intercal\mathbf{X}^\intercal\hat{\boldsymbol{\varepsilon}} + \Delta_{H_0}^\intercal\mathbf{X}^\intercal\mathbf{X}\Delta_{H_0} \\
&= \hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}} + \Delta_{H_0}^\intercal\mathbf{X}^\intercal\mathbf{X}\Delta_{H_0}.
\end{aligned}$$

In the second to last equation we used that the residuals corresponding to the full model and the columns of the design matrix are orthogonal, i.e. $\mathbf{X}^\intercal\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$, see property (ii) in Lemma 2.2.14.
As $\mathbf{X}^\intercal\mathbf{X}$ is positive definite, consequently $\Delta_{H_0}^\intercal\mathbf{X}^\intercal\mathbf{X}\Delta_{H_0}$ is positive and hence, by the way we have shown that under H_0, the squared sum of residuals SSE_{H_0} is always larger than or equal to the squared sum of residuals SSE of the unrestricted model. Finally, for the difference of the squared sum of residuals ΔSSE we obtain:

$$\begin{aligned}
\Delta\mathrm{SSE} &= \hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}} + \Delta_{H_0}^\intercal\mathbf{X}^\intercal\mathbf{X}\Delta_{H_0} - \hat{\boldsymbol{\varepsilon}}^\intercal\hat{\boldsymbol{\varepsilon}} \\
&= \left\{(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal(\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})\right\}^\intercal\mathbf{X}^\intercal\mathbf{X}
\end{aligned}$$

$$\cdot \left\{ (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \right\}$$
$$= (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1}$$
$$\cdot \mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})$$
$$= (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

(iii) a) It holds that

$$E[\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}] = \mathbf{C}\boldsymbol{\beta} - \mathbf{d}$$

and

$$Cov(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) = \sigma^2 \mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal.$$

Applying property 6 from Theorem D.0.2 (Appendix D) on the random vector $\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}$ and on $\mathbf{A} = (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1}$, one obtains

$$\begin{aligned} E[\Delta\text{SSE}] &= E\left[(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \right] \\ &= tr\left\{ \sigma^2 (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} \mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal \right\} + \\ &\quad (\mathbf{C}\boldsymbol{\beta} - \mathbf{d})^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\boldsymbol{\beta} - \mathbf{d}) \\ &= tr(\sigma^2 \mathbf{I}_r) + (\mathbf{C}\boldsymbol{\beta} - \mathbf{d})^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\boldsymbol{\beta} - \mathbf{d}) \\ &= r\sigma^2 + (\mathbf{C}\boldsymbol{\beta} - \mathbf{d})^\intercal (\mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal)^{-1} (\mathbf{C}\boldsymbol{\beta} - \mathbf{d}). \end{aligned}$$

b) We define the random vector $\mathbf{z} = \mathbf{C}\hat{\boldsymbol{\beta}}$. Under H_0 we have

$$E[\mathbf{z}] = \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$$

and

$$Cov(\mathbf{z}) = \sigma^2 \mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal.$$

As $\hat{\boldsymbol{\beta}}$ is normally distributed, it follows

$$\mathbf{z} \sim N(\mathbf{d}, \sigma^2 \mathbf{C}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{C}^\intercal).$$

We can now directly apply part (i) of Theorem D.1.6 (Appendix D) on the random vector $\mathbf{z}$ and obtain the assertion.

c) The difference of the squared sum of residuals is only depending on the LS-estimator $\hat{\boldsymbol{\beta}}$. Hence, the assertion follows directly from part (ii) of Proposition 2.2.21.

□

We know already from part (i) of Proposition 2.2.21 that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Hence, combining this fact and the results of Proposition 2.2.24, with Definition B.1.9 (Appendix B) it follows that under H_0 the test-statistic F

from (2.2.18) is F-distributed with r and $n-p-1$ degrees of freedom, i.e. $F \sim F_{r,n-p-1}$. Thus, a reasonable test can be constructed. For a given significance level α the null-hypothesis is rejected, if the test-statistic exceeds the $(1-\alpha)$-quantile of the corresponding F-distribution, i.e. if

$$F > F_{r,n-p-1,1-\alpha}.$$

Proposition 2.2.25 (Overall-F-test). *For the special case of the null-hypothesis*

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

the test-statistic from (2.2.18) *has the following form:*

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p}.$$

Proof. The null-hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0,$$

which is equivalent to the case that none of the covariates contributes a significant explanation with respect to the response variable y, is tested against

$$H_1 : \beta_j \neq 0 \quad \text{for at least one } j \in \{1, \ldots, p\}.$$

Under H_0, the LS-estimation consists of a single estimation for β_0 with $\hat{\beta}_0 = \bar{y}$. Consequently, we obtain for the squared sum of residuals SSE_{H_0} under the null-hypothesis

$$\mathrm{SSE}_{H_0} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

For the difference between the squared sum of residuals under H_0, SSE_{H_0}, and the squared sum of residuals corresponding to the full model, using the decomposition of the variation (2.2.10) from Lemma 2.2.15, we obtain

$$\Delta\mathrm{SSE} = \mathrm{SSE}_{H_0} - \mathrm{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Hence, for the test-statistic F from (2.2.18) it follows

$$\begin{aligned} F &= \frac{n-p-1}{p}\frac{\Delta\mathrm{SSE}}{\mathrm{SSE}} \\ &= \frac{n-p-1}{p}\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2} \\ &= \frac{n-p-1}{p}\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} \end{aligned}$$

$$= \frac{n-p-1}{p} \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2}{1 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$= \frac{n-p-1}{p} \frac{R^2}{1-R^2}.$$

□

Remark 2.2.26. The test-statistic F of the overall-F-test from Proposition 2.2.25 has the following interesting interpretation: for a small coefficient of determination R^2 the null-hypothesis "no functional relationship" is rather maintained (as F is small), in comparison to the case of a coefficient of determination that is close to one (in this case F is comparatively big).

Testing linear hypotheses

Hypotheses

1. General linear hypotheses

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad \text{vs.} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d},$$

   where $\mathbf{C}$ is a $r \times (p+1)$ matrix with $rk(\mathbf{C}) = r \leq p+1$ (r linear independent restrictions).
2. Significance test for a single influence variable (t-test):

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

3. Test of a subvector:

$$H_0 : \boldsymbol{\beta}_s = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_s \neq \mathbf{0}.$$

4. "No functional relationship":

$$\begin{aligned} H_0 &: \ \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{vs.} \\ H_1 &: \ \beta_j \neq 0 \quad \text{for at least one } j \in \{1, \ldots, p\}. \end{aligned}$$

Test-statistics

Under H_0 and with normally distributed error terms it holds that:

1. $F = 1/r(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\intercal(\hat{\sigma}^2\mathbf{C}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{C}^\intercal)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \sim F_{r,n-p-1}$.
2. $t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{a_{j+1,j+1}}} \sim t_{n-p-1}$, with $a_{j+1,j+1} = (\mathbf{X}^\intercal\mathbf{X})^{-1}_{j+1,j+1}$ denoting the $(j+1)$-th diagonal element.
3. $F = \frac{1}{r}\hat{\boldsymbol{\beta}}_s^\intercal \widehat{Cov(\hat{\boldsymbol{\beta}}_s)}^{-1} \hat{\boldsymbol{\beta}}_s \sim F_{r,n-p-1}$.
4. $F = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \sim F_{p,n-p-1}$.

Test decisions

H_0 is rejected, if

1. $F > F_{r,n-p-1,1-\alpha}$.
2. $|t| > t_{n-p-1,1-\alpha/2}$.
3. $F > F_{r,n-p-1,1-\alpha}$.
4. $F > F_{p,n-p-1,1-\alpha}$.

These tests are quite robust with respect to small discrepancies from the normal distribution. Furthermore, for big sample sizes the tests are also applicable, if the error terms do not follow a normal distribution.

Confidence and prediction intervals

Under the condition of (at least approximately) normally distributed error terms or sufficiently large sample sizes, respectively, we obtain the following confidence and prediction intervals:

Confidence interval for β_j

A confidence interval for β_j for the significance level $1-\alpha$ is given by

$$[\hat{\beta}_j - t_{n-p-1,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{a_{j+1,j+1}}, \hat{\beta}_j + t_{n-p-1,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{a_{j+1,j+1}}],$$

with $a_{j+1,j+1} = (\mathbf{X}^\intercal\mathbf{X})^{-1}_{j+1,j+1}$ denoting the $(j+1)$-th diagonal element.

Confidence interval for a subvector $\boldsymbol{\beta}_s$

A confidence ellipsoid for a subvector $\boldsymbol{\beta}_s = (\beta_{i_1}, \ldots, \beta_{i_r})^\intercal$, with $i_j \in \{0, \ldots, p+1\}$ for $j = 1, \ldots, r$, $r \leq p+1$, for the significance level $1-\alpha$ is given by

$$\left\{\boldsymbol{\beta}_s : \frac{1}{r}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s)^\intercal \widehat{Cov(\hat{\boldsymbol{\beta}}_s)}^{-1}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s) \leq F_{r,n-p-1,1-\alpha}\right\}.$$

Confidence interval for μ_0

A confidence interval for $\mu_0 = E[y_0] = \mathbf{x}_0^\intercal\boldsymbol{\beta}$ at location $\mathbf{x}_0$ for the significance level $1-\alpha$ is given by

$$\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0}.$$

Prediction interval

A prediction interval for a new/future observation y_0 at location $\mathbf{x}_0$ for the significance level $1-\alpha$ is given by

$$\mathbf{x}_0^\intercal\hat{\boldsymbol{\beta}} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_0^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_0 + 1}.$$

2.2.6 Encoding of Categorical Predictors

In order to motivate and explain the special treatment of categorical predictors we start with a simple example.

Example 2.2.27. We regard the general model

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Let x_i be an arbitrary metric regressor, e.g. a person's age, and y a person's reaction time in a reaction test.

- Interpretation of e.g. $\beta_i = 20\frac{\text{msec}}{\text{year}}$: $E[y|x_i]$ increases by 20 msec, if x_i changes by one unit, here one year, while all other regressors are kept constant.
- Motivation:

$$E[y|x_1, \ldots, x_i+1, \ldots, x_p] - E[y|x_1, \ldots, x_i, \ldots, x_p] = (x_i+1)\beta_i - x_i\beta_i = \beta_i$$

Now, let x_1 be binary, e.g.

$$x_1 = \begin{cases} 1, \text{ male} \\ 0, \text{ female} \end{cases}$$

Now suppose a naive application of the linear regression model would yield

$$E[\text{reaction time}|x_1] = \beta_0 + x_1\beta_1, \quad \beta_1 = 30.2 \text{ [msec]}$$

- Question: is β_1 interpretable?
- Answer: yes. Men are on average 30.2 mili-seconds slower than women. β_1 reflects the increment of the response, if the covariate moves from x_i to $x_i + 1$, which in this case corresponds to a changeover in categories. The change of x_i can only be interpreted in a nominal sense.

Next, we regard a multi-categorical covariate $x^M \in \{1, \ldots, m\}$, e.g.

$$x^M = \begin{cases} 1, \text{ in the morning} \\ 2, \text{ at noon} \\ 3, \text{ at night} \end{cases}, \quad \text{or} \quad x^M = \begin{cases} 1, \text{ Hamburg} \\ 2, \text{ Munich} \\ 3, \text{ Berlin} \end{cases}$$

- Question: is the model $E(y|M) = \beta_0 + x^M\beta_1$ reasonable?
- Answer: no, because x^M is a nominally scaled variable!
- Alternative: encoding of the categories (introduction of dichotomous contrast variables).

△

We will now present two different approaches of encoding, the *dummy-encoding* and the *effect-encoding*, which belong to the most commonly used encoding schemes.

For **dummy-encoding** (also called 0-1-encoding) one has to define m contrast variables

$$x_i^M := \begin{cases} 1, \text{ if } M = i, \\ 0, \text{ else.} \end{cases}$$

The naive linear regression model would be of the following form:

$$E[y|M] = \beta_0 + x_1^M\beta_1 + \ldots + x_m^M\beta_m.$$

The main problem is that we obtain m different expectations ($E[y|M = 1], \ldots, E[y|M = m]$), but have to estimate (m+1) parameters ($\beta_0, \beta_1, \ldots, \beta_m$). The model would not be identifiable! An adequate solution is to omit one term, the so-called reference category. In doing so there exist different conventions, which term should be used as reference category. The most commonly used are either

$$\beta_1 = 0 \Longleftrightarrow M = 1 \text{ is reference category}$$

or

$$\beta_m = 0 \Longleftrightarrow M = m \text{ is reference category}\,.$$

Finally, the question arises, how the coefficients can be interpreted. We try to demonstrate this in a short example.

Example 2.2.28. Let $m = 3$ be used as the reference category:

$$\begin{aligned} E[y|M = 1] &= \beta_0 + \beta_1 \\ E[y|M = 2] &= \beta_0 + \beta_2 \\ E[y|M = 3] &= \beta_0. \end{aligned}$$

Consequently, we obtain $\beta_i = E[y|M = i] - E[y|M = 3], i = 1, 2$. The coefficient β_i represents the difference with regard to the reference category $m = 3$. △

An alternative is the **effect-encoding**. We regard the model

$$E[y|M = i] = \beta_0 + \beta_i, \quad i = 1, \ldots, m.$$

As we have already seen, the model is over-parameterized by one parameter and hence, identifiability cannot be guaranteed. The idea of the effect-encoding is to compensate the over-parametrization by an additional restriction, namely

$$\sum_{i=1}^{m} \beta_i = 0.$$

If we sum up the m expectations, $E[y|M = 1] = \beta_0 + \beta_1$ up to $E[y|M = m] = \beta_0 + \beta_m$, due to the restriction we obtain

$$\sum_{i=1}^{m} E[y|M = i] = m\beta_0 + \underbrace{\sum_{i=1}^{m} \beta_i}_{=0} \quad \Longleftrightarrow \quad \beta_0 = \frac{1}{m} \sum_{i=1}^{m} E[y|M = i].$$

This leads to the following interpretation. The constant β_0 corresponds to the mean effect, i.e. averaging over the categories. For $\beta_i, i = 1, \ldots, m$ it holds that

$$\beta_i = E[y|M = i] - \beta_0,$$

i.e. β_i corresponds to the deviation of the expected value of the i-th category from the mean over all categories. An equivalent representation of the effect-encoding is given by

$$x_i^M = \begin{cases} 1, \text{ if } M = i, \\ -1, \text{ if } M = m, \\ 0, \text{ else.} \end{cases}$$

One obtains

$$E[y|M] = \beta_0 + x_1^M\beta_1 + \ldots + x_{m-1}^M\beta_{m-1} \quad \text{and} \quad E[y|M = m] = \beta_0 - \sum_{i=1}^{m-1} \beta_i.$$

Example 2.2.29. Let y be the rent (in Euro per square meter) of flats in Munich and M be the living area with

$$M = \begin{cases} 1, \text{ Schwabing} \\ 2, \text{ Haidhausen} \\ 3, \text{ Neuperlach}\,. \end{cases}$$

Let the following data be given:

y	10.81	7.95	8.50	8.25
M	1	3	2	3

For dummy-encoding with $m = 3$ as the reference category we get the following model

$$E[y|M] = \beta_0 + x_1^M\beta_1 + x_2^M\beta_2 \quad \text{or} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

respectively, with

$$\mathbf{y} = \begin{bmatrix} 10.81 \\ 7.95 \\ 8.50 \\ 8.25 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1\ 1\ 0 \\ 1\ 0\ 0 \\ 1\ 0\ 1 \\ 1\ 0\ 0 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}.$$

The LS-estimator yields

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 8.10 \\ 2.71 \\ 0.40 \end{bmatrix}$$

Interpretation: $\hat{\beta}_0 = 8.10$ reflects the (average) rent in Neuperlach (as it is the reference category!), $\hat{\beta}_1 = 2.71$ represents the difference of the average rents between Schwabing and Neuperlach and $\hat{\beta}_2 = 0.40$ the difference between Haidhausen and Neuperlach.

Using effect-encoding with $m = 3$ as the reference category we obtain the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$\mathbf{y} = \begin{bmatrix} 10.81 \\ 7.95 \\ 8.50 \\ 8.25 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix}.$$

and the LS-estimator yields

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 9.1367 \\ 1.6733 \\ -0.6367 \end{bmatrix}$$

as well as $(\hat{y}|M = 3) = \hat{\beta}_0 - \sum_{i=1}^{2} \hat{\beta}_i = 8.1001 = \hat{\beta}_0 + \hat{\beta}_3 \iff \hat{\beta}_3 = -1.0366$.

Interpretation: $\hat{\beta}_0 = 9.1367$ reflects the mean rent (averaged over all categories). $\hat{\beta}_1 = 1.6733$ represents the difference between Schwabing and the overall mean rent, i.e. the average rent of a flat in Schwabing is increased by 1.6733 Euro per sm compared to the overall mean rent. Equivalently, $\hat{\beta}_2$ and $\hat{\beta}_3$ can be interpreted. △

Of course, also mixtures of metric and categorical covariates can be considered. For illustration, we show another short example:

Example 2.2.30. Let y be the business volume (in 1000 Euro), x_1 the invested capital (in 1000 Euro) and x_2 be an advertising strategy with

$$x_2 = \begin{cases} 1, \text{ advertising strategy 1,} \\ 2, \text{ advertising strategy 2.} \\ 3, \text{ advertising strategy 3.} \end{cases}$$

The model is given by

$$E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2,$$

where advertising strategy 1 is chosen as the reference category and z_1, z_2 are dummy variables corresponding to advertising strategies 2 and 3. Let the LS-estimator yield

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1000 \\ 10 \\ 500 \\ 700 \end{bmatrix}.$$

For $x_2 = 1$ we get $E[y|x_1, x_2 = 1] = \beta_0 + \beta_1 x_1 = 1000 + 10x_1$, for $x_2 = 2$ we get $E[y|x_1, x_2 = 2] = \beta_0 + \beta_1 x_1 + \beta_2 = 1000 + 10x_1 + 500$ and for $x_2 = 3$ we get $E[y|x_1, x_2 = 3] = \beta_0 + \beta_1 x_1 + \beta_3 = 1000 + 10x_1 + 700$. Interpretation: advertising strategy 2 leads to a business volume, which is increased by 500,000 Euro in comparison to advertising strategy 1, while advertising strategy 3 leads to a business volume, which is increased even by 700,000 Euro in comparison to advertising strategy 1, both for fixed invested capital x_1. For a fixed advertising strategy, β_1 can be interpreted as usual. △

2.3 The General Linear Regression Model

So far we have extensively studied classical linear models of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with uncorrelated and homoscedastic error terms, i.e. $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$. But the assumption of uncorrelated and homoscedastic error terms could sometimes be violated or unreasonable in the real world. In this section we extend the class of linear models by allowing for correlated or heteroscedastic error terms. The resulting model class is called *general linear regression model*. Hence, the classical linear model, which has been extensively studied so far, is an important special case. We will see that many inference problems can be solved by reduction to the classical linear model.

2.3.1 Model Definition

In the general linear regression model we substitute

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$$

by the more general assumption

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W},$$

where $\mathbf{W}$ is a positive definite matrix. In the case of heteroscedastic and (still) uncorrelated error terms we obtain the special case

$$\mathbf{W} = diag(w_1, \ldots, w_n).$$

Consequently, the heteroscedastic variances of the error terms yield $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 w_i$.

Usually, with the introduction of a more general model class, more sophisticated inference techniques are necessary in comparison to the simpler special case. For this reason, the question arises, if the use of the more general model is necessary at all. Hence, we want to analyze the effects of using the comparatively simple inference techniques of the classical linear model, if the true underlying model is belonging to the general model class, i.e. with covariance structure $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W}$ instead of $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$.

First, we investigate the effect of using the ordinary LS-estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$, if a general linear regression model is present. Similar to the classical linear model case, we obtain:

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}] = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal E[\mathbf{y}] = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

and

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}) &= Cov((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal Cov(\mathbf{y})((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)^\intercal \\ &= \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{W}\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}. \end{aligned}$$

So the LS-estimator, which was developed for the classical linear model, remains unbiased, even if the true underlying model is a general linear regression model. However, generally the covariance matrix is not consistent any more with the covariance matrix that has been derived for the classical linear model, which was $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$. As a consequence, all quantities that base on the covariance of $\hat{\boldsymbol{\beta}}$ are inaccurate. In particular, we obtain incorrect variances and standard errors of the estimated regression coefficients and consequently, also incorrect test statistics and confidence intervals.

The General Linear Regression Model

The model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is called general linear regression model, if the following assumptions hold:

1. $E[\boldsymbol{\varepsilon}] = 0$.
2. $Cov(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal] = \sigma^2\mathbf{W}$, with a known positive definite matrix $\mathbf{W}$.
3. The design matrix $\mathbf{X}$ has full column rank, i.e. $rk(\mathbf{X}) = p + 1$.

The model is called general normal regression model, if additionally the following assumption holds:

4. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{W})$.

In the following, we discuss several possibilities, how improved estimators in the general linear regression model can be constructed. First, we derive a *weighted LS-estimator* as a generalization of the ordinary LS-estimator, which has the same goodness-of-fit qualities as the ordinary LS-estimator in the classical linear model. However, a necessary condition for the application of the weighted LS-estimator is that the matrix $\mathbf{W}$ is known, which in practice is rather unrealistic. Therefore, in Section 2.3.3 and 2.3.4 we deal with inference techniques for the case that $\mathbf{W}$ is unknown. In doing so, we will confine our analysis to the two important special cases of heteroscedastic and autocorrelated error terms, respectively.

2.3.2 Weighted Least-Squares

In this section we explain an estimation procedure, which is able to avoid the aforementioned problems of the usage of the ordinary LS-estimator. The simple idea is to transform the dependent variable, the design matrix and the error terms in such a way that the transformed quantities follow a classical linear model. For illustration, first of all we regard a model with uncorrelated, but heteroscedastic error terms, i.e. $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W} = \sigma^2 diag(w_1, \ldots, w_n)$. Multiplication of the error terms by $1/\sqrt{w_i}$ yields the transformed error terms $\varepsilon_i^* = \varepsilon_i/\sqrt{w_i}$ with constant variances $Var(\varepsilon_i^*) = \sigma^2$. In order that

the model remains unchanged, also the response variable as well as all covariates (including intercept) have to be transformed accordingly. We obtain $y_i^* = y_i/\sqrt{w_i}, x_{i0}^* = 1/\sqrt{w_i}, x_{i1}^* = x_{i1}/\sqrt{w_i}, \ldots, x_{ip}^* = x_{ip}/\sqrt{w_i}$ and hence, the classical linear model

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_p x_{ip}^* + \varepsilon_i^*,$$

with homoscedastic error terms ε_i^*. Technically, this transformation corresponds to a multiplication of the model equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbf{W}^{-1/2} = diag(1/\sqrt{w_1}, \ldots, 1/\sqrt{w_n})$ from the left, i.e.

$$\mathbf{W}^{-1/2}\mathbf{y} = \mathbf{W}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{-1/2}\boldsymbol{\varepsilon}.$$

Using the transformed quantities $\mathbf{y}^* = \mathbf{W}^{-1/2}\mathbf{y}, \mathbf{X}^* = \mathbf{W}^{-1/2}\mathbf{X}$ and $\boldsymbol{\varepsilon}^* = \mathbf{W}^{-1/2}\boldsymbol{\varepsilon}$, finally one obtains

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*. \tag{2.3.1}$$

Now we are again in the setting of the classical linear model and the corresponding LS-estimator yields

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^{*\intercal}\mathbf{X}^*)^{-1}\mathbf{X}^{*\intercal}\mathbf{y}^* \\ &= (\mathbf{X}^\intercal\mathbf{W}^{-1/2}\mathbf{W}^{-1/2}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{W}^{-1/2}\mathbf{W}^{-1/2}\mathbf{y} \\ &= (\mathbf{X}^\intercal\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{W}^{-1}\mathbf{y}.\end{aligned}$$

The estimator is also known as the so-called *Aitken-estimator*. It can be shown that the Aitken-estimator minimizes the weighted sum of squared residuals

$$WLS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{1}{w_i}(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta})^2.$$

This is the reason why in this context one is also talking about *weighted residuals*. Obviously, observations with bigger variances (w_i big) get smaller weights (w_i^{-1} small) than observations with smaller variances.

Furthermore, under the assumption of normally distributed error terms it can be shown that the weighted LS-estimator is equivalent to the corresponding weighted ML-estimator for $\boldsymbol{\beta}$, i.e. $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}$. The ML-estimator for σ^2 yields

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}^\intercal\mathbf{W}^{-1}\hat{\boldsymbol{\varepsilon}}.$$

Similar to the classical linear model this estimator is biased. An unbiased estimator is given by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\boldsymbol{\varepsilon}}^\intercal\mathbf{W}^{-1}\hat{\boldsymbol{\varepsilon}} = \frac{1}{n-p-1}\sum_{i=1}^{n}\frac{1}{w_i}(y_i - \mathbf{x}_i^\intercal\hat{\boldsymbol{\beta}})^2.$$

Again, this estimate can be regarded as a restricted ML-estimator, compare Remark 2.2.9. All derivations and proofs proceed analogously to the classical linear model. Similarly, the tests and confidence intervals developed for the classical linear model can be carried over to the general setting.

This approach, which has been demonstrated so far for the case of (uncorrelated) heteroscedastic error terms, can easily be carried over to arbitrary covariance matrices $\sigma^2\mathbf{W}$. For this purpose we define the "square root" $\mathbf{W}^{1/2}$ of a matrix $\mathbf{W}$ by $\mathbf{W}^{1/2}(\mathbf{W}^{1/2})^\intercal = \mathbf{W}$. The matrix $\mathbf{W}^{1/2}$ is not unique, but can be obtained for example by using the spectral decomposition

$$\mathbf{W} = \mathbf{P}diag(\lambda_1, \ldots, \lambda_n)\mathbf{P}^\intercal$$

of $\mathbf{W}$:

$$\mathbf{W}^{1/2} = \mathbf{P}diag(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})\mathbf{P}^\intercal,$$

where λ_i denote the eigen values of $\mathbf{W}$. The vector of the response variable, the design matrix and the vector of error terms can then be transformed by multiplication from the left with the matrix

$$\mathbf{W}^{-1/2} = \mathbf{P}diag(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}})\mathbf{P}^\intercal.$$

In doing so we obtain again a model of the form (2.3.1). This is a classical linear model, because

$$E[\boldsymbol{\varepsilon}^*] = E[\mathbf{W}^{-1/2}\boldsymbol{\varepsilon}] = \mathbf{W}^{-1/2}E[\boldsymbol{\varepsilon}] = \mathbf{0}$$

and

$$Cov(\boldsymbol{\varepsilon}^*) = E[\mathbf{W}^{-1/2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal\mathbf{W}^{-\intercal/2}] = \sigma^2\mathbf{W}^{-1/2}\mathbf{W}^{-\intercal/2} = \sigma^2\mathbf{I}_n.$$

Finally, we find that the weighted LS-estimator has the same stochastic properties as the ordinary LS-estimator. The proofs are equivalent to those for the un-weighted LS-estimator.

Example 2.3.1 (Grouped data). Up to now we have assumed that *individual data* or *un-grouped data* are given, i.e. for each individual or object i from the sample with sample size n a single observation $(y_i, \mathbf{x}_i)$ is given. Hence, every value y_i of the response variable and each vector of covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\intercal$ belongs to exactly one unit i:

$$\begin{array}{c} \text{Unit } 1 \\ \vdots \\ \text{Unit } i \\ \vdots \\ \text{Unit } n \end{array} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1} & \ldots & x_{ip} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}$$

If several vectors of covariates or rows of the design matrix, respectively, are identical, the data can be *grouped*: after rearranging and combining, the data

matrix contains only *different* vectors of covariates $\mathbf{x}_i$. Additionally, the number of replicates n_i of $\mathbf{x}_i$ in the original sample of the individual data is given, as well as the arithmetic mean $\bar{y}_i$ of the corresponding individual response realizations, which have been observed for $\mathbf{x}_i$:

$$\begin{array}{c} \text{Group } 1 \\ \vdots \\ \text{Group } i \\ \vdots \\ \text{Group } G \end{array} \quad \begin{bmatrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_G \end{bmatrix} \quad \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_i \\ \vdots \\ \bar{y}_G \end{bmatrix} \quad \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{i1} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots \\ 1 & x_{G1} & \dots & x_{Gp} \end{bmatrix}$$

Here, G denotes the number of *different* vectors of covariates in the sample, which is often clearly smaller than the sample size n. In particular, this applies for binary or categorical covariates.

Grouped data can be easily treated in the context of the general linear model, by specifying $\mathbf{y} = (\bar{y}_1, \dots, \bar{y}_G)^\intercal$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2 diag(1/n_1, \dots, 1/n_G)$. △

Estimators in the General Linear Model

Weighted LS- or ML-estimator, respectively, for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{W}^{-1} \mathbf{y}.$$

Properties of the weighted LS-estimator:

1. *Expectation*: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, the weighted LS-estimator is unbiased.
2. *Covariance matrix*: $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\intercal \mathbf{W}^{-1} \mathbf{X})^{-1}$.
3. *Gauss-Markov-Theorem*: The LS-estimator has minimal variance among all linear and unbiased estimators $\hat{\boldsymbol{\beta}}^L = \mathbf{A}\mathbf{y}$,

$$Var(\hat{\beta}_j) \leq Var(\hat{\beta}_j^L), \quad j = 0, \dots, p.$$

REML-estimator for σ^2:

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \hat{\boldsymbol{\varepsilon}}^\intercal \mathbf{W}^{-1} \hat{\boldsymbol{\varepsilon}}.$$

The REML-estimator is unbiased.

2.3.3 Heteroscedastic Error Terms

In this section linear models with (uncorrelated) heteroscedastic error terms are considered. The structure of the covariance matrix is thus given by $Cov(\boldsymbol{\varepsilon}) = \sigma^2 diag(w_1, \ldots, w_n)$. In the literature, especially in the economical literature, there exist many different estimation and modeling approaches, which cannot all be presented here. A good overview can be found for example in Greene (2000). In the following, we will illustrate a two-stage estimation procedure based on the LS-principle and additionally, we will shortly sketch the ML-estimator. Both alternatives have the disadvantage that the *type* of heteroscedasticity has to be known. Before we introduce the different estimation and modeling approaches, we show first, how heteroscedastic error terms can be detected at all.

Diagnostics of heteroscedastic error terms

Starting point for the diagnostics of heteroscedastic error terms is always the estimation of a classical linear model followed by an investigation of the residuals. In the literature one finds two different approaches, how heteroscedasticity can be detected. In the statistical and rather biometrically orientated literature mainly graphical devices, especially residual plots, are proposed. The economical literature has developed a variety of statistical tests for the detection of heteroscedasticity. We will discuss both alternatives.

Residual plots

For the detection of heteroscedastic errors the standardized or studentized residuals ($\rightarrow$ see exercise) can be confronted in a residual plot with the fitted values $\hat{y}_i$ or with covariates x_{ij}. In doing so, also covariates should be considered, which are not incorporated into the model. The standardized or studentized residuals are to be preferred with regard to the ordinary residuals, as these are themselves heteroscedastic with $Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, compare Proposition 2.2.19, and consequently are less suitable for examination of heteroscedasticity. In the case of homoscedasticity, both the standardized or studentized residuals should sprinkle unsettled and with constant variability around zero. Otherwise, heteroscedastic variances are indicated.

Example 2.3.2 (Munich rent levels - diagnostics of heteroscedastic error terms). We illustrate the diagnostics of heteroscedastic error terms by the help of the Munich rent levels data set. For simplicity, we regard a linear model with the *net rent* as response and only two covariates, *size* of the flat and *year of construction*. The scatter plots between the net rent and both covariates in Figure 2.5 indicate a linear effect of the size and a slightly non-linear effect of the year of construction. Consequently, we model the effect of the year of con-

struction with orthogonal polynomials of degree 3[1], while the size is assumed to have a simple linear influence (we use an orthogonal polynomial of degree 1), and end up with the following classical regression model

$$rent_i = \beta_0 + \beta_1 sizeo_i + \beta_2 yearo_i + \beta_3 yearo2_i + \beta_4 yearo3_i + \varepsilon_i. \quad (2.3.2)$$

We present the corresponding results of the fixed effects in Table 2.1. Figure

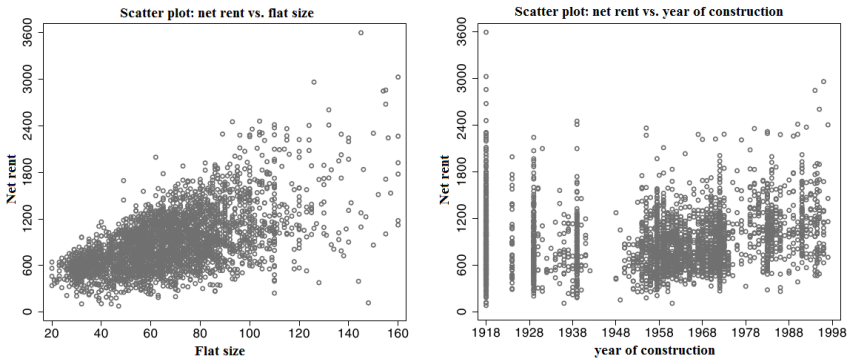

Fig. 2.5: Munich rent levels: scatter plots between net rent and the explanatory variables *size* of the flat and *year of construction*.

variable	$\hat{\beta}_k$	$\hat{\sigma}_{\hat{\beta}_k}$	t-value	p-value
intercept	895.903	5.131	174.600	< 0.001
sizeo	237.544	5.347	44.430	< 0.001
yearo	105.956	5.277	20.080	< 0.001
yearo2	61.394	5.203	11.800	< 0.001
yearo3	-0.387	5.131	-0.080	0.940

Table 2.1: Munich rent levels: estimation results for an un-weighted regression.

2.6 (a) and (b) show the estimated effect of the flat size and the construction year (in each case together with the partial residuals). The studentized residuals depending on the fitted net rent, size of the flat and year of construction are illustrated in Figures (c), (d) and (e). There is clear evidence of heteroscedastic variances, but probably the variances do not only depend on the size of the flat, but also on the construction year. △

[1] If certain covariates are developed in orthogonal polynomials, the corresponding columns of the design matrix are centered and orthogonalized, in this case the columns corresponding to *size* as well as to *year*, $year^2$ and $year^3$. In **R**, this can be done e.g. by use of the `poly` function. Usually, the computation of the LS-estimator is numerically more stable, if orthogonal polynomials are used. In the exercises it is shown, how orthogonal polynomials can be easily generated by application of the properties of the LS-method.

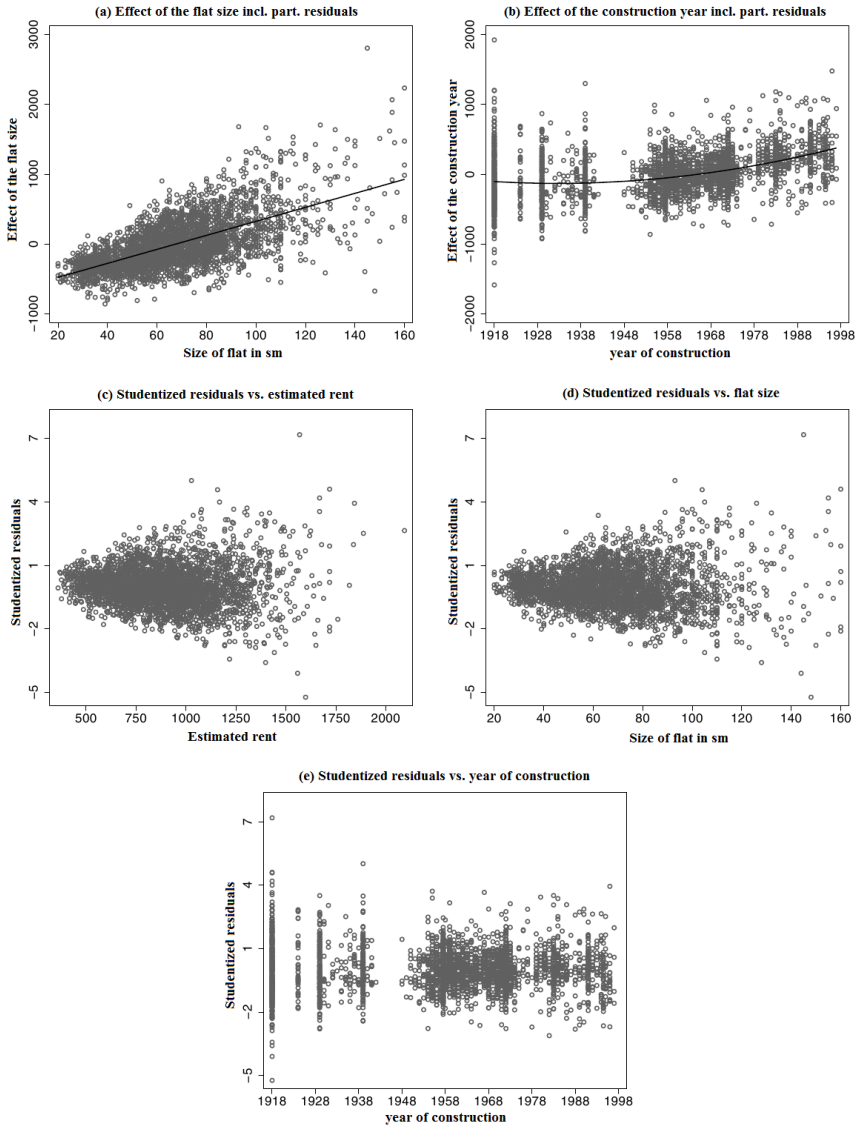

Fig. 2.6: Figures (a) and (b) show the effect of the flat size and the construction year including the partial residuals. Figures (c), (d) and (e) illustrate the studentized residuals depending on the fitted *net rent*, *size* of the flat and *year of construction*.

Tests on heteroscedasticity

Tests on heteroscedasticity are described in detail in the economical literature, see for example Greene (2000) and Judge et al. (1980). Exemplarily, we want to describe a test following Breusch and Pagan. The test is based on the assumption of a multiplicative model for the variance of the error terms:

$$\sigma_i^2 = \sigma^2 \cdot h(\alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_q z_{iq}). \tag{2.3.3}$$

Here, h is a function, which is not depending on i, and $\mathbf{z}_1, \ldots, \mathbf{z}_q$ are covariates that are assumed to have a possible influence on the variance. The hypothesis of homoscedastic variances is equivalent to $\alpha_1 = \cdots = \alpha_q = 0$. So the Breusch-Pagan-test is testing the hypotheses

$$H_0 : \alpha_1 = \cdots = \alpha_q = 0 \quad \text{vs.} \quad H_1 : \alpha_j \neq 0 \text{ for at least one } j.$$

The test is based on the execution of an auxiliary regression between the response variable

$$g_i = \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}_{ML}^2}$$

and the covariates $\mathbf{z}_1, \ldots, \mathbf{z}_q$. The quantities $\hat{\varepsilon}_i$ and $\hat{\sigma}_{ML}^2$ denote the residuals and the ML-estimator for σ^2 of the basic classical linear model $y = \mathbf{x}^\intercal \boldsymbol{\beta} + \varepsilon$ with homoscedastic error terms. A suitable test-statistic is

$$T = \frac{1}{2} \sum_{i=1}^{n} (\hat{g}_i - \bar{g})^2.$$

If none of the potential influence variables has an effect on the variance, i.e. if no heteroscedasticity is present, it should hold that $\hat{g}_i \approx \bar{g}$. The bigger T, i.e. the bigger the sum of squared discrepancies between the estimated values $\hat{g}_i$ and the mean $\bar{g}$, the more plausible becomes the alternative hypothesis of heteroscedastic variances. Under H_0, the distribution of T is independent from the function h and asymptotically χ^2-distributed with q degrees of freedom, i.e. $T \sim \chi_q^2$. For a given significance level of α the null hypothesis is rejected, if the realization of the test statistic exceeds the $(1-\alpha)$-quantile of the χ_q^2-distribution.

Example 2.3.3 (Munich rent levels - Breusch-Pagan-test). As it has to be assumed that the variances may depend (possibly non-linearly) on both the size of the flat and the year of construction, we imply the following underlying variance model

$$\begin{aligned}\sigma_i^2 = \sigma^2 \cdot h(&\alpha_0 + \alpha_1 sizeo_i + \alpha_1 sizeo2_i + \alpha_1 sizeo3_i \\ &+\alpha_4 yearo_i + \alpha_5 yearo2_i + \alpha_6 yearo3_i),\end{aligned}$$

where $sizeo, sizeo2$ and $sizeo3$ as well as $yearo, yearo2$ and $yearo3$ are orthogonal polynomials of degree 3 of the flat size and the year of construction, respectively. Based on this model we obtain the test-statistic realization $T = 997.16$ for the Breusch-Pagan-test. The corresponding p-value is very close to zero. Consequently, the Breusch-Pagan-test gives further clear evidence (besides the studentized residuals) for heteroscedastic variances. △

Finally, we want to critically comment on the heteroscedasticity tests. For some scientists the application of a technical test my seem more exact than an analysis of residual plots. Nevertheless, heteroscedasticity is diagnosed in most cases using explorative tools, and in only very few cases there exist special theories about type and degree of the heteroscedasticity. Hence, there is even more uncertainty for the modeling of the error variances in the linear model in comparison to the modeling of the expectation. But the validity of tests is especially depending on the fact, whether the underlying model is specified correctly or not. For example, in the test situation of the Breusch-Pagan-test, a multiplicative variance with an exactly specified vector of covariates is assumed. But the number of covariates that are used in (2.3.3) directly affects the distribution of the test-statistic. Consequently, these tests should be used mainly as explorative (heuristic) tools and in no case as the single instrument to detect heteroscedasticity!

Procedures for heteroscedasticity

If evidence for heteroscedasticity is found, some suitable procedures have to be considered, in order to avoid false conclusions. In the following we shortly sketch two of the most common ones.

Transformation of variables

One possibility is to use a model with multiplicative error terms. One of the most common multiplicative error models is based on an exponential structure, i.e.

$$\begin{aligned} y_i &= \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i) \\ &= \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_p x_{ip})\exp(\varepsilon_i). \end{aligned} \tag{2.3.4}$$

If the error terms follow a normal distribution, $\varepsilon_i \sim N(0, \sigma^2)$, then the quantities $\exp(\varepsilon_i)$ and y_i are logarithmically normally distributed, see Definition B.1.10 from Appendix B. Under consideration of the variance for logarithmically normally distributed random variables, we obtain

$$Var(\exp(\varepsilon_i)) = \exp(\sigma^2) \cdot (\exp(\sigma^2) - 1)$$

and hence

$$Var(y_i) = (\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta}))^2 \exp(\sigma^2) \cdot (\exp(\sigma^2) - 1).$$

We find that in this model the variances of y_i are heteroscedastic, although the variances of the error terms themselves are homoscedastic. Note that simply by taking the logarithm, the exponential model can be transformed into an ordinary linear model $\log(y_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i$, with homoscedastic variances. Thus, if a multiplicative model seems adequate, a simple procedure to deal with heteroscedastic variances is to logarithmize the response variable and fit an ordinary linear model with the transformed response.

Two-stage estimation

In general, for heteroscedastic variances, beside the regression coefficients $\boldsymbol{\beta}$ and the variance parameter σ^2, also the weights w_i are unknown. Hence, an obvious procedure in the case of heteroscedastic variances would be to *jointly* estimate all unknown parameters. However, in the following we first sketch a simple two-stage estimation scheme.

Due to $E[\varepsilon_i] = 0$, we have $E[\varepsilon_i^2] = Var(\varepsilon_i) = \sigma_i^2$ and are able to represent ε_i^2 in the following form

$$\varepsilon_i^2 = \sigma_i^2 + v_i,$$

with v_i denoting the discrepancy of the quadratic error from its expectation. In most cases, σ_i^2 is depending on several covariates. Thus, an obvious strategy is to assume

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_q z_{iq} = \mathbf{z}_i^\intercal \boldsymbol{\alpha}.$$

The vector $\mathbf{z}$ contains all covariates that have an influence on the variance. In many applications the vector $\mathbf{z}$ will be identical with the covariates $\mathbf{x}$. But it is also possible to include new, other variables into $\mathbf{z}$, which are not yet contained in $\mathbf{x}$. For the estimation of the vector of unknown parameters $\boldsymbol{\alpha}$, in principle, it is possible to use the LS-approach and fit a linear model with the quadratic error terms ε_i^2 as response variable and with regressors $\mathbf{z}_i$. As the error terms are unobserved, they need to be estimated by the residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}$, which can be obtained by a preceding unweighted regression between y and $\mathbf{x}$. This results in the following two-stage approach:

(i) Perform an unweighted regression between y and $\mathbf{x}$, yielding temporary estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\varepsilon}_i$.

(ii) Perform an unweighted regression between $\hat{\varepsilon}_i^2$ and the variables $\mathbf{z}_i$, which explain the variance, yielding parameter estimates $\hat{\boldsymbol{\alpha}}$. Afterwards, fit a general linear model, using the estimated weights

$$\hat{w}_i = \mathbf{z}_i^\intercal \hat{\boldsymbol{\alpha}}.$$

Note that this procedure is not always applicable in practice, because it is not guaranteed that $\mathbf{z}_i^\intercal \hat{\boldsymbol{\alpha}}$ is always positive. But if $\mathbf{z}_i^\intercal \hat{\boldsymbol{\alpha}} < 0$, this would yield negative weights $\hat{w}_i$ and consequently negative expected variances $\hat{\sigma}^2$. Alternatively, the model

$$\sigma_i^2 = \exp(\mathbf{z}_i^\mathsf{T}\boldsymbol{\alpha})$$

can be established. Here, the exponential function guarantees that the estimated variances are strictly positive. In this situation the parameter estimates $\hat{\boldsymbol{\alpha}}$ are obtained fitting the regression model

$$\log(\varepsilon_i^2) = \mathbf{z}_i^\mathsf{T}\boldsymbol{\alpha} + v_i.$$

This yields the weights

$$\hat{w}_i = \exp(\mathbf{z}_i^\mathsf{T}\hat{\boldsymbol{\alpha}})$$

for the final general linear model, which is fitted in a terminal step.

Example 2.3.4 (Munich rent levels - two-stage estimation with heteroscedasticity). We illustrate the two-stage estimation approach by the help of the Munich rent level data. In the Examples 2.3.2 and 2.3.3 we detected heteroscedastic variances for the model from (2.3.2). The two-stage estimates are obtained by the following three steps:

Step 1: Classical linear model:
We estimate a classical linear model, compare Example 2.3.2. Figures 2.7 (a) and (b) illustrate scatter plots between $\log(\hat{\varepsilon}_i^2)$ and the flat size and the year of construction, respectively. Both figures provide evidence that the variances σ_i^2 vary depending both on size and year of construction of the flat.

Step 2: Auxiliary regression:
For the determination of the estimated weights $\hat{w}_i$ for the weighted regression, we fit the regression model

$$\begin{aligned}\log(\hat{\varepsilon}_i^2) = \alpha_0 + \alpha_1 sizeo_i + \alpha_2 sizeo2_i + \alpha_3 sizeo3_i \\ + \alpha_4 yearo_i + \alpha_5 yearo2_i + \alpha_6 yearo3_i + v_i\end{aligned}$$

using least squares. Figures 2.7 (c) and (d) illustrate the estimated effects of the flat size and the year of construction together with the partial residuals. Obviously, the variability of the net rent is increasing with increasing size of the flats, but it is also seen that this increase slows down for very large flats. The effect of the year of construction is S-shaped, but weaker than the effect of the flat size. The effect can be interpreted quite well: the net rent of older flats is more varying than for modern flats.

Step 3: Weighted linear model:
In the last step the regression model from (2.3.2) is fitted again using the weights

$$\hat{w}_i = \exp(\hat{\eta}_i),$$

with $\hat{\eta}_i = \hat{\alpha}_0 + \hat{\alpha}_1 sizeo_i + \ldots + \hat{\alpha}_6 yearo3_i$. This yields the estimates presented in Table 2.2. The estimated standard errors base on the estimated covariance matrix

$$\widehat{Cov(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\mathbf{X}^\intercal diag(\hat{1}/w_1, \ldots, \hat{1}/w_n)\mathbf{X})^{-1}.$$

Overall, these are smaller than for the un-weighted regression model from Example 2.3.2. On the basis of the corrected standard errors, the tests and confidence intervals developed for the classical linear model can be used. This would result in slightly smaller confidence intervals here in comparison to Example 2.3.2.

△

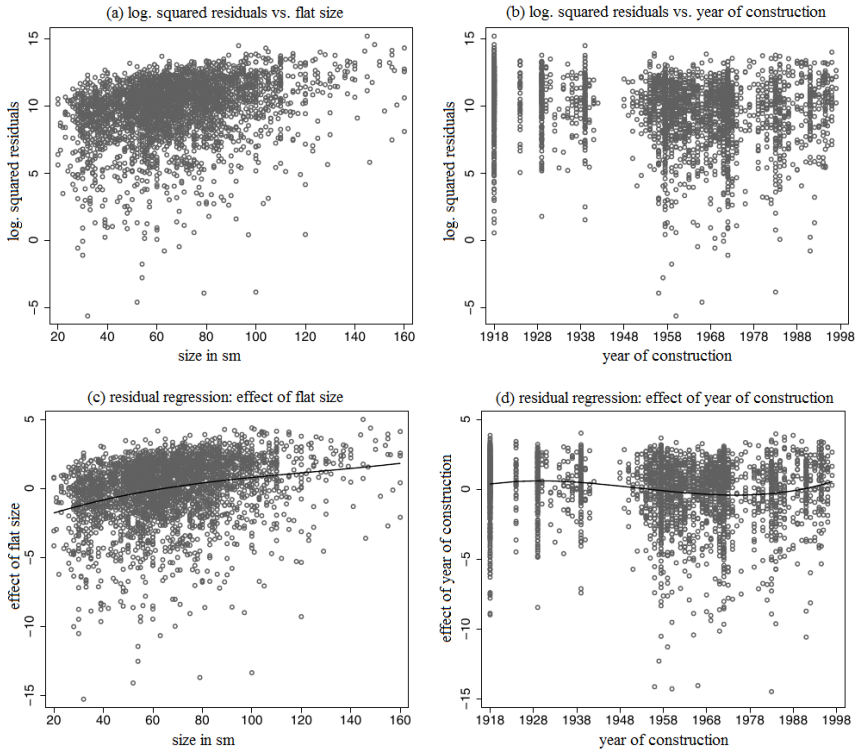

Fig. 2.7: Munich rent levels with heteroscedastic errors. Figures (a) and (b) show scatter plots between $\log(\hat{\varepsilon}_i^2)$ and the explanatory variables *size* of the flat and *year of construction*. In Figures (c) and (d) the effects of the flat size and the construction year are illustrated for the residual regression, including the partial residuals.

Joint ML-estimation

In general, a *simultaneous estimation* of $\boldsymbol{\beta}, \boldsymbol{\alpha}$ and σ^2 would be more elegant in comparison to the two-stage approach presented here. Possible approaches

variable	$\hat{\beta}_k$	$\hat{\sigma}_{\hat{\beta}_k}$	t-value	p-value
intercept	894.739	5.087	175.900	< 0.001
sizeo	229.585	4.645	49.430	< 0.001
yearo	94.959	5.180	18.330	< 0.001
yearo2	49.301	4.551	10.830	< 0.001
yearo3	-2.582	4.642	-0.560	0.578

Table 2.2: Munich rent levels: estimation results for the weighted regression.

would be complete ML-estimation or combinations of ML-estimation for $\boldsymbol{\beta}$ and REML-estimation for the variance parameters $\boldsymbol{\alpha}$ and σ^2. We abstain on presenting these approaches in detail, as these go beyond the scope of most basic software programs and are not implemented standardly.

2.3.4 Autocorrelated Error Terms

Beside heteroscedastic error terms, correlation among errors is one of the main reasons for violations of the model assumptions of the classical linear model, with the consequence that the general linear model has to be used for the fitting. For example, autocorrelated errors can occur, if the model is misspecified. There are two major reasons for a misspecification of a linear regression model:

- The influence of a covariate is not modeled correctly. Here, several approaches for the modeling of non-linear effects for metric covariates, such as polynomials, orthogonal polynomials or smoothing splines[2] can be helpful. Also interaction effects could be considered. If autocorrelated error terms are detected this source of error should be taken into account first.
- Some relevant covariates can not be observed and thus are not considered in the model. In the context of autocorrelated error terms this problem becomes especially relevant for panel or longitudinal data, if the disregarded covariates manifest a time trend. In this case, the estimation techniques presented in this section can be used, at least to get more exact results and improved predictions.

[2] There is a vast amount of literature available concerning smoothing splines. A nice introduction is provided e.g. by Ruppert et al. (2003) or Ramsay and Silverman (2005).

First-order autocorrelation

We will confine our analysis to the simplest *error process*, which is most frequently applied in practice, namely autocorrelated error terms with first-order autocorrelation. In particular, we assume that the error terms follow a first-order autoregressive process, denoted by AR(1), i.e.

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i,$$

with $-1 < \rho < 1$. With respect to u_i we make the following assumptions:

1. $E[u_i] = 0$.
2. $Var(u_i) = E[u_i^2] = \sigma_u^2, \quad i = 1, \ldots, n.$
3. $Cov(u_i, u_j) = E[u_i u_j] = 0, \quad i \neq j.$

Additionally, we assume that the process has already started in the past. From these assumptions we can deduce a particular linear model with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and covariance matrix

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W} = \frac{\sigma_u^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix}. \tag{2.3.5}$$

The derivation of the covariance matrix will be done in the exercises.

By the help of the covariance matrix it is easy to compute the correlation between ε_i and the error term ε_{i-j} lagged by j time periods. We obtain the so-called *autocorrelation function*

$$ACF(j) = \frac{Cov(\varepsilon_i, \varepsilon_{i-j})}{Var(\varepsilon_i)} = \rho^j, \quad j = 0, 1, 2, \ldots$$

Hence, a typical characteristic of autocorrelated error terms with first-order autocorrelation is a slowly declining correlation between ε_i and the other lagged errors. If ρ is positive, the correlation is decreasing geometrically. If ρ is negative, the correlation is decreasing with alternating sign, compare the left column in Figure 2.8, which illustrates the autocorrelation function for different values of ρ. Such graphical representations are also known as *correlograms*.

Beside the autocorrelation function, the so-called *partial autocorrelation function* is another characteristic, which is useful in the context of correlated errors or general stochastic processes, respectively. The partial autocorrelation function $PACF(j)$ between ε_i and ε_{i-j} is defined as the regression coefficient α_j in the regression model

$$\varepsilon_i = \alpha_1\varepsilon_{i-1} + \ldots + \alpha_j\varepsilon_{i-j} + v_i. \tag{2.3.6}$$

It follows that $PACF(1) = ACF(1)$. For autocorrelated error terms with first-order autocorrelation obviously it holds that $PACF(1) = \rho$ and $PACF(j) = 0$, for $j > 1$. The right column in Figure 2.8 shows the partial autocorrelation function for some $AR(1)$-processes. Typical is the abrupt decline of the partial autocorrelation for $j > 1$. The coefficient α_j in Equation (2.3.6) can be interpreted as the correlation between $\varepsilon_i - \alpha_1\varepsilon_{i-1} - \ldots - \alpha_{j-1}\varepsilon_{i-j+1}$ and ε_{i-j}. So the notation *partial* autocorrelation comes from the fact that not simply the correlation between ε_i and ε_{i-j} is derived, but the correlation between ε_i and ε_{i-j} under exclusion of the influence of all the error terms in between.

Diagnostics of autocorrelated error terms

Similar to the diagnostics of heteroscedastic error terms the instruments of diagnostics base on the residuals of a classical linear model, which is initially assumed to be the underlying model. The following instruments have proofed of value in practice:

Graphical presentation of the residuals over time

The most nearby possibility for the detection of correlation is provided by residual plots, in the present case scatter plots of the residuals or studentized residuals over time can be used. If positive residuals are systematically followed by positive ones, this in an indicator for positive autocorrelation. If on the other hand positive residuals are systematically followed by negative ones, this in an indicator for negative autocorrelation.

Empirical autocorrelation functions

Hints on the type of correlation are also provided by the *empirical* autocorrelations, partial autocorrelations and their visualization in correlograms. The empirical autocorrelations are estimates of $ACF(j)$ and can be derived by

$$\widehat{ACF}(j) = \frac{\widehat{Cov}(\varepsilon_i, \varepsilon_{i-j})}{\widehat{Var}(\varepsilon_i)} \quad \text{with} \quad \widehat{Cov}(\varepsilon_i, \varepsilon_{i-j}) = \frac{1}{n-j} \sum_{i=j+1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-j}.$$

Basically, the empirical autocorrelations are the empirical correlations between the residuals and the residuals lagged by j time periods.

The partial empirical autocorrelations can be determined by repeated estimation of the regression model from (2.3.6) for $j = 1, 2, 3, \ldots$ and by fixing $\widehat{PACF}(j) = \hat{\alpha}_j$. In doing so, the error terms ε_i in (2.3.6) are substituted by their estimates, namely the residuals $\hat{\varepsilon}_i$. An analysis of the correlograms as well as the partial correlograms allows to draw some conclusions concerning the presence of autocorrelation. If all empirical autocorrelations and partial autocorrelations are close to zero, on can rather expect uncorrelated errors. An anew estimation of the model due to autocorrelation then becomes not

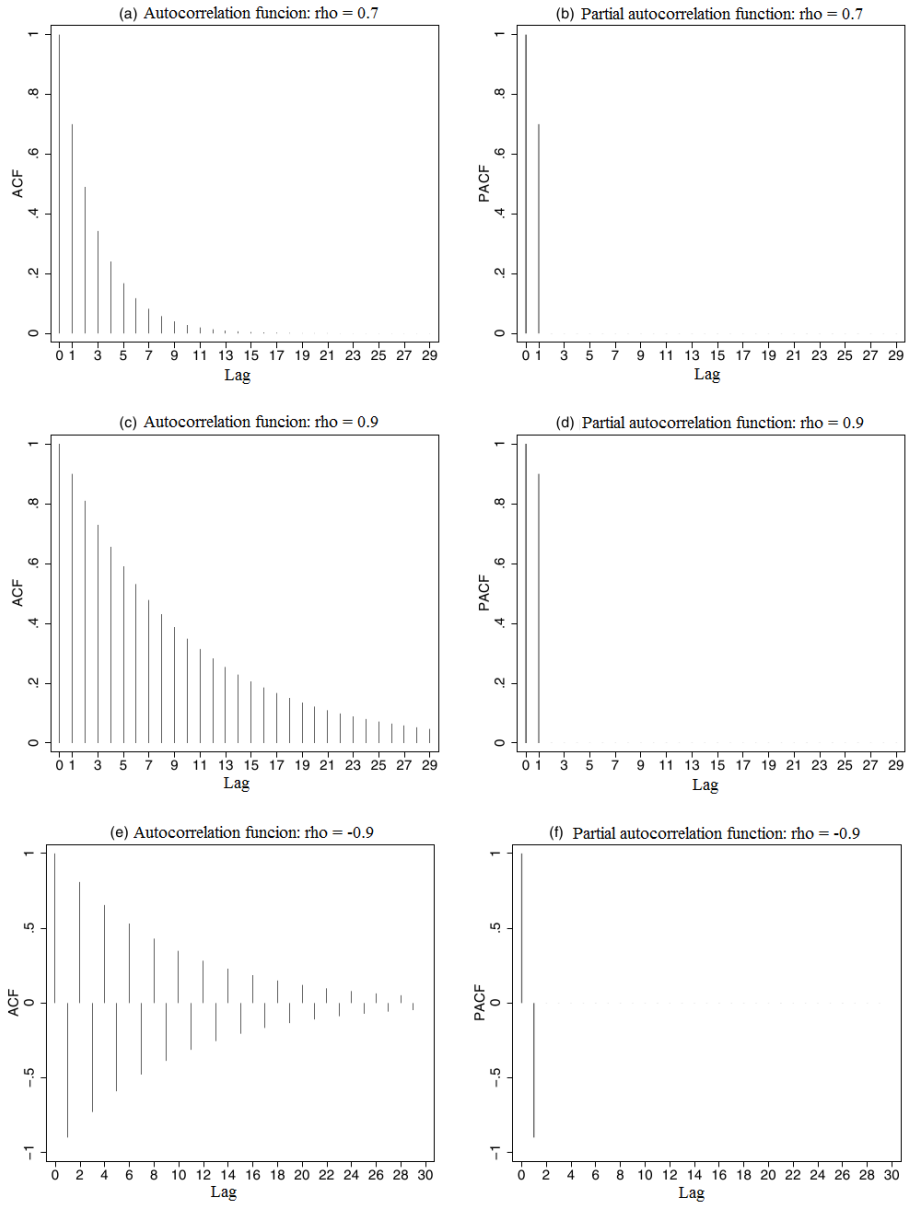

Fig. 2.8: Autocorrelation function and partial autocorrelation function for autocorrelated error terms with first-order autocorrelation for $\rho = 0.7, 0.9, -0.9$.

necessary. If the empirical correlations are similar to the theoretical autocorrelations and partial autocorrelations of $AR(1)$-error terms, this reflects

an indicator on the presence of autocorrelated error terms of first-order autocorrelation. If the correlograms manifest features that are not typical for $AR(1)$-error terms, this indicates more complex correlation structures, which will be further analysed in the context of time series analysis later on in this course.

Example 2.3.5 (Simulated data - graphical diagnostics of autocorrelated errors). We simulate a regression model $y_i = -1 + 2x_i + \varepsilon_i$ with $x_i \sim U[0,1]$ and for the error terms we set up the $AR(1)$-process $\varepsilon_i = 0.9\varepsilon_{i-1} + u_i$, $u_i \sim N(0, 0.05^2)$. For illustration of the diagnostics of autocorrelated error terms we fit the data using a classical linear regression model and obtain the estimate $\hat{y}_i = -1.010 + 1.981x_i$. Figure 2.9 shows the residuals over time as well as the empirical autocorrelation and partial autocorrelation functions of the residuals. The figure shows almost ideal indications of first-order autocorrelation. The residuals are highly correlated, the empirical autocorrelation is decreasing geometrically and the partial autocorrelations are almost equal to zero for $j > 1$. △

Tests on autocorrelation - Durbin-Watson-test

Beside some graphical tools, there are also statistical tests available for the detection of autocorrelation. The most frequently used test in this context is a test on serial correlation, developed by Durbin and Watson (Durbin and Watson, 1950, 1951, 1971). The Durbin-Watson-test is testing the hypotheses

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0$$

and is based on the test-statistic

$$d = \frac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}.$$

This test-statistic originates from the following deliberation: in particular for large sample sizes n, the test-statistic can be approximated as

$$d = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i^2 + \sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2 - 2\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2} \approx 1 + 1 - 2\hat{\rho} = 2(1 - \hat{\rho}),$$

and due to $-1 < \hat{\rho} < 1$ it follows that $0 < d < 4$. If d is close to 2, this means that $\hat{\rho} \approx 0$, and one will keep the null hypothesis. The closer the test-statistic d gets to 0 or 4, respectively, the closer $\hat{\rho}$ gets to 1 or -1, respectively, and the null hypothesis is rather rejected. The determination of the corresponding distribution of d under H_0 is quite complicated, because it is depending on the design matrix. Hence, a test-decision can possibly be difficult. Durbin and Watson found at least a partial solution to this problem. For certain intervals in the range of d a test decision can be made. These intervals are depending on the upper and lower bounds d_l and d_u, which can be found in collection books

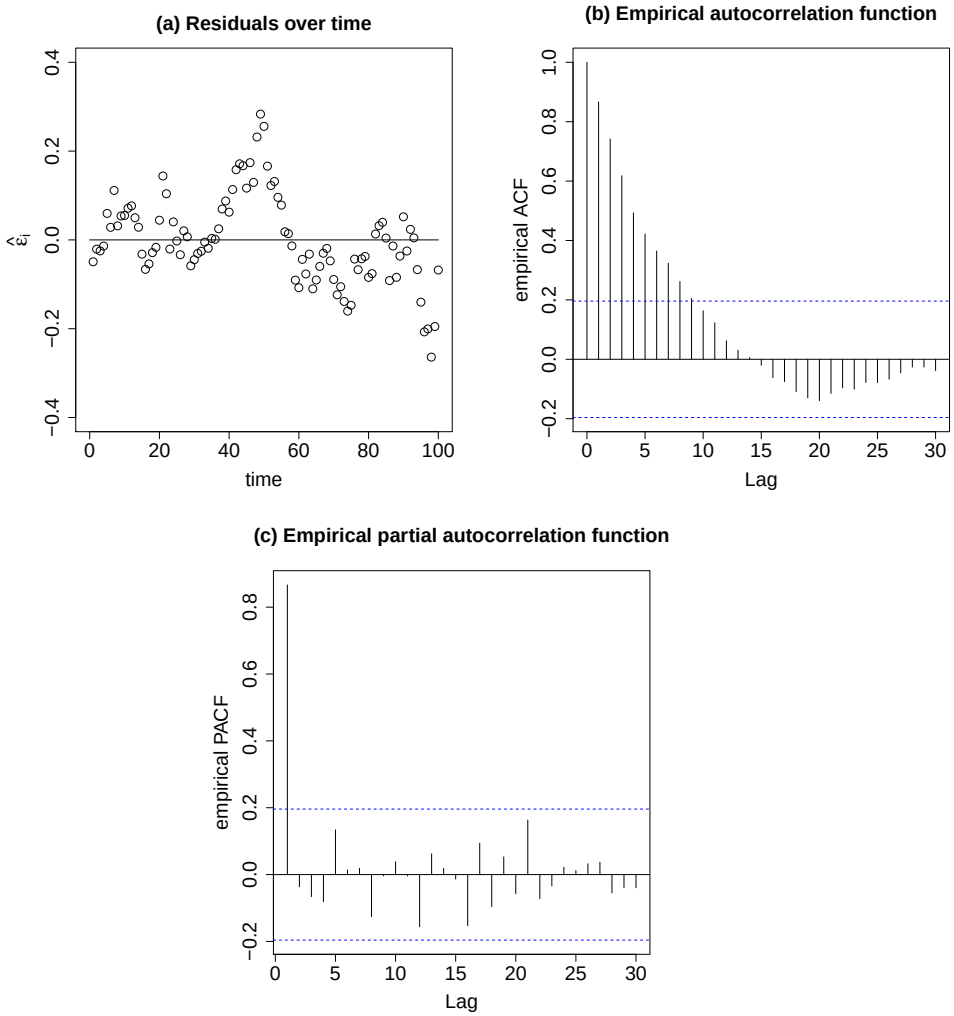

Fig. 2.9: Residuals over time, empirical autocorrelation function and partial autocorrelation function of the residuals for simulated data with positive first-order autocorrelation ($\rho = 0.9$).

containing tables with different sample sizes n and different numbers $p + 1$ of regressors. Figure 2.10 graphically illustrates the acceptance and rejection areas of the Durbin-Watson-test.

Furthermore, thanks to modern computer technologies, the p-values of the Durbin-Watson-test can be computed numerically. However, in some software programs these techniques are not yet integrated. There is an implementation

available e.g. in the software program **R**, namely the `dwtest`-function of the `lmtest`-package.

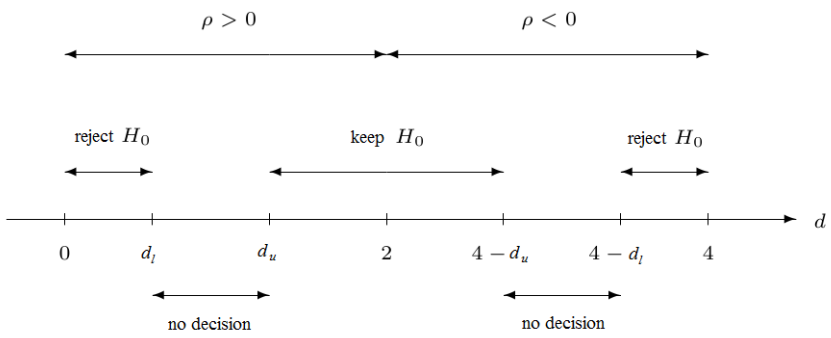

Fig. 2.10: Acceptance and rejection areas of the Durbin-Watson-test.

Example 2.3.6 (Simulated data - Durbin-Watson-test). For the simulated data from Example 2.3.5 the `dwtest`-function yields a realisation $d = 0.2596$ of the test-statistic together with a very low p-value. Hence, the Durbin-Watson-test confirms the clear evidence of autocorrelation, which has been already indicated by our preceding graphical analysis. △

Procedures for first-order autocorrelation

As already mentioned, autocorrelated error terms indicate a misspecification (of whatsoever nature) of the model. Hence one should investigate, if it is possible to eliminate this misspecification, before the estimation procedures described below are established. Possible improvements can be gained for example by the incorporation of additional covariates (which have not been considered so far) or by non-linear modeling of metric covariates.

If the correlation cannot be removed, one can revert to estimation procedures for models with autocorrelated error terms. In a model with autocorrelated error terms of first-order autocorrelation we need estimates for the regression parameters $\boldsymbol{\beta}$, the variance parameter σ^2 and the correlation parameter ρ. In the literature a variety of estimation techniques exists. We will first sketch a two-stage estimation scheme. In the first step, the correlation parameter ρ is estimated based on ordinary LS-estimation. Afterwards, the regression parameters are estimated based on the weighted LS-approach. Finally, we shortly sketch possible ML-approaches. A good overview can be found in Judge et al. (1980).

Two-stage estimation

An estimation of the model with autocorrelated error terms is obtained as follows:

1. Perform un-weighted linear regression between y and $\mathbf{x}$, yielding $\hat{\boldsymbol{\beta}}$ and $\hat{\varepsilon}_i$.
2. Estimate the correlation parameter ρ using the empirical correlation coefficient between $\hat{\varepsilon}_i$ and $\hat{\varepsilon}_{i-1}$, i.e.

$$\hat{\rho} = \frac{\sum_{i=2}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\sqrt{\sum_{i=2}^{n} \hat{\varepsilon}_i^2} \sqrt{\sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2}}. \tag{2.3.7}$$

3. Insert $\hat{\rho}$ into the weight matrix $\mathbf{W}$ from (2.3.5) and obtain an estimate $\hat{\mathbf{W}}$. Use $\hat{\mathbf{W}}$ for an anew estimation of the regression parameters, now using weighted LS-estimation.

The estimations can be improved by iterating steps 2 and 3 until convergence. This procedure is known as *Prais-Winsten-estimation*. One can show under quite general conditions that the resulting estimator for $\boldsymbol{\beta}$ is consistent. In addition to the method originally proposed by Prais and Winston, several modifications exist, compare Greene (2000).

ML-estimation

For the ML-estimation, again we need the additional assumption of normally distributed error terms. The corresponding likelihood is then given by

$$L(\boldsymbol{\beta}, \sigma^2, \rho) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} |\mathbf{W}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal \mathbf{W}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right). \tag{2.3.8}$$

First we compute

$$\mathbf{W}^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}$$

and $|\mathbf{W}| = \frac{1}{1-\rho^2}$. Hence, we obtain the log-likelihood function

$$l(\boldsymbol{\beta}, \sigma^2, \rho) = -\frac{n}{2} \log(\sigma^2) + \frac{1}{2} \log(1-\rho^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal \mathbf{W}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.3.9}$$

The maximum of the log-likelihood cannot be derived in closed form, in fact, the ML-estimator has to be determined iteratively. Details can be found in Greene (2000) or in the original literature that is referred there.

Example 2.3.7 (Simulated data - Prais-Winsten- and ML-estimators). For the simulated data from Example 2.3.5 we present the results for the Prais-Winsten-estimator in Table 2.3 and for the ML-estimator in Table 2.4, based on maximization of the log-likelihood (2.3.9) using the **R**-function `bobyqa` from the `minqa`-package. △

variable	$\hat{\beta}_k$	$\hat{\sigma}_{\hat{\beta}_k}$	t-value	p-value
Intercept	-1.030	0.037	-27.958	< 0.001
X	2.008	0.013	154.127	< 0.001

Table 2.3: Prais-Winsten-estimator for simulated data.

variable	$\hat{\beta}_k$	$\hat{\sigma}_{\hat{\beta}_k}$	t-value	p-value
Intercept	-1.030	0.035	-29.316	< 0.001
X	2.008	0.013	155.149	< 0.001

Table 2.4: ML-estimator for simulated data based on the **R**-function `bobyqa` from the `minqa`-package.

Example 2.3.8 (Supermarket-Scanner-Data). The data are served by supermarkets, which are nowadays collected as a matter of routine during the payment transactions at the supermarket counters. In this context the relationship between the weekly disposal of a certain product (here a coffee brand) and its price and the price of a competitive product, respectively, should be investigated. In the following, the multiplicative model from Equation (2.3.4) is used to model this relationship. Figure 2.11 shows both scatter plots of the disposal of the product versus its price (a) and versus the price of a competitive product (b), respectively, for five different supermarkets. We will now use this data to illustrate a possible procedure, if correlated errors are present. For simplicity we analyse just the data of a single supermarket, and due to the multiplicative model for the error terms, we use the logarithmized disposal as response instead of the untransformed disposal. Figure 2.12 shows the scatter plot between the logarithmic disposal and the product's price. Additionally, the estimated log-disposal based on an ordinary LS-regression is shown, if the influence of the regressor *price* is modeled linearly (solid line) or using a cubic polynomial (dashed line), respectively. The corresponding residuals

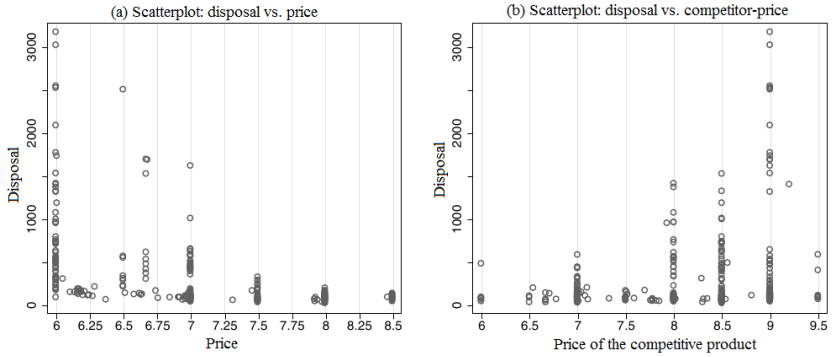

Fig. 2.11: Scatter plots between the disposal of a certain product and its price (left) and the price of a competitive product (right), respectively.

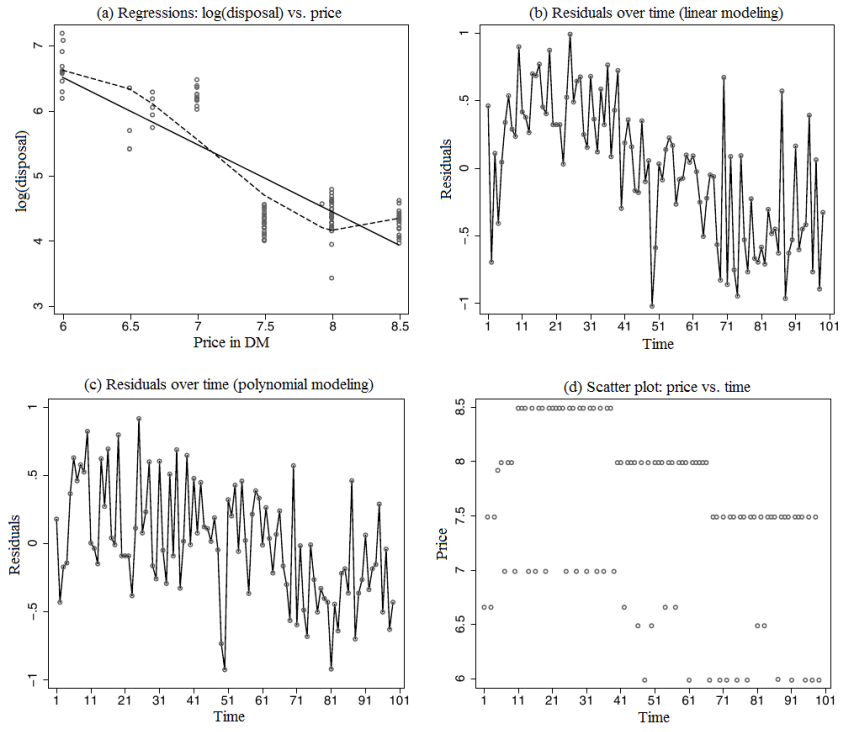

Fig. 2.12: Figure (a): scatter plot between log-disposal and price, including the estimated log-disposal for linear (solid line) and cubic (dashed line) modeling of the price effect. Figures (b) and (c) show the corresponding residuals over time. Figure (d) shows the price of the coffee brand over time.

over time are illustrated in Figures 2.12 (b) and (c). In both cases the residuals are clearly correlated. The test-statistic of the Durbin-Watson test yields $d = 1.1751$ (p-value $= 1.534 \cdot 10^{-5}$) for linear modeling and $d = 1.504$ (p-value $= 0.01142$) for polynomial modeling. So for both modeling approaches the null hypothesis $H_0 : \rho = 0$ can be rejected on the level of significance $\alpha = 0.05$. For the level $\alpha = 0.01$ the null hypothesis can only be rejected for linear modeling. This already indicates that a more sophisticated modeling can reduce the autocorrelation. In order to get a more reliable validation we also investigate the empirical autocorrelation and partial autocorrelation functions, see Figure 2.13. The plots shown there additionally contain point-wise 95%-confidence intervals, but we don't go into detail about their derivation here. The correl-

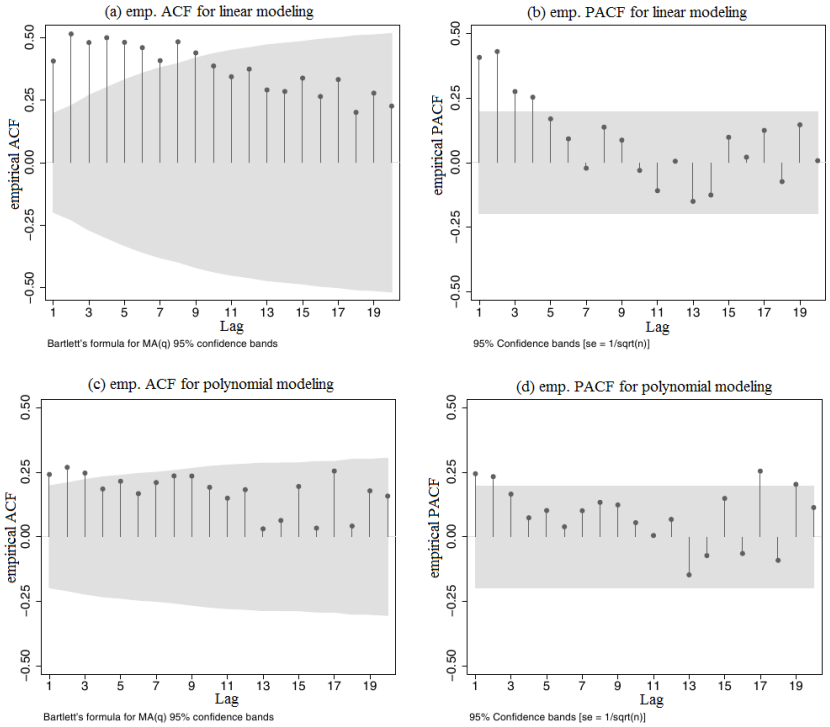

Fig. 2.13: Empirical autocorrelation and partial autocorrelation functions of the residuals of the regressions between log-disposal and price.

ograms also clearly indicate autocorrelation, especially for the model with a linear effect of the price. The autocorrelation functions are just slowly decreasing for increasing lags. The partial autocorrelation functions are quite small not until for lags bigger than two. Hence, the correlograms do not necessarily

suggest a first order autocorrelation, optionally a more complicated autocorrelation structure (second order autocorrelation) should be considered. Here, one advantage of graphical tools over technical tests becomes evident. Indeed, the Durbin-Watson test provides hints on the presence of autocorrelation, but cannot detect the specific type of autocorrelation.

After having succeeded to diagnose autocorrelated error terms with a substantial degree of certainty, the following strategy is nearby: first it has to be analysed, if the correlation can be eliminated (or at least distinctly diminished) by improved model specifications. Fitting a model with autocorrelated error terms should always be the last resort. In the present case a glance on the price development over time (Figure 2.12 (d)) gives the key hint on how to improve the model specification. Obviously the prices can be divided into three periods with different "customary" prices of the coffee brand (the marketing literature refers to *regular prices*). At first, the regular price was around 8.5 DM, then it was permanently reduced to 8 DM, and finally even to 7.5 DM. In each period, several sales campaigns with low prices have been organized. It seems reasonable that inobservance of the time development of the regular price is problematic. A proposal of the marketing literature (compare Leeflang et al., 2000) is thus, to replace the regressor *price* by the quotient of the actual and the current regular price, i.e. to use the new covariate

$$price.rel = \frac{\text{actual price}}{\text{current regular price}},$$

with $0 \leq price.rel \leq 1$. If $price.rel = 1$, the actual price is equal to the current regular price and there is no sales campaign during that week. Figure 2.14 (a) shows the scatter plot between the log-disposal and the new covariate *price.rel*. Again, additionally, the estimated log-disposal based on an ordinary LS-regression is shown, if the influence of *price.rel* is linear (solid line) or cubic (dashed line), respectively. At least the residuals corresponding to the polynomial modeling approach (Figure 2.14 (c)) are approximately uncorrelated. This finding is supported by the Durbin-Watson-test: for linear modeling we obtain $d = 1.5946$ (p-value $= 0.04559$) and for polynomial modeling $d = 1.4853$ (p-value $= 0.9437$). Consequently, the Durbin-Watson-tests provides no hints for autocorrelation in the case of a non-linear, cubic effect of *price.rel*. △

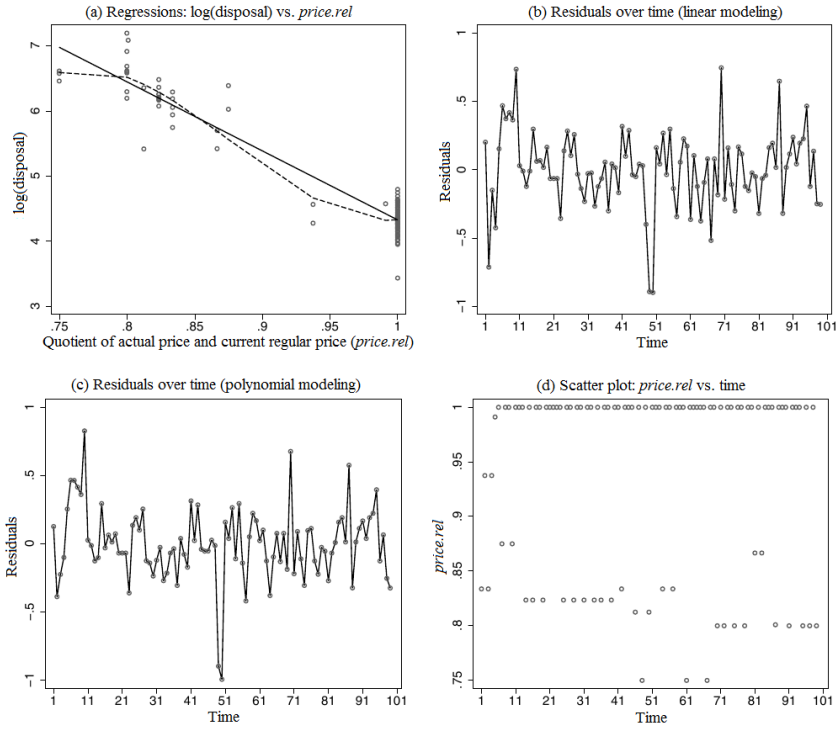

Fig. 2.14: Figure (a): scatter plot between log-disposal and the proportion of the actual and the current regular price *price.rel*, including the estimated log-disposal for linear (solid line) and cubic (dashed line) modeling of the price effect. Figures (b) and (c) show the corresponding residuals over time. Figure (d) shows the proportion *price.rel* over time.

3

Generalized Linear Models

Classical statistical models for regression are generally useful in situations where data are approximately Gaussian - at least after a suitable transformation - and can be explained by some linear structure. These models are usually easy to interpret and the methods are theoretically well understood and investigated. However, in many applications the underlying assumptions may be too stringent and the methods may be misleading in situations where data are clearly non-normal, such as categorical or counted data. Possible examples would be:

- Company bankrupt (yes/no);
- Tumor benign or malignant;
- Individual is unemployed, part-time employed or full-time employed;
- The number of insurance claims, defaults in credit business or individuals infected by a disease in a certain time period.

Statistical modeling aims at providing more flexible model-based tools for data analysis.

The generalized linear model (GLM) has been introduced by Nelder and Wedderburn (1972) as a unifying family of models for non-standard cross-sectional regression analysis with non-normal responses. Its further development had a major influence on statistical modeling in a wider sense, and it has been extended in various ways and for more general situations. An introduction into GLMs can be found for example in Chapter 4 of Fahrmeir et al. (2007) and Chapter 3 of Tutz (2011). A review over large parts of recent advances in statistical modeling that are based on or related to GLMs is provided by Fahrmeir and Tutz (2001). A detailed exposition is found in McCullagh and Nelder (1989).

3.1 Basic Structure of Univariate Generalized Linear Models

In this section we explain the basic structure of univariate GLMs, following Tutz (2011) and Fahrmeir and Tutz (2001). The classical linear model for (ungrouped) normal response and deterministic covariates, which has been analysed in Chapter 2, is

$$y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the design vector $\mathbf{x}_i$ is an appropriate function of certain covariates. These covariates can be metric or binary ($\rightarrow$ the function is the identity), but can also be transformed or extended. For example, for multi-categorical covariates or a mixture of metric and qualitative variables, dummy variables have to be included, or if a smoothing spline representation is chosen for certain covariates, the covariate is evaluated on several basis functions. The error terms ε_i are assumed to be i.i.d., in most cases it is even postulated

$$\varepsilon_i \sim N(0, \sigma^2).$$

This model can be rewritten in a form that leads to GLMs in a natural way: the observations y_i are independent and normally distributed,

$$y_i \sim N(\mu_i, \sigma^2), \qquad i = 1, \ldots, n, \tag{3.1.1}$$

with $\mu_i = E[y_i]$. The means μ_i are given by the linear combinations $\mathbf{x}_i^\intercal \boldsymbol{\beta}$,

$$\mu_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}, \qquad i = 1, \ldots, n. \tag{3.1.2}$$

If covariates are stochastic, we assume the pairs $(y_i, \mathbf{x}_i)$ to be i.i.d.. Then the model is understood conditionally, i.e., the density from (3.1.1) is the conditional density of y_i given $\mathbf{x}_i$, and the y_i are conditionally independent.

Now we are able to give the definition of GLMs, where the preceding assumptions are relaxed in the following way:

(i) *Random component/distributional assumption:*
Given $\mathbf{x}_i$, the y_i are (conditionally) independent observations from a simple exponential family with (conditional) expectation $E[y_i|\mathbf{x}_i] = \mu_i$. This family has a probability density function or mass function of the form

$$f(y_i|\mathbf{x}_i, \theta_i, \phi, w_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\right), \tag{3.1.3}$$

where θ_i is the natural parameter of the family, ϕ is a scale or dispersion parameter, w_i is a weight[1] and $b(\cdot)$ and $c(\cdot)$ are specific functions corre-

[1] The introduction of suitable weights allows to treat the cases of individual and grouped data consistently. One uses $w_i = 1, i = 1, \ldots, n$, for ungrouped data and $w_i = n_i, i = 1, \ldots, G$ for grouped data with individual group sizes n_i, if the *average* is considered as response (or $w_i = 1/n_i$, if the *sum* of individual responses is considered).

sponding to the type of the family. A detailed treatise of the exponential family can be found e.g. in Pruscha (2000).

(ii) *Systematic component/structural assumption:*
The systematic component is determined by two structuring components, the linear term and the link between response and covariates. The linear part, which gives the GLM its name, specifies that the variables $\mathbf{x}_i$ enter the model in linear form by specifying the linear predictor

$$\eta_i = \mathbf{x}_i^\intercal \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the unknown parameter vector of dimension $p+1$. The relation between the linear part and the conditional expectation $\mu_i = E(y_i|\mathbf{x}_i)$ is determined by the transformation

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i^\intercal \boldsymbol{\beta}) \tag{3.1.4}$$

or

$$\eta_i = g(\mu_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}, \tag{3.1.5}$$

respectively, where h is a known one-to-one (sufficiently smooth) *response function*, and g is the so-called *link function*, i.e. the inverse of h.

Equations (3.1.4) and (3.1.5) reflect equivalent ways of specifying how the mean of the response variable is linked to the linear predictor. The response function h in (3.1.4) shows how the linear predictor has to be transformed to determine the mean response. Equation (3.1.5) shows for which transformation of the mean the model becomes linear.

Example 3.1.1 (Logistic model). A simple example is the logistic model where the mean μ_i corresponds to the probability of success π_i. One has the two forms

$$\pi_i = \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})},$$

yielding the response function $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$ and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^\intercal \boldsymbol{\beta},$$

where the link function $g = h^{-1}$ is specified by $g(\mu) = g(\pi) = log(\pi/(1-\pi)) = log(\mu/(1-\mu))$. The form of the link function, which corresponds to (3.1.5), shows that a GLM is a linear model for the transformed mean. Additionally, it is assumed that the response follows a distribution that belongs to the simple exponential family.

△

Altogether, a specific GLM is fully determined by three components, namely

- the type of the exponential family, which specifies the particular distribution of $y_i|\mathbf{x}_i$,

- the form of the linear predictor, i.e. the selection and coding of covariates,
- and finally, the response or link function.

Before considering the various models which fit into this framework, we want to remark on simple exponential families.

Remark 3.1.2. In simple exponential families the natural parameter is linked to the mean of the distribution. Hence, the parameter θ_i may be seen as $\theta_i = \theta(\mu_i)$, where θ is considered as a transformation of the mean. Certain distributions usually use a specific notation for parameters, like for example μ for the normal or π for the binomial distribution. These parameters determine the mean. The corresponding function $\mu_i = \mu(\cdot)$ is a function of parameters, which characterize a specific distribution, e.g. $\mu_i = \mu(\lambda_i) = \lambda_i$ for the Poisson distribution parameter λ_i. When considering the function $\theta(\mu_i)$, the dependence on these parameters is given by $\theta(\mu(.))$.

In general, the choice of the link function depends on the distribution of the response. For example, if y is non-negative, a link function is appropriate, which specifies non-negative means without contemporaneously restricting the parameters. For each distribution within the simple exponential family there exists one link function that has some technical advantages, the so-called canonical link:

Definition 3.1.3 (Canonical link). *A link function is called* canonical *or* natural link *function, if it relates the natural parameter θ directly to the linear predictor:*

$$\theta(\mu) = \eta = \mathbf{x}^\intercal \boldsymbol{\beta},$$

i.e., $g(\mu) \equiv \theta(\mu)$.

Note that though the canonical link functions lead to models with convenient mathematical and statistical properties, this should not be the main criterion for choosing them. In particular applications, non-natural link functions may often be more appropriate.

Following Tutz (2011), we will give an overview of GLMs with continuous responses in the next subsection, and afterwards will regard GLMs for discrete responses. For each distribution we will identify the canonical link function.

3.1.1 GLMs for Continuous Responses

We will start with the simplest case of the ordinary linear regression model, which has already been extensively studied in Chapter 2. Next, we will also considered several other distributions for continuous responses, which not necessarily need to have support $\mathbb{R}$.

Gaussian Distribution

We know that the corresponding regression model is usually specified for $i = 1, \ldots, n$, as

$$y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

resulting in normally distributed response variables

$$y_i | \mathbf{x}_i \sim N(\mathbf{x}_i^\intercal \boldsymbol{\beta}, \sigma^2).$$

Alternatively, now we can specify this model in the GLM terminology by

$$y_i | \mathbf{x}_i \sim N(\mu_i, \sigma^2) \qquad \text{and} \qquad \mu_i = \eta_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}.$$

In this case, the link function is the identity link. It can easily be shown that the normal distribution is within the distribution class of the exponential family ($\rightarrow$ see exercises), with the natural parameter θ, the dispersion parameter ϕ and function b given by

$$\theta(\mu) = \mu, \quad b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}, \quad \phi = \sigma^2.$$

As $\theta(\mu) = \mu$, the canonical link is given as $g(\mu) = \eta = \mu$.

The GLM framework provides an extension of the normal linear model by considering alternative link functions. If responses are expected to be positive (i.e. positive mean), more appropriate response functions might be

$$h(\eta) = \eta^2 \qquad \text{or} \qquad h(\eta) = \exp(\eta).$$

Of course, in these scenarios the influence of covariates and consequently the interpretation of parameters is quite different from that in the linear model. In contrast to the classical linear model

$$\mu_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip},$$

where the change of a certain covariate x_j by one unit results in an *additive* effect of β_j on the expectation, for the modified relationship (link)

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}) = e^{\beta_0} e^{\beta_1 x_{i1}} \cdot \ldots \cdot e^{\beta_p x_{ip}},$$

a change of x_j by one unit now has a *multiplicative effect* on μ by the factor e^{β_j}, since $e^{(x_j+1)\beta_j} = e^{x_j \beta_j} e^{\beta_j}$. In Figure 3.1 the regression model for normally distributed response is illustrated for a single explanatory variable and two different link functions. The left graphic in Figure 3.1 shows the linear model, whereas in the right graphic the log-link model is depicted. In both diagrams, the straight line (left) and the curve (right), respectively, show the means. Additionally, the corresponding normal densities of the response variable are shown at three distinct x-values.

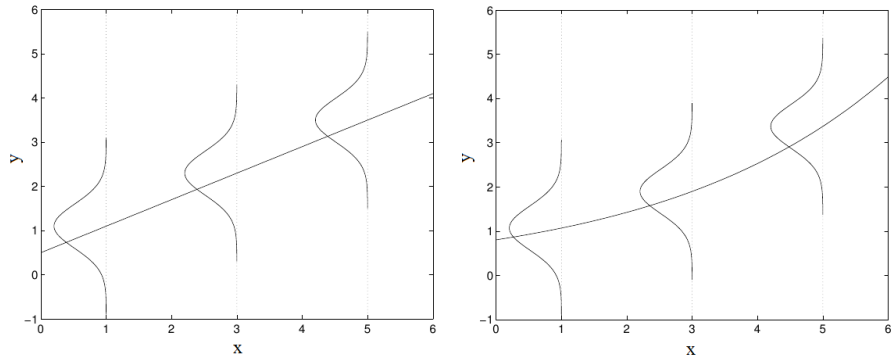

Fig. 3.1: Regression for normally distributed response with the identity link (left) and log-link (right)

Exponential Distribution

In cases where responses are strictly non-negative, for example in the analysis of durations up to a certain event or survival times of certain individuals, the normal distribution model is rarely adequate. A classical distribution, which is often used when time is the response variable, is the exponential distribution

$$f(y) = \lambda e^{-\lambda y} = \exp(-\lambda y + \log(\lambda)), \qquad y \geq 0.$$

With $\theta = -\lambda, \phi = 1$, $b(\theta) = -\log(-\theta)$ and $c(y, \phi, w) \equiv 0$, the exponential distribution is of the simple exponential family. It is well known and can be easily derived that for an exponentially distributed random variable with parameter λ, i.e. $Y \sim Exp(\lambda)$, it holds that $E[Y] = 1/\lambda$ and $Var(Y) = 1/\lambda^2$. So in contrast to the normal distribution, here the variance increases with increasing expectation. Thus, although there is a fixed link between expectation and variance, this distribution model captures an essential characteristic that is often found in real data sets. Here, the canonical link function is given by

$$g(\mu) = -\frac{1}{\mu} \qquad \text{or} \qquad h(\eta) = -\frac{1}{\eta}.$$

Since $\mu > 0$, the linear predictor is restricted to $\eta = \mathbf{x}^\intercal \boldsymbol{\beta} < 0$. As this condition implies severe restricions on $\boldsymbol{\beta}$, often a more adequate link function is given by the log-link

$$g(\mu) = \log(\mu) \qquad \text{or} \qquad h(\eta) = \exp(\eta),$$

yielding $\mu = \exp(\eta) = \exp(\mathbf{x}^\intercal \boldsymbol{\beta})$.

Gamma-Distributed Responses

Since the exponential distribution is a one parameter distribution, its flexibility is rather restricted. According to Tutz (2011), a more flexible distribution

model for non-negative responses, like for example durations or insurance claims, is the gamma-distribution. With expectation $\mu > 0$ and shape parameter $\nu > 0$, the density function of the gamma-distribution is given by

$$\begin{aligned} f(y) &= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right) \\ &= \exp\left(\frac{-(1/\mu)y - \log(\mu)}{1/\nu} + \nu \log(\nu) + (\nu - 1)\log(y) - \log(\Gamma(\nu))\right). \end{aligned}$$

With respect to the exponential family parametrization, we obtain the dispersion parameter $\phi = 1/\nu$ and $\theta(\mu) = -1/\mu$ as well as $b(\theta) = -\log(-\theta)$. We find that in comparison to the exponential distribution only the dispersion parameter is different: while it is fixed by $\phi = 1$ for the exponential distribution, it is more flexible for the gamma-distribution. Figure 3.2 illustrates, how ν controls the shape of the distribution. For $0 < \nu \leq 1$, $f(y)$ decreases monotonically, whereas for $\nu > 1$ the density has a mode at $y = \mu - \mu/\nu$ and is positively skewed. Usually, for a gamma-distributed random variable Y one uses the abbreviation $Y \sim \Gamma(\nu, \alpha)$, with the parametrization $\alpha = \nu/\mu$. Using the expectation as parameter, one can write $\Gamma(\nu, \frac{\nu}{\mu})$.

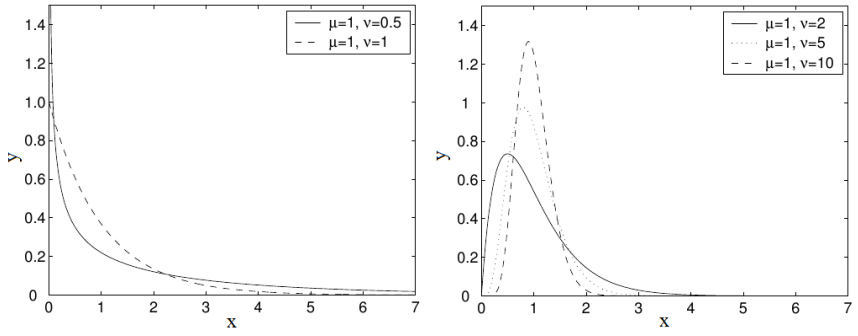

Fig. 3.2: Gamma distributions for different choices of μ and ν.

The variance of a gamma-distributed random variable Y is given by $Var(Y) = \nu/\alpha^2 = \mu^2/\nu$. Thus, the variance depends strongly on the expectation. This effect is often found in practice. The dependence may be characterized by the coefficient of variation, which is defined by $c = \sigma/\mu$ and is a specific measure of variation, scaling the standard deviation by the expectation. For gamma-distributions the coefficient of variation is given by $c = \sigma/\mu = \mu/(\sqrt{\nu}\mu) = 1/\sqrt{\nu}$. For the link function nothing has changed in comparison to the exponential distribution and one usually chooses $g(\mu) = \log(\mu)$ or $h(\eta) = \exp(\eta)$, respectively.

Figure 3.3 shows the exponential and the gamma-regression model for the log-link function. It is seen how the shifting of the mean along the logarithmic function changes the form of the distribution. In contrast to the normal model, where densities are simply shifted, for gamma-distributed (and hence also for exponentially distributed) responses the form of the densities depends on the mean. Moreover, we find that densities have support $\mathbb{R}^+$. For the normal model shown in Figure 3.1 the log-link ascertains that the mean is positive, but nevertheless the model also allows for negative values. Thus, for strictly positive valued response a normal model would be doubtful. Of course, the adequacy of the model depends on the covariates x, which are incorporated into the model, and the variance of the response.

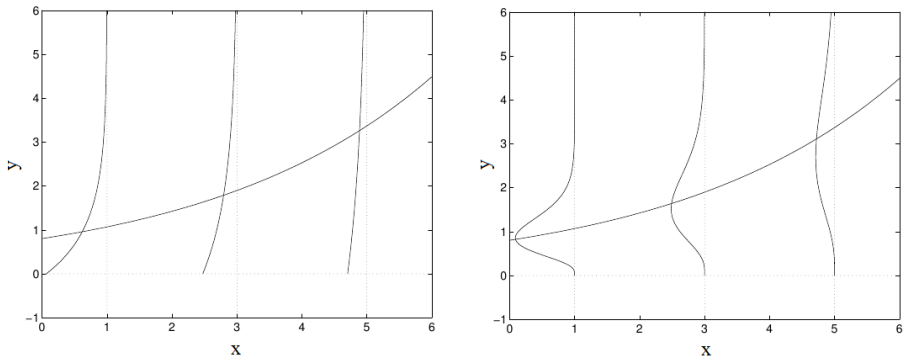

Fig. 3.3: Exponential (left) and gamma-distributed regression model with $\nu = 5$ (right), both with log-link.

Inverse Gaussian Distribution

An alternative distribution with strictly non-negative response, which can be used to model e.g. duration, is the inverse Gaussian distribution. In its usual form it is given by the following density function

$$f(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2\mu^2 y}(y-\mu)^2\right), \qquad y > 0,$$

which we abbreviate by $IG(\mu, \lambda)$, with determining parameters $\mu, \lambda > 0$. One can easily transform the density function into

$$f(y) = \exp\left(\frac{y(-1/(2\mu^2)) + 1/\mu}{1/\lambda} - \frac{\lambda}{2y} - \frac{1}{2}\log(2\lambda\pi) - \frac{3}{2}\log(y)\right),$$

yielding

$$\begin{aligned}
\theta &= -\frac{1}{2\mu^2}, \\
b(\theta) &= -\sqrt{-2\theta} = -\frac{1}{\mu}, \\
\phi &= \frac{1}{\lambda}, \\
c(y, \phi, w) &= -\frac{1}{2y\phi} - \frac{1}{2}\log(2\pi/\phi) - \frac{3}{2}\log(y).
\end{aligned}$$

The canonical link, for which $\theta(\mu) = \eta$ holds, is then given by

$$g(\mu) = -\frac{1}{2\mu^2} \qquad \text{or} \qquad h(\eta) = \frac{1}{\sqrt{-2\eta}},$$

which implies the severe restriction $\eta = \mathbf{x}^\intercal\boldsymbol{\beta} > 0$. A link function avoiding these problems is again the log-link function $g(\mu) = \log(\mu)$ or $h(\eta) = \exp(\eta)$, respectively, compare Figure 3.4.

The inverse Gaussian distribution has several interesting properties, for example that the ML-estimates of the mean μ and the dispersion $1/\lambda$, given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{and} \qquad \frac{1}{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{y_i} - \frac{1}{\bar{y}}\right),$$

are independent. This is similar to the normal distribution, for which the sample mean and the sample variance are also independent.

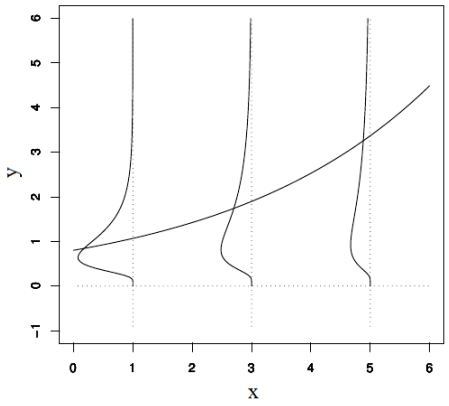

Fig. 3.4: Inverse Gaussian distributed regression model with $\lambda = 3$ and log-link.

3.1.2 GLMs for Discrete Responses

Following Fahrmeir and Tutz (2001), in this section we provide an introduction into GLMs for several scenarios with discrete responses, which are commonly used. We will start with the simplest case of a discrete response, when only "success" or "failure" is measured, encoded with $y \in \{0, 1\}$.

Models for Binary Data

The corresponding distribution is the Bernoulli distribution, which for $Y \in \{0, 1\}$ has the following well-known probability mass function

$$f(y) = \pi^y (1 - \pi)^{1-y}.$$

Here, $\pi = P(Y = 1)$ denotes the probability of "success". It can be shown ($\rightarrow$ see exercises) that with $\mu = \pi$ this probability mass function belongs to the exponential family class with $\theta(\pi) = \log(\frac{\pi}{1-\pi}), b(\theta) = \log(1 + \exp(\theta)) = -\log(1-\pi)$ and $\phi = 1$. As π is a probability in $[0, 1]$, it would be a reasonable choice to relate π to the linear predictor η by

$$\pi = F(\eta),$$

where F is any strictly monotone distribution function on the whole real axis, with the preferable consequence that no restrictions on η and hence, $\boldsymbol{\beta}$ have to be imposed. We will now give an overview of the most common choices for F.

Probit model

Perhaps the most nearby choice for F would be the distribution function of the standard normal distribution, which is usually denoted by Φ. This results in the so-called probit model, which is defined by

$$\pi = \Phi(\eta) = \Phi(\mathbf{x}^\intercal \boldsymbol{\beta}).$$

This model imposes no restrictions on η, but has the disadvantage that it is computationally more demanding, if corresponding likelihoods need to be computed.

Logit model

The logit model corresponds to the canonical link function

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \eta,$$

with the logistic distribution function

$$\pi = h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

as resulting response function. The logistic distribution function also has support on the entire real axis and is symmetric, but has somewhat heavier tails compared to the standard normal, compare Figure 3.5, where the values of η are plotted against π. The Figure also shows that apart from values for π close to 0 or 1, which correspond to the tails, fits using probit or logit models are generally quite similar. As the logistic distribution function is easier to compute than the standard normal (and as it reflects the canonical link!), the logit model is often preferred in practice.

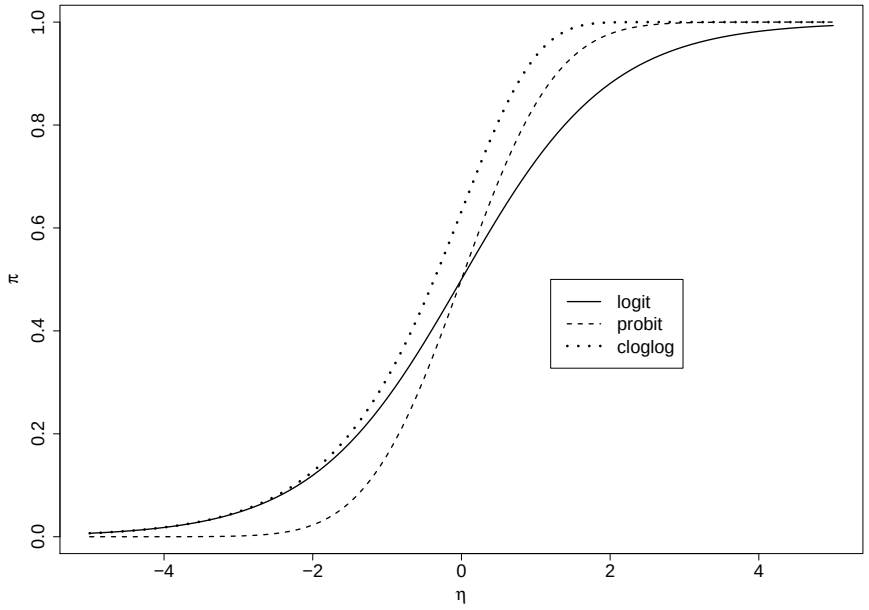

Fig. 3.5: Most common response functions for binary responses.

Complementary log-log model

This model has the link function

$$g(\pi) = \log(-\log(1 - \pi)),$$

and the so-called extreme minimal-value distribution function

$$h(\eta) = 1 - \exp(-\exp(\eta))$$

as corresponding response function, which is non-symmetric, close to the logistic function for small π, but with a considerably less heavy right tail.

Another link function that is sometimes useful and that should be mentioned in this context is the complementary log model, see Fahrmeir and Tutz (2001) for more details or Piegorsch (1992) for further motivations and applications.

At a glance on Figure 3.5 it seems that the response functions are quite different, but the comparison done there is quite unfair, as the models should be compared for an appropriate scaling of the predictor η, e.g. by transforming F so that the mean and variance according to the different distribution functions are the same. Recall that a standardized version of a distribution function with mean μ and standard deviation σ is obtained via $\tilde{F}(u) = F(\frac{u-\mu}{\sigma})$. Table 3.1 summarizes the means and variances corresponding to the distribution functions used in the probit, logit and cloglog model.

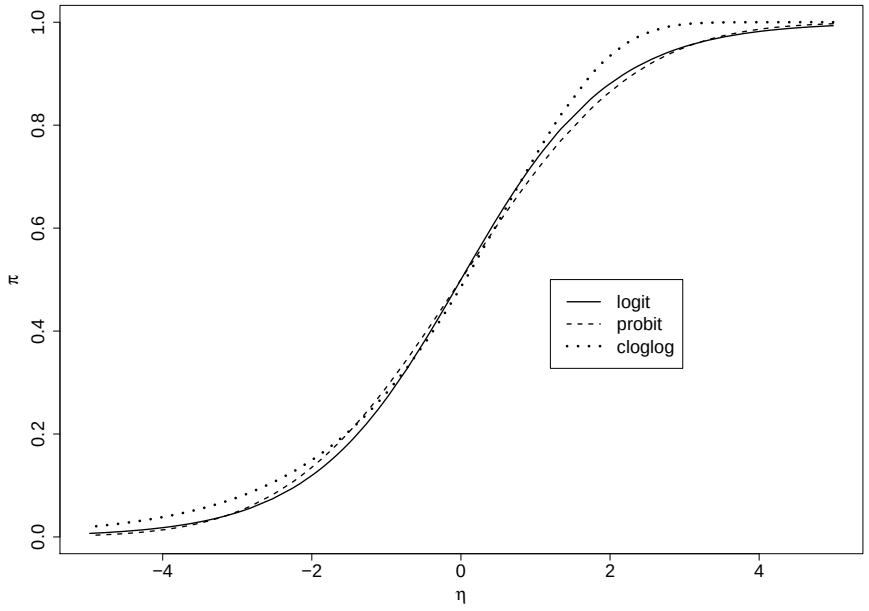

Fig. 3.6: Most common response functions for binary responses adjusted to the logistic function (such that after linear transformation all distributions have mean zero and variance $\pi^2/3$.

Figure 3.6 displays the three response functions with η adjusted such that all three distributions have the mean and variance of the logistic distribution. In contrast to Figure 3.5, the logistic and probit response functions, which

Response function F	Mean	Variance
probit	0.0	1
logit	0.0	$\pi^2/3$
cloglog	-0.5772	$\pi^2/6$

Table 3.1: Means and variances corresponding to the distribution functions used in the probit, logit and cloglog model.

now have both variance $\pi^2/3$, are almost indistinguishable. Therefore, fits of probit and logit models are generally quite similar after the adjustment of η, which is often implicitly done in estimation. However, the cloglog function is steeper than the other two, even after adjustment. Thus, for small values of η it approaches 0 more slowly, and if η goes to infinity it approaches 1 faster than the logistic and adjusted probit functions.

Remark 3.1.4 (Interpretation of parameters). In general, the interpretation of parameters becomes more difficult for direct interpretation of covariate effects on a binary response y. For the logit model, e.g., a linear model for the "log odds"[2] is assumed. Consequently, the interpretation of covariate effects on the log odds is equivalent to the interpretation of covariate effects on the expectation $\mu = E[y]$ in the linear model. Let a simple linear predictor be given of the form $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, with

$$\frac{\pi}{1-\pi} = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2).$$

Obviously, the exponentials of covariate effects have a multiplicative effect on the relative risk. However, this interpretation is not possible for other link functions. Fahrmeir and Tutz (2001) suggest to break up the interpretation in two stages:

1. Interpret covariate effects on the linear predictor $\mathbf{x}^\intercal\boldsymbol{\beta}$ in the same way as in linear models.
2. Transform this linear effect on η into a non-linear effect on π with the aid of a graph of the response function $\pi = h(\eta)$, as given in Figures 3.5 and 3.6.

Extension: Binary Models as Threshold Models of Latent Linear Models

As presented so far, our models for binary responses seem to be ad hoc specifications with some useful properties. However, Fahrmeir and Tutz (2001) show how all these models can be derived as threshold models, where binary

[2] i.e. the logarithm of the odds $\pi/(1-\pi)$, which are defined as the proportion of the probability for a success and the probability for a failure. In a medical context, the odds are often called "relative risk".

responses y are based on a latent continuous variable y^* that obeys a linear model

$$y^* = \alpha_0 + \mathbf{w}^\mathsf{T}\boldsymbol{\alpha} + \sigma\varepsilon,$$

where ε is distributed according to $F(\cdot)$, e.g. a logistic or standard normal distribution function, and σ is a scale parameter. The relation between y and y^* is given by

$$y = \begin{cases} 1, & y^* \leq \tau, \\ 0, & y^* > \tau, \end{cases}$$

with a threshold value τ. From this assumption we obtain

$$P(y = 1) = P(\alpha_0 + \mathbf{w}^\mathsf{T}\boldsymbol{\alpha} + \sigma\varepsilon \leq \tau) = F\left(\frac{\tau - \alpha_0 - \mathbf{w}^\mathsf{T}\boldsymbol{\alpha}}{\sigma}\right).$$

Defining

$$\boldsymbol{\beta} = \left(\frac{\tau - \alpha_0}{\sigma}, \frac{\boldsymbol{\alpha}^\mathsf{T}}{\sigma}\right)^\mathsf{T}, \qquad \mathbf{x}^\mathsf{T} = (1, -\mathbf{w}^\mathsf{T}),$$

one obtains the general model $\pi = F(\eta)$. Note that the covariate effects $\boldsymbol{\alpha}$ of the underlying linear model can be identified only up to the common but generally unknown factor $1/\sigma$, and that the global mean parameter α_0 cannot be identified at all, as τ is usually unknown. Consequently, not absolute values $\boldsymbol{\beta}$ are meaningful, but relative values, e.g. β_1/β_2.

Models for Binomial Data

If experiments which distinguish only between "success" and "failure" are repeated independently, often the number of successes or its proportion is of interest and hence used as response variable. For m independently repeated Bernoulli trials one obtains the binomial distributed response $\tilde{y} \in \{0, \ldots, m\}$. The probability mass function has the parameters m and the probability π of success in one trial and is given for $\tilde{Y} \in \{0, \ldots, m\}$ by

$$\begin{aligned} f(\tilde{y}) &= \binom{m}{\tilde{y}} \pi^{\tilde{y}}(1-\pi)^{m-\tilde{y}} \\ &= \exp\left\{\frac{\frac{\tilde{y}}{m}\log(\frac{\pi}{1-\pi}) + \log(1-\pi)}{1/m} + \log\binom{m}{\tilde{y}}\right\}. \end{aligned}$$

If we consider the proportion of successes $y = \tilde{y}/m$ as response instead of the number of successes $\tilde{y}$, we find that the probability mass function belongs to the exponential family with $\mu = E[\tilde{y}/m] = \pi, \theta(\pi) = log(\pi/(1-\pi)), b(\theta) = \log(1 + \exp(\theta)) = -\log(1-\pi)$ and $\phi = 1/m$. The variance is given by $Var(\tilde{y}/m) = \pi(1-\pi)/m$.

For this reason we consider the proportion $y = \tilde{y}/m$ as response, with outcomes in the set $\{0, 1/m, 2/m, \ldots, 1\}$. The distribution of y has the usual binomial form

$$f(y) = \binom{m}{my} \pi^{my}(1-\pi)^{m-my} = \binom{m}{\tilde{y}} \pi^{\tilde{y}}(1-\pi)^{m-\tilde{y}},$$

but with $y \in \{0, 1/m, 2/m, \ldots, 1\}$. It consists of a simple rescaling of the number of successes to proportions and hence, reflects a changing of the support, which is why it is called the *scaled binomial distribution*.

At this point we want to mention an aspect concerning the dispersion parameter, where the binomial distribution differs from all other distributions that have been considered so far. For different observations $y_i = \tilde{y}_i/m_i$, denoted by the subscript i, one has the dispersion parameters $\phi_i = 1/m_i$, m_i denoting the number of replications, which is actually in contrast to our basic formulation of the exponential family class in (3.1.3), where we simply use the notation ϕ. Since m_i is fixed, the dispersion is fixed (and known), but may depend on the observations since the number of replications may vary across observations. So in contrast to the other distributions, the dispersion depends on i.

An alternative way of looking at binomial data is by considering them as grouped observations, i.e. a grouping of repeated Bernoulli observations. For the special case $m = 1$ the ordinary binomial and rescaled binomial distribution are equivalent. Besides, the Bernoulli case may of course be treated as a special case of the binomial case. Consequently, link and response function are treated in the same way as in the binary case.

Poisson Models for Count Data

Discrete responses appear in many applications, often in the form of count data, e.g., as the number of certain events within a fixed period of time (insurance claims, accidents, deaths, births, etc.) or in the form of frequencies in the cells of contingency tables. Under certain circumstances, such data may be approximately modelled by models for normal data. If only a small number of response values $0, 1, \ldots, q$ is observed, models for multi-categorial data could be used, see Tutz (2000).

But in general, the Poisson distribution or some modification should be the first choice. For integer values $y \in \{0, 1, \ldots\} = \mathbb{N}_0$ and $\lambda > 0$, the corresponding probability mass function is given by

$$f(y) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

It can be easily shown ($\rightarrow$ see exercises) that this distribution belongs to the exponential family. For the expectation $\mu = \lambda$ the parameters are identified by $\theta(\mu) = \log(\mu), b(\theta) = \exp(\theta)$ and $\phi = 1$.

A sensible choice of the link function should guarantee $\lambda > 0$. Thus, a widely used link function is the canonical link, which is the log link, yielding

$$\log(\lambda) = \mathbf{x}^\intercal \boldsymbol{\beta} \qquad \text{or} \qquad \lambda = \exp(\mathbf{x}^\intercal \boldsymbol{\beta}), \qquad \text{respectively.}$$

In Figure 3.7 the distribution is shown for three distinct x-values. It is seen that for varying means the shape of the distribution changes. While the distribution of the response is skewed for low means, it is nearly symmetric for large values of the mean.

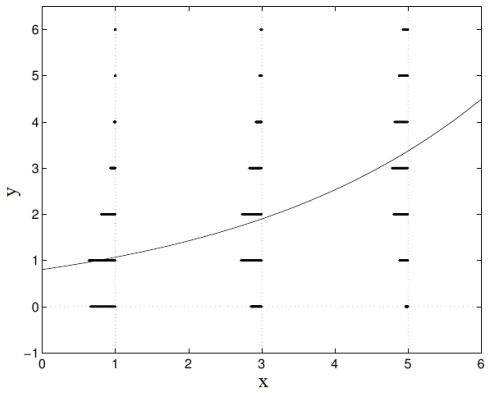

Fig. 3.7: Poisson regression with log link

Negative Binomial Models for Count Data

An alternative distribution for count data with support $y \in \mathbb{N}_0$ is the negative binomial distribution with probability mass function

$$f(\tilde{y}) = \frac{\Gamma(\tilde{y}+\nu)}{\Gamma(\tilde{y}+1)\Gamma(\nu)} \left(\frac{\nu}{\tilde{\mu}+\nu}\right)^{\nu} \left(\frac{\tilde{\mu}}{\tilde{\mu}+\nu}\right)^{\tilde{y}}, \qquad \tilde{y} = 0, 1, \ldots \tag{3.1.6}$$

with $\nu, \tilde{\mu} > 0$. There are two major ways to motivate this distribution.

(i) The probability mass function may be seen as a mixture of Poisson distributions in the so-called gamma-Poisson-model. The model assumes that the parameter λ of the Poisson distribution is itself a random variable which is gamma-distributed with $\lambda \sim \Gamma(\nu, \frac{\nu}{\tilde{\mu}})$ with shape parameter ν and expectation $\tilde{\mu}$. The conditional distribution of $\tilde{y}$ is then assumed to follow the Poisson distribution $\tilde{y}|\lambda \sim Po(\lambda)$, resulting in the marginal distribution (3.1.6) for $\tilde{y}$. The assumption that the total counts result from heterogeneous sources with individual parameters increases the flexibility, so the negative binomial model is an attractive alternative to the Poisson model. From the variance of the gamma distribution, which is given by $\tilde{\mu}^2/\nu$, it follows that for $\nu \to \infty$ the mixture of Poisson distributions reduces to just one Poisson distribution as the limiting case and we are

back in the simple Poisson model. Expectation and variance of the negative binomial are given by

$$E[\tilde{y}] = \tilde{\mu}, \qquad Var(\tilde{y}) = \tilde{\mu} + \tilde{\mu}^2/\nu.$$

So it is seen that $\lim_{\nu\to\infty} E[\tilde{y}] = Var(\tilde{y})$, which is in accordance with the Poisson distribution. The parameter ν may be seen as an additional dispersion parameter which yields larger variation for small values. Thus, $1/\nu$ is an indicator for the amount of variation.

(ii) For integer valued ν the negative binomial has the simpler form

$$f(\tilde{y}) = \binom{\nu + \tilde{y} - 1}{\nu - 1} \pi^{\nu} (1-\pi)^{\tilde{y}}, \qquad \tilde{y} = 0, 1, \ldots \tag{3.1.7}$$

where $\pi = \nu/(\tilde{\mu} + \nu) \in (0, 1)$ may be seen an alternative parameter with simple interpretation. Let us consider independent Bernoulli variables with success probability π and let X be the number of trials that are necessary until the ν-th success has occurred, ($\nu \in \mathbb{N}$). We define the events $A :=$"exactly $\nu - 1$ times success in the first $n - 1$ trials" and $B :=$"the n-th trial is a success". Due to the Bernoulli-chain, we get

$$P(X = n) = P(A \cap B) = P(A)P(B).$$

The probabilities of both events can be easily derived

$$P(A) = B(\nu - 1|n - 1, \pi) = \binom{n-1}{\nu-1} \pi^{\nu-1} (1-\pi)^{n-\nu}$$

and

$$P(B) = \pi.$$

Hence, we get

$$\begin{aligned} P(X = n) = P(A)P(B) &= \binom{n-1}{\nu-1} \pi^{\nu-1} (1-\pi)^{n-\nu} \pi \\ &= \binom{n-1}{\nu-1} \pi^{\nu} (1-\pi)^{n-\nu}, \; n = \nu, \nu+1, \ldots \end{aligned}$$

Consequently, the negative binomial distribution in the parametrization (3.1.7) reflects the probability for the number of trials which in addition to ν are necessary in order to obtain ν hits. The most familiar case is $\nu = 1$, resulting in the *geometric distribution*. In this case, $\tilde{y}$ (plus one) reflects the number of trials, which are necessary until the first hit occurs. It is the standard distribution for example in fertility studies, where the number of trials until conception is modeled.

From (3.1.6) we obtain

$$f(\tilde{y}) = \exp\left\{\frac{\log(\pi) + (\tilde{y}/\nu)\log(1-\pi)}{1/\nu} + \log\left(\frac{\Gamma(\tilde{y}+\nu)}{\Gamma(\tilde{y}+1)\Gamma(\nu)}\right)\right\}.$$

For fixed ν this is a simple exponential family for the scaled response $y = \tilde{y}/\nu$ and dispersion $\phi = 1/\nu$. Since we consider $\tilde{y}/\nu$ as the response, we have expectation $\mu = E[y] = \tilde{\mu}/\nu$ and hence $\theta(\mu) = \log(1-\pi) = \log(\mu/(\mu+1))$ and $b(\theta) = \log(1-\exp(\theta))$. Then the corresponding canonical link is

$$\log\left(\frac{\mu}{\mu+1}\right) = \eta \qquad \text{or} \qquad \mu = \frac{\exp(\eta)}{1-\exp(\eta)}.$$

The canonical link may cause problems for $\eta = 0$, since for $\eta \to 0$ we get $\mu \to \infty$. Again, for the log link

$$\log(\mu) = \mathbf{x}^\intercal\boldsymbol{\beta} \qquad \text{or} \qquad \mu = \exp(\mathbf{x}^\intercal\boldsymbol{\beta}),$$

respectively, the linear predictor $\eta = \mathbf{x}^\intercal\boldsymbol{\beta}$ needs not be restricted.

Note that as the negative binomial response $y = \tilde{y}/\nu$ is scaled by ν, this parameter has to be fixed in advance, when a GLM is fitted.

3.1.3 Means and Variances

One of the key assumptions in GLMs is that the distribution of the responses is in the exponential family

$$f(y_i|\theta_i, \phi, w_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}w_i + c(y_i, \phi, w_i)\right).$$

In the previous sections several examples have been given for the dependence of the natural parameters θ_i on μ_i and the parameters which characterize the distribution. For example in the Bernoulli case one obtains $\theta_i = \theta(\mu_i) = \log(\mu_i/(1-\mu_i))$ and since $\mu_i = \pi_i$, we have $\theta_i = \log(\pi_i/(1-\pi_i))$.

A nice property of the exponential family is that the mean is directly related to the function $b(\theta_i)$ in the form

$$E[y_i|\mathbf{x}_i] = \mu_i = \frac{\partial b(\theta_i)}{\partial\theta} = b'(\theta_i) \tag{3.1.8}$$

For the variance one obtains

$$Var(y_i|\mathbf{x}_i) = \sigma_i^2 = \frac{\phi}{w_i}\frac{\partial^2 b(\theta_i)}{\partial\theta^2} = \frac{\phi b''(\theta_i)}{w_i}. \tag{3.1.9}$$

A proof of these results can be found for example in Pruscha (2000). It is seen that the variances are composed from the dispersion parameter ϕ, the weights w_i and the so-called variance function $v(\mu_i) = b''(\theta_i)$, which is a function of the mean due to Equations (3.1.8) and (3.1.9). These equations hold for every GLM and strictly link the mean μ_i and the variance since both are based on

derivatives of the function $b(\theta)$. An overview of the variance functions for the distributions of the exponential family can be found in Table 3.2 (b). Some flexibility for the link between mean and variance is provided by the dispersion parameter ϕ. However, the latter is not always an additional parameter, as it is fixed e.g. for the exponential, Bernoulli, binomial and Poisson distribution. For the normal, Gamma, negative binomial and inverse Gaussian on the contrary it is a parameter which may be chosen to fit the data. Note here again that for the (rescaled) binomial model, which may be considered as replications of Bernoulli variables, the dispersion parameter ϕ exceptionally depends on i, namely $\phi_i = 1/m_i$, which is in contrast to our basic formulation of the exponential family class in (3.1.3), where we simply use the notation ϕ.

Generalized linear model (GLM)

Distribution assumption

For given covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n$, the responses y_i are (conditional) independent and their (conditional) density belongs to the exponential family with

$$f(y_i|\theta_i, \phi, w_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\right).$$

θ_i is the natural parameter, ϕ is the dispersion parameter (usually independent of i). For the weights we have $w_i = 1$ for ungrouped data and $w_i = n_i, i = 1, \ldots, G$ for grouped data with individual group sizes n_i, if the average is considered as response (or $w_i = 1/n_i$, if the sum of individual responses is considered) and $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of the family.
For $E[y_i|\mathbf{x}_i] = \mu_i$ and $Var(y_i|\mathbf{x}_i) = \sigma_i^2$ one obtains

$$E[y_i|\mathbf{x}_i] = \mu_i = b'(\theta_i), \qquad Var(y_i|\mathbf{x}_i) = \sigma_i^2 = \frac{\phi b''(\theta_i)}{w_i}.$$

Systematic component

The (conditional) expectation μ_i is linked to the linear predictor $\eta_i = \mathbf{x}_i^\intercal\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ via

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i^\intercal\boldsymbol{\beta}) \qquad \text{or} \qquad \eta_i = g(\mu_i) = \mathbf{x}_i^\intercal\boldsymbol{\beta},$$

respectively, where

- h is a (unique and twice differentiable) *response function*
- g is the *link function*, i.e. the inverse $g = h^{-1}$ of h

$$f(y_i|\theta_i, \phi, w_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\right),$$

(a) Components of the exponential family

Distribution	Notation	μ_i	$\theta(\mu_i)$	$b(\theta_i)$	ϕ
Normal	$N(\mu_i, \sigma^2)$	μ_i	μ_i	$\theta_i^2/2$	σ^2
Exponential	$Exp(\lambda_i)$	$1/\lambda_i$	$-1/\mu_i$	$-\log(-\theta_i)$	1
Gamma	$\Gamma(\nu, \frac{\nu}{\mu_i})$	μ_i	$-1/\mu_i$	$-\log(-\theta_i)$	$\frac{1}{\nu}$
Inverse Gaussian	$IG(\mu_i, \lambda)$	μ_i	$-1/2\mu_i^2$	$-\sqrt{-2\theta_i}$	$\frac{1}{\lambda}$
Bernoulli	$Ber(1, \pi_i)$	π_i	$\log \frac{\mu_i}{1-\mu_i}$	$\log(1+\exp(\theta_i))$	1
Binomial (rescaled)	$B(m_i, \pi_i)/m_i$	π_i	$\log \frac{\mu_i}{1-\mu_i}$	$\log(1+\exp(\theta_i))$	$\frac{1}{m_i}$
Poisson	$Po(\lambda_i)$	λ_i	$\log(\mu_i)$	$\exp(\theta_i)$	1
Negative binomial (rescaled)	$NB(\nu, \frac{\nu(1-\pi_i)}{\pi_i})/\nu$	$\frac{1-\pi_i}{\pi_i}$	$\log \frac{\mu_i}{1+\mu_i}$	$\log(1+\exp(\theta_i))$	$\frac{1}{\nu}$

(b) Expectation and variance

Distribution	$\mu_i = b'(\theta_i)$	$v(\mu_i) = b''(\theta_i)$	$\sigma_i^2 = \frac{\phi b''(\theta_i)}{w_i}$
Normal	$\mu_i = \theta_i$	1	$\frac{\sigma^2}{w_i}$
Exponential	$\lambda_i = -\frac{1}{\theta_i}$	μ_i^2	$\frac{\mu_i^2}{w_i}$
Gamma	$\mu_i = -\frac{1}{\theta_i}$	μ_i^2	$\frac{\mu_i^2}{\nu w_i}$
Inverse Gaussian	$\mu_i = \frac{1}{\sqrt{-2\theta_i}}$	μ_i^3	$\frac{\mu_i^3}{\lambda w_i}$
Bernoulli	$\pi_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$	$\pi_i(1-\pi_i)$	$\frac{\pi_i(1-\pi_i)}{w_i}$
Binomial (resc.)	$\pi_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$	$\pi_i(1-\pi_i)$	$\frac{\pi_i(1-\pi_i)}{m_i w_i}$
Poisson	$\lambda_i = \exp(\theta_i)$	λ_i	$\frac{\lambda_i}{w_i}$
Negative binomial (resc.)	$\mu_i = \frac{\exp(\theta_i)}{1-\exp(\theta_i)}$	$\mu_i(1+\mu_i)$	$\frac{\mu_i(1+\mu_i)}{\nu w_i}$

Table 3.2: Distributions of the exponential family

3.2 Likelihood Inference

Regression analysis with GLMs is based on likelihoods. In this section we give an overview of the basic inference tools for parameter estimation, hypothesis testing and possible criteria for checking the goodness-of-fit. Following Fahrmeir and Tutz (2001), the methods rely on the genuine method of ML, i.e. we assume that the model is completely and correctly specified in the context of Section 3.1.

3.2.1 Maximum-Likelihood Estimation

For all GLMs maximum likelihood estimation has a common form. This is due to the assumption that the responses come from an exponential family. We found that for the simple exponential family with density (3.1.3) the mean and variance are given by

$$E[y_i|\mathbf{x}_i] = \mu_i = b'(\theta_i), \qquad Var(y_i|\mathbf{x}_i) = \sigma_i^2 = \frac{\phi b''(\theta_i)}{w_i},$$

where the parametrization is specified by the canonical parameter θ_i.

Log-likelihood and score function

From the exponential family (3.1.3) one obtains for (conditionally) independent observations $y_1, \ldots, y_n$ the log-likelihood, up to an additive constant,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\theta_i) = \sum_{i=1}^{n} \frac{(y_i\theta_i - b(\theta_i))w_i}{\phi}.$$

When computing the derivatives it is useful to consider the parameters to result from transformations in the form $\theta_i = \theta(\mu_i), \mu_i = h(\eta_i), \eta_i = \mathbf{x}_i^\intercal\boldsymbol{\beta}$. One has the transformation structure

$$\begin{array}{ccccc} & \overset{h}{\longrightarrow} & & \overset{\theta}{\longrightarrow} & \\ \eta_i & & \mu_i & & \theta_i \\ & \underset{g = h^{-1}}{\longleftarrow} & & \underset{\mu = \theta^{-1}}{\longleftarrow} & \end{array}$$

yielding $\theta_i = \theta(\mu_i) = \theta(h(\eta_i))$. We define the *score function* as

$$\mathbf{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l(\theta_i)}{\partial \theta} \frac{\partial \theta(\mu_i)}{\partial \mu} \frac{\partial h(\eta_i)}{\partial \eta} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}.$$

With $\mu_i = \mu(\theta_i)$ denoting the transformation of θ_i into μ_i we obtain

$$\begin{aligned} \frac{\partial l(\theta_i)}{\partial \theta} &= \frac{(y_i - b'(\theta_i))w_i}{\phi} = \frac{(y_i - \mu_i)w_i}{\phi}, \\ \frac{\partial \theta(\mu_i)}{\partial \mu} &= \left(\frac{\partial \mu(\theta_i)}{\partial \theta}\right)^{-1} = \left(\frac{\partial^2 b(\theta_i)}{\partial \theta^2}\right)^{-1} = \frac{\phi}{\sigma_i^2 w_i} \\ \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} &= \mathbf{x}_i. \end{aligned}$$

Defining $d_i := \frac{\partial h(\eta_i)}{\partial \eta}$, altogether we have

$$\mathbf{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \frac{d_i(y_i - \mu_i)w_i\phi}{\phi w_i \sigma_i^2} = \sum_{i=1}^{n} \mathbf{x}_i \frac{d_i(y_i - \mu_i)}{\sigma_i^2}. \qquad (3.2.1)$$

Note that as $E[y_i|\mathbf{x}_i] = \mu_i$, it follows that $E[\mathbf{s}(\boldsymbol{\beta})|\mathbf{x}_i] = \mathbf{0}$, and the corresponding estimation equation, called *ML-equation* and given by $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, yields

$$\sum_{i=1}^{n} \mathbf{x}_i \frac{d_i(y_i - \mu_i)}{\sigma_i^2} = \sum_{i=1}^{n} \mathbf{x}_i \frac{d_i(y_i - \mu_i)w_i}{\phi v(\mu_i)} = \mathbf{0}\,. \tag{3.2.2}$$

Equation (3.2.2) shows that the response (or link) function is found in the specification of the mean $\mu_i = h(\mathbf{x}_i^\intercal \boldsymbol{\beta})$ and in the derivative $d_i = \partial h(\eta_i)/\partial\eta$, whereas from higher moments of the distribution of y_i only the variance $\sigma_i^2 = \phi v(\mu_i)/w_i$ is needed. As the dispersion parameter ϕ usually does not depend on i, it may be canceled out and the estimate $\hat{\boldsymbol{\beta}}$ does not depend on ϕ. For the canonical link the estimation equation simplifies. Since $\theta_i = \eta_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}$, the score function reduces to

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial l(\theta_i)}{\partial \theta} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\mathbf{x}_i (y_i - \mu_i) w_i}{\phi}.$$

In particular, the equality $\partial h(\eta_i)/\partial\eta = \sigma_i^2 w_i/\phi$ holds.
In matrix notation the score function (3.2.1) is given by

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{X}^\intercal \mathbf{D} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}),$$

where $\mathbf{X}^\intercal = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is the design matrix, with the diagonal matrix of derivatives $\mathbf{D} = diag(\partial h(\eta_1)/\partial\eta, \ldots, \partial h(\eta_n)/\partial\eta)$, $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ is the covariance matrix, and $\mathbf{y}^\intercal = (y_1, \ldots, y_n)$, $\boldsymbol{\mu}^\intercal = (\mu_1, \ldots, \mu_n)$ are the vectors of observations and means. Sometimes it is useful to combine $\mathbf{D}$ and $\boldsymbol{\Sigma}$ into the so-called weight matrix $\mathbf{W} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^\intercal$, which yields

$$\mathbf{s}(\boldsymbol{\beta}) = \mathbf{X}^\intercal \mathbf{W} \mathbf{D}^{-1} (\mathbf{y} - \boldsymbol{\mu}).$$

Furthermore, we define $\mathbf{F}(\boldsymbol{\beta}) = \mathbf{X}^\intercal \mathbf{W} \mathbf{X}$. By using the notation $\mathbf{W}$, $\mathbf{D}$ and $\boldsymbol{\Sigma}$, the dependence on $\boldsymbol{\beta}$ is suppressed. Actually, we have $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta})$, $\mathbf{D} = \mathbf{D}(\boldsymbol{\beta})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\beta})$.

Information matrix

In maximum likelihood theory the information matrix determines the asymptotic variance. The observed information matrix is given by

$$\mathbf{F}_{obs}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal} = \left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j} \right)_{i,j}.$$

Its explicit form is not needed in the sequel (it is given e.g. in Appendix A.1 from Fahrmeir and Tutz, 2001), but it shows that the observed information matrix depends on the observations and therefore is random. The (expected) information matrix, which is also called *Fisher matrix*, is given by

$$\mathbf{F}(\boldsymbol{\beta}) = E[\mathbf{F}_{obs}(\boldsymbol{\beta})].$$

For the derivation we use $E[\mathbf{s}(\boldsymbol{\beta})|\mathbf{x}_i] = \mathbf{0}$ and it is essential that $E[-\partial^2 l_i/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal] = E[(\partial l_i/\partial\boldsymbol{\beta})(\partial l_i/\partial\boldsymbol{\beta}^\intercal)]$, which holds under general assumptions (see for example Cox and Hinkley, 1974). Thus, we obtain

$$\begin{aligned}\mathbf{F}(\boldsymbol{\beta}) &= E[\sum_{i=1}^{n} s_i(\boldsymbol{\beta}) s_i(\boldsymbol{\beta})^\intercal] = E\left[\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\intercal \left(\frac{\partial h(\eta_i)}{\partial\eta}\right)^2 \frac{(y_i-\mu_i)^2}{\sigma_i^4}\right] \\ &= \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\intercal \left(\frac{\partial h(\eta_i)}{\partial\eta}\right)^2 /\sigma_i^2,\end{aligned}$$

where $\sigma_i^2 = Var(y_i|\mathbf{x}_i)$. Using the definition of the design matrix $\mathbf{X}^\intercal = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, we see that $\mathbf{F}(\boldsymbol{\beta})$ is given by

$$\mathbf{F}(\boldsymbol{\beta}) = \mathbf{X}^\intercal\mathbf{W}\mathbf{X}, \tag{3.2.3}$$

where $\mathbf{W} = diag\left(\left(\frac{\partial h(\eta_1)}{\partial\eta}\right)^2/\sigma_1^2, \ldots, \left(\frac{\partial h(\eta_n)}{\partial\eta}\right)^2/\sigma_n^2\right)$ is a diagonal weight matrix that has the matrix form $\mathbf{W} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^\intercal$, where again the dependence on $\boldsymbol{\beta}$ is suppressed.
For the canonical link the corresponding simpler form is

$$\mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\intercal \sigma_i^2 w_i^2/\phi^2 = \mathbf{X}^\intercal\mathbf{W}\mathbf{X},$$

with weight matrix $\mathbf{W} = (\sigma_1^2 w_1^2, \ldots, \sigma_n^2 w_n^2)/\phi^2$. Besides, in this case the observed information is identical to the information matrix, i.e. $\mathbf{F}_{obs}(\boldsymbol{\beta}) = \mathbf{F}(\boldsymbol{\beta})$. It is immediately seen that for the normal distribution model with (canonical) identity link and with $\phi = \sigma^2$ and $w_i = 1$ one obtains the familiar form

$$\mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\intercal/\sigma^2 = \mathbf{X}^\intercal\mathbf{X}/\sigma^2.$$

We know that for the normal distribution model with (canonical) identity link the covariance is given by $cov(\hat{\beta}) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1} = \mathbf{F}(\boldsymbol{\beta})^{-1}$.
However, for GLMs this result holds only asymptotically for $n \to \infty$ with

$$\widehat{cov}(\hat{\boldsymbol{\beta}}) \approx (\mathbf{X}^\intercal\hat{\mathbf{W}}\mathbf{X})^{-1},$$

where the notation $\hat{\mathbf{W}}$ indicates the $\mathbf{W}$ is evaluated at $\hat{\boldsymbol{\beta}}$, i.e. $\partial h(\eta_i)/\partial\eta$ is replaced by $\partial h(\hat{\eta}_i)/\partial\eta$ with $\hat{\eta}_i = \mathbf{x}_i^\intercal\hat{\boldsymbol{\beta}}$, and $\sigma_i^2 = \phi v(\hat{\mu}_i)$ with $\hat{\mu}_i = h(\hat{\eta}_i)$.
Note that in the grouped observations case one obtains the same form of the likelihood, score function and Fisher matrix, only the summation index n has to be replaced by the number of groups and corresponding weights have to be adjusted.

If ϕ is unknown, as for example in the case of a normal or Gamma distribution, the moments estimate is given by

$$\hat{\phi} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}.$$

For the normal model with $w_i = 1$, $\hat{\phi}$ reduces to the usual unbiased and consistent estimate $\hat{\phi} = \hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/(n-p-1)$. For the Gamma distribution with $w_i = 1$ we obtain

$$\hat{\phi} = \frac{1}{\hat{\nu}} = \frac{1}{n-p-1} \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2.$$

In the following approximation, ϕ has to be replaced by $\hat{\phi}$ when computing $\mathbf{F}(\hat{\boldsymbol{\beta}})$. Table 3.3 summarizes the log-likelihood, the score function, the Fisher matrix and the approximative covariance matrix of $\hat{\boldsymbol{\beta}}$.

Log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(y_i\theta_i - b(\theta_i))w_i}{\phi}.$$

Score-function

$$\mathbf{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial h(\hat{\eta}_i)}{\partial \eta} \frac{(y_i - \mu_i)}{\sigma_i^2} = \mathbf{X}^\intercal \mathbf{W} \mathbf{D}^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

Information matrix

$$\mathbf{F}(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal} \right] = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\intercal \left(\frac{\partial h(\eta_i)}{\partial \eta} \right)^2 / \sigma_i^2 = \mathbf{X}^\intercal \mathbf{W} \mathbf{X}$$

Approximate covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\widehat{cov}(\hat{\boldsymbol{\beta}}) \approx F(\hat{\boldsymbol{\beta}})^{-1} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\intercal \left(\frac{\partial h(\eta_i)}{\partial \eta} \right)^2 w_i / (\hat{\phi} v(\mu_i)) \right)^{-1}$$

Table 3.3: Log-likelihood, score function and Fisher matrix for GLMs.

The unifying concept of GLMs may be seen in the common form of the log-likelihood, the score function (which determines the estimation equation) and the information matrix (which determines the variances of estimators). Specific forms result from specific choices of

- the link or response function, yielding $\partial h(\eta_i)/\partial \eta$
- the distribution, yielding $\sigma_i^2 = var(y_i)$.

3.2.2 Computation of Maximum-Likelihood Estimates

Maximum likelihood estimates are obtained by solving the equation $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. In general there is no closed form of the estimate available and iterative procedures have to be applied. The *Newton-Raphson method* is such an iterative method for solving non-linear equations. Following Fahrmeir and Tutz (2001), one starts with an initial guess $\boldsymbol{\beta}^{(0)}$ and the solution is then found by successive improvement. Let $\hat{\boldsymbol{\beta}}^{(k)}$ denote the estimate in the k-th iteration step, where $k = 0$ is the initial estimate. If $\mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)}) \neq \mathbf{0}$, linear Taylor approximation is considered:

$$\mathbf{s}(\boldsymbol{\beta}) \approx \mathbf{s}_{lin}(\boldsymbol{\beta}) = \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{\partial \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)})}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(k)}).$$

Instead of solving $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, we solve $\mathbf{s}_{lin}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, yielding

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)} - \left(\frac{\partial \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)})}{\partial \boldsymbol{\beta}} \right)^{-1} \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)}).$$

Since $\partial \mathbf{s}(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \partial^2 l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\intercal}$, with the corresponding Hessian matrix $\mathbf{H}(\boldsymbol{\beta}) = \partial^2 l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\intercal}$ we get the new estimate

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - \mathbf{H}(\hat{\boldsymbol{\beta}}^{(k)})^{-1} \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)}),$$

or alternatively, by using the observed information matrix $\mathbf{F}_{obs}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta})$, the new estimate yields

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \mathbf{F}_{obs}(\hat{\boldsymbol{\beta}}^{(k)})^{-1} \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)}).$$

Finally, iterations are carried out until convergence, which means that changes between successive steps are smaller than a pre-specified threshold ε, i.e., the procedure is stopped, if

$$\frac{||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||}{||\hat{\boldsymbol{\beta}}^{(k)}||} < \varepsilon.$$

Usually, convergence is quite fast and the number of correct decimals in the approximation is roughly doubling at each iteration.
An alternative method is the *Newton method with Fisher scoring*. The essential difference is that the observed information matrix $\mathbf{F}_{obs}(\boldsymbol{\beta})$ is replaced by the Fisher matrix $\mathbf{F}(\boldsymbol{\beta}) = E[\mathbf{F}_{obs}(\boldsymbol{\beta})]$, or equivalently, $\mathbf{H}(\boldsymbol{\beta})$ by $-\mathbf{F}(\boldsymbol{\beta})$, yielding

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \mathbf{F}(\hat{\boldsymbol{\beta}}^{(k)})^{-1} \mathbf{s}(\hat{\boldsymbol{\beta}}^{(k)}). \tag{3.2.4}$$

Note that the iterative scheme (3.2.4) may alternatively be seen as an iterative weighted least-squares fitting procedure. Let *pseudo* or *working observations* be defined by

$$\tilde{\eta}_i(\hat{\boldsymbol{\beta}}) = \mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}} + \left(\frac{\partial h(\hat{\eta}_i)}{\partial \eta}\right)^{-1} (y_i - \mu_i(\hat{\boldsymbol{\beta}})),$$

then the corresponding vector of pseudo observations $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})^\intercal = (\tilde{\eta}_1(\hat{\boldsymbol{\beta}}), \ldots, \tilde{\eta}_n(\hat{\boldsymbol{\beta}}))$ is given by

$$\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})^\intercal = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{D}(\hat{\boldsymbol{\beta}})^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}).$$

By simple substitution, we obtain

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)}).$$

Consequently, $\hat{\boldsymbol{\beta}}^{(k+1)}$ has the form of a weighted least-squares estimate corresponding to the working observations $(\tilde{\eta}_i(\hat{\boldsymbol{\beta}}^{(k)}), \mathbf{x}_i), i = 1, \ldots, n$, with a weight matrix $\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})$, which is depending on the current iteration.
For a canonical link one obtains $\mathbf{F}(\boldsymbol{\beta}) = \mathbf{X}^\intercal \mathbf{W}(\boldsymbol{\beta})\mathbf{X}$ with $\mathbf{W}(\boldsymbol{\beta}) = \mathbf{R}\boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{R}/\phi^2, \mathbf{R} = diag(w_1, \ldots, w_n)$ and score function $\mathbf{s}(\boldsymbol{\beta}) = \mathbf{X}^\intercal \mathbf{R}(\mathbf{y} - \boldsymbol{\mu})/\phi$ and therefore

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + (\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{R}(\mathbf{y} - \hat{\boldsymbol{\mu}})/\phi,$$

which corresponds to the least-square fitting

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{W}\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)}),$$

with $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{W}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{R}(\mathbf{y} - \hat{\boldsymbol{\mu}})/\phi$.

Hat Matrix for GLMs

With weight matrix $\mathbf{W}(\boldsymbol{\beta}) = \mathbf{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})^{-1}\mathbf{D}(\boldsymbol{\beta})^\intercal$ the iterative fitting procedure has the form

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}^{(k)}),$$

with the pseudo response $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})^\intercal = \hat{\boldsymbol{\eta}} + \mathbf{D}(\hat{\boldsymbol{\beta}})^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ and $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. At convergence we obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{W}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}). \tag{3.2.5}$$

Hence, $\hat{\boldsymbol{\beta}}$ constitutes a weighted least-squares solution to the linear problem $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, or an unweighted least-squares solution to the linear problem

$$\mathbf{W}^{\frac{\intercal}{2}}\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \mathbf{W}^{\frac{\intercal}{2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where in $\mathbf{W} = \mathbf{W}(\hat{\boldsymbol{\beta}})$ again the dependence on $\hat{\boldsymbol{\beta}}$ is suppressed. The corresponding hat matrix is given by

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^\intercal\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{W}^{\frac{1}{2}}.$$

Since the matrix $\mathbf{H}$ is idempotent and symmetric it may be seen as a projection matrix for which $tr(\mathbf{H}) = rk(\mathbf{H})$ holds. Moreover, one obtains for the diagonal elements of $\mathbf{H} = (h_{ij})_{i,j} : 0 \leq h_{ii} \leq 1$ and $tr(\mathbf{H}) = p + 1$ (if $\mathbf{X}$ has full rank). It should be noted that, in contrast to the normal regression model, the hat matrix depends on $\hat{\boldsymbol{\beta}}$, since $\mathbf{W} = \mathbf{W}(\hat{\boldsymbol{\beta}})$. It follows from Equation (3.2.5) that $\mathbf{W}^{\frac{1}{2}}\hat{\boldsymbol{\eta}} = \mathbf{H}\mathbf{W}^{\frac{1}{2}}\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$ holds, which shows how the hat matrix maps the adjusted variable $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$ into the fitted values $\hat{\boldsymbol{\eta}}$. Thus, $\mathbf{H}$ may be seen as the matrix that maps the adjusted observation vector $\mathbf{W}^{\frac{1}{2}}\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$ into the vector of "fitted" values $\mathbf{W}^{\frac{1}{2}}\hat{\boldsymbol{\eta}}$, which is a mapping on the transformed predictor space. For the linear model the hat matrix represents a simple projection having the form $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$. In the present case of generalized linear models it may be shown that approximately

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \approx \mathbf{H}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}) \tag{3.2.6}$$

holds, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$. Hence, $\mathbf{H}$ may be seen as a measure of the influence of $\mathbf{y}$ on $\hat{\boldsymbol{\mu}}$ in standardized units of changes. From (3.2.6) follows

$$\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \approx \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}), \tag{3.2.7}$$

such that the influence in unstandardized units is given by the projection matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}^{-\frac{1}{2}}$, which is idempotent but not symmetric in general. Note that for the normal regression model with identity link we simply get $\mathbf{W} = \mathbf{I}/\sigma^2, \mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ and both (3.2.6) and (3.2.7) hold exactly.

3.2.3 Asymptotic Properties of the ML Estimator

Following Fahrmeir and Tutz (2001), we give a short overview of the asymptotic properties of the ML estimator (MLE). Inferential methods for GLMs rely on these properties. Under comparatively weak "regularity assumption", which are discussed e.g. in Fahrmeir and Tutz (2001), the following properties hold:

- *Asymptotic existence and uniqueness*:
  The probability that $\hat{\boldsymbol{\beta}}$ exists and is (locally) unique tends to 1 for $n \to \infty$.
- *Consistency*:
  For an unknown true parameter vector $\boldsymbol{\beta}$, for $n \to \infty$ we have that $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ in probability (weak consistency) or with probability 1 (strong consistency).

- *Asymptotic normality*:
  The distribution of the (normed) MLE is normal for $n \to \infty$, or, more informally, for n large
  $$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}),$$
  i.e., $\hat{\boldsymbol{\beta}}$ is approximately normal with approximate (or "asymptotic") covariance matrix
  $$cov(\hat{\boldsymbol{\beta}}) \overset{a}{=} \mathbf{F}(\hat{\boldsymbol{\beta}})^{-1},$$
  where $\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}$ is the inverse Fisher matrix.

Note that for an unknown scale parameter ϕ all results remain valid, if it is replaced by a consistent estimate $\hat{\phi}$. Furthermore, the MLE is asymptotically efficient compared to a wide class of other estimators. Fahrmeir and Tutz (2001) give some heuristic arguments for consistency and asymptotic normality of the MLE. As we know, $E[\mathbf{s}(\boldsymbol{\beta})] = \sum_{i=1}^{n} E[\mathbf{s}_i(\boldsymbol{\beta})] = \mathbf{0}$ and hence, by some law of large numbers, $\mathbf{s}(\boldsymbol{\beta})/n \to 0$ in probability. Since $\mathbf{s}(\hat{\boldsymbol{\beta}})/n = \mathbf{0}$ holds for the MLE $\hat{\boldsymbol{\beta}}$, one obtains $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}$ in probability, i.e. (weak) consistency, by continuity arguments.
With $E[\mathbf{s}(\boldsymbol{\beta})] = \mathbf{0}$, it is easy to derive

$$cov(\mathbf{s}(\boldsymbol{\beta})) = E[\mathbf{s}(\boldsymbol{\beta})\mathbf{s}(\boldsymbol{\beta})^{\mathsf{T}}] = \mathbf{F}(\boldsymbol{\beta})$$

with the (expected) Fisher information $\mathbf{F}(\boldsymbol{\beta})$ given in (3.2.3). Applying a central limit theorem to the sum $\mathbf{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{s}_i(\boldsymbol{\beta})$, we obtain approximate normality of the score function, i.e.

$$\mathbf{s}(\boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, \mathbf{F}(\boldsymbol{\beta})) \tag{3.2.8}$$

for large n.
First-order Taylor expansion of $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ about $\boldsymbol{\beta}$ yields

$$\mathbf{0} = \mathbf{s}(\hat{\boldsymbol{\beta}}) \overset{a}{\sim} \mathbf{s}(\boldsymbol{\beta}) + \mathbf{H}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where $\mathbf{H}(\boldsymbol{\beta}) = \partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathsf{T}} = -\mathbf{F}_{obs}(\boldsymbol{\beta})$. Replacing the observed information matrix by its expectation $\mathbf{F}(\boldsymbol{\beta})$, we get

$$\mathbf{s}(\boldsymbol{\beta}) \overset{a}{\sim} \mathbf{F}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

and

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \overset{a}{\sim} \mathbf{F}(\boldsymbol{\beta})^{-1}\mathbf{s}(\boldsymbol{\beta}).$$

From the approximate normality of $\mathbf{s}(\boldsymbol{\beta})$ in (3.2.8), the approximate normality of the MLE,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \overset{a}{\sim} N(\mathbf{0}, \mathbf{F}(\boldsymbol{\beta})^{-1})$$

follows, where the approximate covariance matrix is obtained via $\mathbf{A}(\boldsymbol{\beta}) = \mathbf{F}(\boldsymbol{\beta})^{-1}\mathbf{F}(\boldsymbol{\beta})\mathbf{F}(\boldsymbol{\beta})^{-1} = \mathbf{F}(\boldsymbol{\beta})^{-1}$. Replacing $\boldsymbol{\beta}$ by its consistent estimate $\hat{\boldsymbol{\beta}}$ gives $\mathbf{A}(\hat{\boldsymbol{\beta}}) = \mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}$.
Under appropriate regularity assumptions, these heuristic arguments are the basis for rigorous proofs, see Appendix A.2 in Fahrmeir and Tutz (2001).

3.3 Diagnostics and Goodness-of-Fit

According to Tutz (2011), main questions concerning the diagnostics and the goodness-of-fit are

- the adequacy of the model or goodness-of-fit of the model,
- the relevance of explanatory variables,
- the explanatory value of the model.

In the following, these questions are considered in different order. First the *deviance* is introduced, which measures the discrepancy between observations and the fitted model. The deviance is a tool for various purposes. The relevance of explanatory variables may be investigated by comparing the deviance of two models, the model which contains the variable in question and the model where this variable is omitted. Moreover, for grouped observations the deviance may be used as a goodness-of-fit statistic.

3.3.1 The Deviance

When fitting a GLM we want some measure for the discrepancy between the fitted model and the observations. The deviance is a measure for the discrepancy, which is based on the likelihood ratio statistic for comparing nested models. The nested models which are investigated are the GLM under investigation and the most general possible model. This so-called *saturated model* fits the data exactly by assuming as many parameters as observations. Let $l(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi)$ denote the maximum of the log-likelihood of the model where $\mathbf{y}^\intercal = (y_1, \ldots, y_n)$ represents the data and $\hat{\boldsymbol{\mu}}^\intercal = (\hat{\mu}_1, \ldots, \hat{\mu}_1)$ with $\hat{\mu}_i = h(\mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}})$ represents the fitted values based on the MLE $\hat{\boldsymbol{\beta}}$. For the saturated model, which matches the data exactly, one has $\hat{\boldsymbol{\mu}} = \mathbf{y}$ and the log-likelihood is given by $l(\mathbf{y}, \mathbf{y}, \phi)$. With $\theta(\hat{\mu}_i), \theta(y_i)$ denoting the canonical parameters of the GLM under investigation and of the saturated model, respectively, the deviance is given by

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= -2\phi\{l(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi) - l(\mathbf{y}, \mathbf{y}, \phi)\} \\ &= 2\sum_{i=1}^{n}\{y_i[\theta(y_i) - \theta(\hat{\mu}_i)] - [b(\theta(y_i)) - b(\theta(\hat{\mu}_i))]\}. \end{aligned}$$

Note that while $D(\mathbf{y}, \hat{\boldsymbol{\mu}})$ is known as deviance of the model under consideration, $D(\mathbf{y}, \hat{\boldsymbol{\mu}})/\phi$ is the so-called *scaled deviance*. The deviance is linked to the likelihood ratio statistic $\lambda = -2\{l(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi) - l(\mathbf{y}, \mathbf{y}, \phi)\}$, which compares the current model to the saturated model by $D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \phi\lambda$ and which is examined in detail in Section 3.3.2.
Simple derivation yields the deviances given in Table 3.4. Obviously, for the normal model the deviance is identical to the error or residual sum of squares SSE and the scaled deviance takes the form SSE/σ^2.

Normal	$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$
Gamma	$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} -\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right]$
Inverse Gaussian	$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{(\hat{\mu}_i^2 y_i)}$
Bernoulli	$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i) \log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)$
Poisson	$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - [(y_i - \hat{\mu}_i)]$

Table 3.4: Deviance for a selected set of distributions.

Adopting the convention $\infty \cdot 0 = 0$, with $\hat{\mu}_i = \hat{\pi}_i$ we obtain for the Bernoulli distribution the form

$$D(\mathbf{y}, \hat{\mu}) = 2 \sum_{i=1}^{n} d(y_i, \hat{\pi}_i)$$

with

$$d(y_i, \hat{\pi}_i) = \begin{cases} -\log(\hat{\pi}_i) & y_i = 1 \\ -\log(1 - \hat{\pi}_i) & y_i = 0. \end{cases}$$

The simpler form

$$d(y_i, \hat{\pi}_i) = -\log(1 - |y_i - \hat{\pi}_i|)$$

shows that for the binary data the deviance uses the difference between observations and fitted values. In the case of the Bernoulli distribution there is no difference between deviance and scaled deviance. In the cases of the Poisson and the Gamma deviances the last term given in brackets [...] can be omitted, if the model includes a constant term, because then the sum over the terms is zero (no proof; but compare the results of Lemma 2.2.14 (iii) and Exercise 5, Problem Sheet 1).
Following Tutz (2011), we want to give some general remarks on the concept of deviance.

Remark 3.3.1. 1) The deviance as a measure of discrepancy between the observations and the fitted model may be used in an informal way to compare the fit of two models. For example, two models with the same predictor but differing link functions can be compared by considering which one has the smaller deviance. But unfortunately, there is no simple way to interpret the difference between the deviances of these models.

2) However, if the difference of deviances is used for nested models, for example to investigate the relevance of terms in the linear predictor, it has a simple interpretation. The comparison of models with and without the term in question allows a decision based on significance tests with known

asymptotic distribution. The corresponding *analysis of deviance* (see Section 3.3.2) generalizes the concept of *analysis of variance*, which is in common use for normal linear models, to GLMs.

3) For ungrouped data the interpretation of the deviance as a goodness-of-fit measure must be treated with caution. As an absolute measure of goodness-of-fit, which allows to decide, whether the model achieves a satisfactory fit or not, the deviance for ungrouped observations is appropriate only in some cases. For the interpretation of the value of the deviance it would be useful to have a benchmark in the form of an asymptotic distribution. Since the deviance may be derived as a likelihood ratio statistic it is tempting to assume that the deviance is asymptotically χ^2-distributed. However, in general the deviance does not have an asymptotic χ^2-distribution in the limit for $n \to \infty$. Standard asymptotic theory of likelihood ratio statistics for nested models assumes that the ranks of the design matrices that build the two models and therefore the degrees of freedom are fixed for increasing sample size. In the present case this theory does not apply, since the degrees of freedom of the saturated model increase with n.

We want to give a short example for the scenario mentioned in Remark 3.3.1 3), where the degrees of freedom of the saturated model increase with n.

Example 3.3.2. In the case of the normal distribution, we already know that $D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = (n - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$. Hence, for $n \to \infty$ the limiting distribution does not have a χ^2-distribution with fixed degrees of freedom. △

Similar effects occur for binary data.

3.3.2 Analysis of Deviance and Hypothesis Testing

Let us consider the nested models $\tilde{M} \subset M$, where M is a given GLM with $\mu_i = h(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})$ and $\tilde{M}$ a submodel characterized by a linear restriction of the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{C}$ is a known $r \times (p+1)$-matrix with $rk(\mathbf{C}) = r \leq p+1$ and $\mathbf{d}$ is a r-dimensional vector. This means that $\tilde{M}$ corresponds to the null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ and thus specifies a simpler structure of the predictor.

Analysis of Deviance

Let $\tilde{\boldsymbol{\mu}}^\mathsf{T} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_n)$ denote the fitted values for the restricted model and $\hat{\boldsymbol{\mu}}^\mathsf{T} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ those corresponding to the fit for model M. Then we obtain the corresponding two deviances

$$D(M) = -2\phi\{l(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi) - l(\mathbf{y}, \mathbf{y}, \phi)\},$$
$$D(\tilde{M}) = -2\phi\{l(\mathbf{y}, \tilde{\boldsymbol{\mu}}, \phi) - l(\mathbf{y}, \mathbf{y}, \phi)\},$$

The difference of deviances, which is given by

$$D(\tilde{M}|M) = D(\tilde{M}) - D(M),$$

is strongly connected to the likelihood ratio statistic for testing H_0. The likelihood ratio statistic is equivalent to the difference of scaled deviances

$$\begin{aligned} l(\tilde{M}|M) &= -2\{l(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi) - l(\mathbf{y}, \tilde{\boldsymbol{\mu}}, \phi)\}, \\ &= \frac{1}{\phi} D(\tilde{M}|M). \end{aligned}$$

Similar to the partitioning of the sum of squares in linear regression one may consider the partitioning of the deviance of the restricted model $\tilde{M}$

$$D(\tilde{M}) = D(\tilde{M}|M) + D(M),$$

where $D(\tilde{M}|M)$ represents the increase in discrepancy between data and fit, if instead of model M the more restrictive model $\tilde{M}$ is fitted. For normal distributions this corresponds to the partitioning of the sum of squares

$$SSE(\tilde{M}) = SSE(\tilde{M}|M) + SSE(M)\,,$$

which we have already used in Section 2.2.5 in the form $\text{SSE}_{H_0} = \Delta\text{SSE}+\text{SSE}$. From Proposition 2.2.21 we know already that in the normal case $SSE(M) = \text{SSE}$ follows a $\sigma^2\chi^2_{n-p-1}$-distribution, if M holds. We also know that if $\tilde{M}$ holds, then $SSE(\tilde{M}|M)$ and $SSE(M)$ are independent with $SSE(\tilde{M}|M) \sim \sigma^2\chi^2_r$ and for testing H_0 we can use the F-statistic

$$\frac{n-p-1}{r} \frac{SSE(\tilde{M}) - SSE(M)}{SSE(M)} \sim F_{r,n-p-1}.$$

In the general case of GLMs one uses

$$\frac{D(\tilde{M}) - D(M)}{\phi} = \frac{D(\tilde{M}|M)}{\phi},$$

which under mild restrictions is asymptotically χ^2_r-distributed.
When using the χ^2-approximation, the deviance has to be scaled by $1/\phi$. For the Bernoulli, exponential, Poisson (all $\phi = 1$) and the binomial ($\phi = 1/m_i$) distribution we can use the difference $D(\tilde{M}) - D(M)$ directly, whereas for the normal, the Gamma, the inverse Gaussian and the negative binomial distribution the dispersion parameter has to be estimated. For the normal regression case we already know that $\hat{\phi} = \hat{\sigma}^2$ is a consistent estimate and that $D(\tilde{M}|M)/r\hat{\phi}$ then follows an $F_{r,n-p-1}$-distribution. In general, the approximation by the F-distribution may be used, if $\hat{\phi}$ is consistent for ϕ with a scaled χ^2-distribution as approximate distribution and if $D(\tilde{M}) - D(M)$ and $\hat{\phi}$ are independent (see Jorgensen, 1987).

	df	cond. deviance	df
$D(\tilde{M})$	$n-p-1+r$		
$D(M)$	$n-p-1$	$D(\tilde{M}\|M)$	r

Table 3.5: Structure of the table for an analysis of deviance.

In analogy to the ANOVA table in normal regression one obtains a table for the analysis of deviance. Note that only the difference of deviances $D(\tilde{M}|M)$ has asymptotically a $\phi\chi^2_r$-distribution. The degrees of freedom of $D(M)$ have the basic structure "number of observations minus number of fitted parameters". In $\tilde{M}$ by considering an additional r dimensional restriction the effective parameters in the model are reduced to $p+1-r$, yielding $df = n-(p+1-r) = n-p-1+r$. In the case of grouped data the deviances $D(\tilde{M})$ and $D(M)$ themselves are asymptotically distributed with $D(\tilde{M}) \sim \chi^2_{g-p-1+r}$ and $D(M) \sim \chi^2_{g-p-1}$, where g denotes the number of grouped observations. While $D(\tilde{M})$ and $D(M)$ have different distributions for grouped and ungrouped data, the difference $D(\tilde{M}) - D(M)$ remains unchanged.
In the following we present a short summary of results on distributions and compare the classical linear case and the deviances within the GLM framework. For the classical linear model we have

$$\underbrace{SSE(\tilde{M})}_{\substack{\sigma^2\chi^2_{n-p-1+r} \\ \text{if } \tilde{M} \text{ holds}}} = \underbrace{SSE(\tilde{M}|M)}_{\substack{\sigma^2\chi^2_r \\ \text{if } \tilde{M} \text{ holds}}} + \underbrace{SSE(M)}_{\substack{\sigma^2\chi^2_{n-p-1} \\ \text{if } M \text{ holds}}}.$$

For grouped data within the GLM framework, at least asymptotic distributions can be derived, namely

$$\underbrace{D(\tilde{M})}_{\substack{\phi\chi^2_{g-p-1+r} \\ \text{if } M \text{ holds} \\ \text{grouped data}}} = \underbrace{D(\tilde{M}|M)}_{\substack{\phi\chi^2_r \\ \text{if } \tilde{M} \text{ holds}}} + \underbrace{D(M)}_{\substack{\phi\chi^2_{g-p-1} \\ \text{if } M \text{ holds} \\ \text{grouped data}}}.$$

This approach can be used to test a sequence of nested models

$$M_1 \subset M_2 \subset \ldots \subset M_m,$$

by using the successive differences $(D(M_k) - D(M_{k+1}))/\phi$. The deviance of the most restrictive model yields

$$\begin{aligned} D(M_1) = (D(M_1) - D(M_2)) + (D(M_2) - D(M_3)) \\ + \ldots + (D(M_{m-1}) - D(M_m)) + D(M_m) \end{aligned}$$

$$= D(M_1|M_2) + \ldots + D(M_{m-1}|M_m) + D(M_m)$$

and hence can be written as sum of these differences. Altogether, the discrepancy of the model M_1 is given by the the sum of the "conditional" deviances $D(M_i|M_{i+1}) = D(M_i) - D(M_{i+1})$ and the discrepancy between the most general model M_m and the saturated model. However, if one starts from a model M_m and considers sequences of simpler models, one should be aware that also different sequences of submodels are possible.

Alternative Test Statistics for Linear Hypothesis

The analysis of deviance leads to a test, which analyzes, if a model can be reduced to a model that has a simpler structure of covariates. The simplified structure can be specified by the null hypothesis H_0 of the following pair of hypothesis

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \qquad \text{vs.} \qquad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d},$$

where $rk(\mathbf{C}) = r$. Following Tutz (2011), alternative test statistics can be used, namely the Wald test and the score statistic.

Wald test

The *Wald statistic* has the form

$$w = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\intercal (\mathbf{C}\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}^\intercal)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

It is based on the weighted distance between the unrestricted estimate $\mathbf{C}\hat{\boldsymbol{\beta}}$ of $\mathbf{C}\boldsymbol{\beta}$ and its hypothetical value $\mathbf{d}$ under H_0. The weight is derived from the distribution of the difference $(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})$, for which one obtains asymptotically $Cov(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) = \mathbf{C}\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}^\intercal$. Hence, w is the squared length of the standardized estimate $(\mathbf{C}\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}^\intercal)^{-1/2}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})$, and we obtain an asymptotic χ^2_r-distribution for w under H_0.
An advantage of the Wald statistic is the fact that it is based on the MLEs of the full model. Thus, it is not necessary to compute an additional fit under H_0. This is the reason why most software packages give significance tests for single regression parameters in terms of the Wald statistic. Similar to the ordinary linear model from Chapter 2, if one wants to test a single parameter with $H_0 : \beta_j = 0$, the corresponding matrix $\mathbf{C}$ has the simple form $\mathbf{C} = (0, \ldots, 0, 1, 0, \ldots, 0)$, with a single one at position $j+1$. Then the Wald statistic reduces to

$$w = \frac{\hat{\beta}_j^2}{\hat{a}_{j+1,j+1}},$$

where $\hat{a}_{j+1,j+1}$ is the $(j+1)$-th diagonal element of the estimated inverse Fisher matrix $\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}$. Since w is asymptotically χ^2_1-distributed, one may also consider its square root

$$z = \sqrt{w} = \frac{\hat{\beta}_j}{\sqrt{\hat{a}_{j+1,j+1}}},$$

which follows asymptotically a standard normal distribution. Thus, for the significance tests of single regression parameters software packages usually provide the standard error $\sqrt{\hat{a}_{j+1,j+1}}$ together with the p-value based on z.

Score statistic

The score statistic is based on the following consideration. The score function $\mathbf{s}(\boldsymbol{\beta})$ of the unrestricted model is the zero vector, if it is evaluated at the unrestricted MLE $\hat{\boldsymbol{\beta}}$. If, however, $\hat{\boldsymbol{\beta}}$ is replaced by the MLE $\tilde{\boldsymbol{\beta}}$ under H_0, $\mathbf{s}(\tilde{\boldsymbol{\beta}})$ will be significantly different from zero, if H_0 is not true. Since the covariance of the score function is approximately the Fisher matrix, we define the *score statistic* as follows:

$$u = \mathbf{s}(\tilde{\boldsymbol{\beta}})^\intercal \mathbf{F}(\tilde{\boldsymbol{\beta}})^{-1} \mathbf{s}(\tilde{\boldsymbol{\beta}}),$$

which is the squared weighted score function evaluated at $\tilde{\boldsymbol{\beta}}$.
One advantage of the Wald and score statistics is that both are properly defined for models with over-dispersion, since only first and second moments are involved. An overview of the test statistics is presented in Table 3.6. All test statistics have the same asymptotic distribution. If they are differing strongly, this is a hint that the conditions for asymptotic results may not hold. For a survey on asymptotics for test statistics, see Fahrmeir (1987).

Test statistics for linear hypothesis

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \qquad \text{vs.} \qquad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d},$$

where $rk(\mathbf{C}) = r$:

Likelihood ratio statistic

$$\begin{aligned} \lambda &= -2\{l(\mathbf{y}, \tilde{\boldsymbol{\mu}}, \hat{\phi}) - l(\mathbf{y}, \hat{\boldsymbol{\mu}}, \hat{\phi})\} \\ &= (D(\mathbf{y}, \tilde{\boldsymbol{\mu}}, \hat{\phi}) - D(\mathbf{y}, \hat{\boldsymbol{\mu}}, \hat{\phi}))/\hat{\phi} \end{aligned}$$

Wald statistic

$$w = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\intercal (\mathbf{C}\mathbf{F}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{C}^\intercal)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})$$

Score statistic

$$u = \mathbf{s}(\tilde{\boldsymbol{\beta}})^\intercal \mathbf{F}(\tilde{\boldsymbol{\beta}})^{-1} \mathbf{s}(\tilde{\boldsymbol{\beta}})$$

Approximation:

$$\lambda, w, u \sim \chi^2_r$$

Table 3.6: Test statistics for linear hypothesis.

A

Addendum from Linear Algebra, Analysis and Stochastic

In this chapter we present a short summary with some fundamental properties from Linear Algebra, Analysis and Stochastic, see e.g. Pruscha (2000).

Proposition A.0.1. *Let* $\mathbf{x} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^n$ *and* $\mathbf{a} \in \mathbb{R}^m$ *be vectors of dimensiona* n *and* m*, respectively. Then the following derivation rules hold:*

$\frac{d}{d\mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}$ $[\mathbf{A}\ m \times n - matrix]$

$\frac{d}{d\mathbf{x}}(\mathbf{x}^\intercal \mathbf{A}) = \mathbf{A}$ $[\mathbf{A}\ n \times m - matrix]$

$\frac{d}{d\mathbf{x}}(\mathbf{x}^\intercal \mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$ $[\mathbf{A}\ symmetric\, n \times n - matrix]$

$\frac{d^2}{d\mathbf{x}d\mathbf{x}^\intercal}(\mathbf{x}^\intercal \mathbf{A}\mathbf{x}) = 2\mathbf{A}$ $[\mathbf{A}\ symmetric\, n \times n - matrix]$

$\frac{d}{d\mathbf{x}}((\mathbf{A}\mathbf{x} - \mathbf{a})^\intercal \cdot (\mathbf{A}\mathbf{x} - \mathbf{a})) = 2\mathbf{A}^\intercal(\mathbf{A}\mathbf{x} - \mathbf{a})$ $[\mathbf{A}\ m \times n - matrix]$.

$\frac{d}{d\mathbf{x}}(\boldsymbol{b}^\intercal \mathbf{x}) = \boldsymbol{b}$.

Proposition A.0.2. *Let* $\boldsymbol{A}$ *be a* $n \times n$ *and* $\boldsymbol{Q}$ *be a* $n \times m$ *matrix. Then:*

1. *If* $\boldsymbol{A}$ *is positive semi-definite, then also* $\boldsymbol{Q}^\intercal \boldsymbol{A}\boldsymbol{Q}$ *is positive semi-definite.*
2. *If* $\boldsymbol{A}$ *is positive definite and* $\boldsymbol{Q}$ *has full column rank, then also* $\boldsymbol{Q}^\intercal \boldsymbol{A}\boldsymbol{Q}$ *is positive definite.*

B

Important Distribtuions and Parameter Estimation

In this chapter we present a short summary of the most important properties of estimation functions as well as some concepts of estimation theory. A general extensive introduction into the basic concepts of inductive statistics can be found in Fahrmeir et al. (2007) or Mosler and Schmid (2005).

B.1 Some one-dimensional distributions

Definition B.1.1 (Gamma-distribution). *A continuous, non-negative random variable X is called* gamma-distributed *with parameters $a > 0$ and $b > 0$, abbreviated by the notation $X \sim G(a, b)$, if it has a density function of the following form*

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \quad x > 0.$$

Lemma B.1.2. *Let $X \sim \mathcal{G}(a, b)$ be a continuous, non-negative random variable. Then its expectation and variance are given by:*

- $E[X] = \frac{a}{b}$
- $Var(X) = \frac{a}{b^2}$

Definition B.1.3 (χ^2-distribution). *A continuous, non-negative random variable X with density*

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}x\right), \quad x > 0,$$

is called χ^2-distributed with n degrees of freedom*, abbreviated by the notation $X \sim \chi^2_n$.*

Lemma B.1.4. *Let $X \sim \chi^2_n$ be a continuous, non-negative random variable. Then its expectation and variance are given by:*

- $E[X] = n$
- $Var(X) = 2n$

Remark B.1.5. The χ^2-distribution is a special gamma-distribution with $a = n/2$ and $b = 1/2$.

Lemma B.1.6. *Let $X_1, \ldots, X_n$ be independent and identically standard normally distributed, then*

$$Y_n = \sum_{i=1}^{n} X_i^2$$

is χ^2-distributed with n degrees of freedom.

Definition B.1.7 (t-distribution). *A continuous random variable X with density*

$$f(x) = \frac{\Gamma(n+1)/2}{\sqrt{n\pi}\Gamma(n/2)(1+x^2/n)^{(n+1)/2}}$$

is called t-distributed with n degrees of freedom*, abbreviated by the notation $X \sim t_n$.*

Lemma B.1.8. *Let $X \sim t_n$ be a continuous, non-negative random variable. Then its expectation and variance are given by:*

- $E[X] = n, \quad n > 1$
- $Var(X) = n/(n-2), \quad n > 2.$

The t_1-distribution is also called Cauchy-distribution*. If $X_1, \ldots, X_n$ are iid with $X_i \sim N(\mu, \sigma^2)$, it follows that*

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \sim t_{n-1},$$

with

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \textit{and} \quad \bar{X} = \sum_{i=1}^{n} X_i.$$

Definition B.1.9 (F-distribution). *Let X_1 and X_2 be independent random variables with χ^2_n- and χ^2_m-distributions, respectively. Then the random variable*

$$F = \frac{X_1/n}{X_2/m}$$

is called F-distributed *with n and m degrees of freedom, abbreviated with the notation $F \sim F_{n,m}$.*

Definition B.1.10 (Log-normal distribution). *A continuous, non-negative random variable X is called* logarithmically normally distributed*, $X \sim LN(\mu, \sigma^2)$, if the transformed variable $Y = \log(X)$ is following a normal distribution, $Y \sim N(\mu, \sigma^2)$. The density of X is given by*

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right), \qquad x > 0.$$

The expectation and the variance yield

$$E[X] = \exp(\mu + \sigma^2/2)$$
$$Var(X) = \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1).$$

B.2 Some Important Properties of Estimation Functions

Definition B.2.1. *An* estimation function *or* statistic *for a parameter θ in the population is a function*

$$\hat{\theta} = g(X_1, \ldots, X_n)$$

of the sample variables $X_1, \ldots, X_n$. The numerical value

$$g(x_1, \ldots, x_n)$$

obtained from the realisations $x_1, \ldots, x_n$ is called estimate.

Definition B.2.2. *An estimation function $\hat{\theta} = g(X_1, \ldots, X_n)$ is called* unbiased for θ, *if*

$$E_\theta[\hat{\theta}] = \theta.$$

It is called asymptotically unbiased for θ, *if*

$$\lim_{n\to\infty} E_\theta[\hat{\theta}] = \theta.$$

The bias *is defined by*

$$Bias_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta.$$

Definition B.2.3. *The* mean squared error *is defined by*

$$MSE_\theta(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

and can be expressed in the following form

$$MSE_\theta(\hat{\theta}) = Var(\hat{\theta}) + Bias_\theta(\hat{\theta})^2.$$

Definition B.2.4. *An estimation function $\hat{\theta}$ is called* MSE-consistent *or simply* consistent, *if*

$$MSE_\theta(\hat{\theta}) \overset{n\to\infty}{\longrightarrow} 0.$$

Definition B.2.5. *An estimation function $\hat{\theta}$ is called* weak consistent, *if for any $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

or

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = 0,$$

respectively.

C

Central Limiting Value Theorems

For dealing with asymptotic statistical methods one needs definitions, concepts and results for the convergence of sequences of random variables. An extensive overview of the different convergence definitions can be found in Pruscha (2000). There, also references for the corresponding proofs are given. In the following we only present those definitions, which are important for this course.

Definition C.0.1. *Let the sequence of $p-$dimensional random vectors $X_n, n \geq 1$, and X_0 be given with distribution functions $F_n(x), x \in \mathbb{R}^p$, and $F_0(x), x \in \mathbb{R}^p$, respectively. The sequence $X_n, n \geq 1$, converges* in distribution *to X_0, if*

$$\lim_{n\to\infty} F_n(x) = F_0(x) \quad \forall x \in C_0,$$

with $C_0 \subset \mathbb{R}^p$ denoting the set of all points where $F_0(x)$ is continuous. In this case one also uses the notation $X_n \xrightarrow{d} X_0$.

D

Probability Theory

This chapter contains a short summary of some important results from stochastic and probability theory.

Proposition D.0.1. *Uni- and multivariate normally distributed random variables can be transformed in the following way:*

- *Univariate case:*

$$x \sim N(\mu, \sigma^2) \quad \Longleftrightarrow \quad ax \sim N(a\mu, a^2\sigma^2) \quad (\text{a } \textit{is a constant}).$$

- *Multivariate case:*

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \Longleftrightarrow \quad \mathbf{A}\mathbf{x} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal),$$

  where $\mathbf{A}$ *is a deterministic* $(q \times p)$*-dimensional matrix.*

Theorem D.0.2. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be random vectors,* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{a}, \boldsymbol{b}$ *matrices and vectors, respectively, of suitable dimension, and let* $E[\boldsymbol{x}] = \boldsymbol{\mu}$ *and* $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma}$*. Then:*

1. $E[\boldsymbol{x} + \boldsymbol{y}] = E[\boldsymbol{x}] + E[\boldsymbol{y}]$.
2. $E[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A} \cdot E[\boldsymbol{x}] + \boldsymbol{b}$.
3. $Cov(\boldsymbol{x}) = E[\boldsymbol{x}\boldsymbol{x}^\intercal] - E[\boldsymbol{x}]E[\boldsymbol{x}]^\intercal$.
4. $Var(\boldsymbol{a}^\intercal \boldsymbol{x}) = \boldsymbol{a}^\intercal Cov(\boldsymbol{x})\boldsymbol{a} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j \sigma_{ij}$.
5. $Cov(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) = \boldsymbol{A}Cov(\boldsymbol{x})\boldsymbol{A}^\intercal$.
6. $E[\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x}] = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\intercal \boldsymbol{A}\boldsymbol{\mu}$.

D.1 The Multivariate Normal Distribution

Definition D.1.1. *A* p*-dimensional random vector* $\mathbf{x} = (x_1, \ldots, x_p)^\intercal$ *is called* multivariate normally distributed*, if* $\mathbf{x}$ *has the density function*

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}, \tag{D.1.1}$$

with $\boldsymbol{\mu} \in \mathbb{R}^p$ *and positive semi-definite* $p \times p$ *matrix* $\boldsymbol{\Sigma}$*.*

Theorem D.1.2. *The expectation and the covariance matrix of a multivariate normally distributed vector* $\mathbf{x}$ *with density function* (D.1.1) *are given by* $E[\mathbf{x}] = \boldsymbol{\mu}$ *and* $Cov(\mathbf{x}) = \boldsymbol{\Sigma}$*. Hence, we use the notation*

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

similarly to the univariate case. Often the index p *is suppressed, if the dimension is clear by context. For* $\boldsymbol{\mu} = \mathbf{0}$ *and* $\boldsymbol{\Sigma} = \boldsymbol{I}$ *the distribution is called* (multivariate) standard normal distribution*.*

Theorem D.1.3. *Let* $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *be multivariate normally distributed. Now regard a partition of* $\boldsymbol{X}$ *into two sub-vectors* $\boldsymbol{Y} = (X_1, \ldots, X_r)^\intercal$ *and* $\boldsymbol{Z} = (X_{r+1}, \ldots, X_n)^\intercal$*, i.e.*

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_Z \end{pmatrix}.$$

Then the sub-vector $\boldsymbol{Y}$ *is again* r*-dimensional normally distributed wih* $\boldsymbol{Y} \sim N_r(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$*. Besides, the conditional distribution of* $\boldsymbol{Y}$ *given* $\boldsymbol{Z}$ *is also a multivariate normal distribution with expectation*

$$\boldsymbol{\mu}_{Y|Z} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YZ} \cdot \boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_Z)$$

and covariance matrix

$$\boldsymbol{\Sigma}_{Y|Z} = \boldsymbol{\Sigma}_Y - \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_{ZY}.$$

Furthermore, for normally distributed random variables the situations of independency and of being uncorrelated are equivalent: $\boldsymbol{Y}$ *and* $\boldsymbol{Z}$ *are independent if and only if* $\boldsymbol{Y}$ *and* $\boldsymbol{Z}$ *are uncorrelated, i.e.* $\boldsymbol{\Sigma}_{YZ} = \boldsymbol{\Sigma}_{ZY} = \mathbf{0}$*. For non-normal random variables this equivalence does not hold in general. Here, from independency merely follows uncorrelatedness.*

D.1.1 The Singular Normal Distribution

Definition D.1.4. *Let* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. The distribution of* $\mathbf{x}$ *is called* singular*, if* $rk(\boldsymbol{\Sigma}) < p$ *holds. In such cases, the distribution is often characterized by the precision matrix* $\boldsymbol{P}$ *(with* $rk(\boldsymbol{P}) < p$*) instead of the covariance matrix (see e.g. the field of spatial statistics for examples). The random vector* $\mathbf{x}$ *then has a density function*

$$f(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\intercal \boldsymbol{P}(\boldsymbol{x} - \boldsymbol{\mu})\right],$$

which is only defined up to proportionality.

Theorem D.1.5. *Let* $\mathbf{x}$ *be a* p*-dimensional random vector, which is singularly distributed, i.e.* $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $rk(\boldsymbol{\Sigma}) = r < p$*. Let* $(\boldsymbol{GB})$ *be an orthogonal matrix, in which the rows of the* $p \times r$ *matrix* $\boldsymbol{G}$ *form a basis of the column space of* $\boldsymbol{\Sigma}$ *and the rows of* $\boldsymbol{B}$ *a basis of the null-space of* $\boldsymbol{\Sigma}$*. Consider the transformation*

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = (\boldsymbol{GB})^\intercal \mathbf{x} = \begin{pmatrix} \boldsymbol{G}^\intercal \mathbf{x} \\ \boldsymbol{B}^\intercal \mathbf{x} \end{pmatrix}.$$

Then $\boldsymbol{y}_1$ *is the stochastic part of* $\mathbf{x}$ *and not singular with*

$$\boldsymbol{y}_1 \sim N_r(\boldsymbol{G}^\intercal \boldsymbol{\mu}, \boldsymbol{G}^\intercal \boldsymbol{\Sigma} \boldsymbol{G}),$$

$\boldsymbol{y}_2$ *is the deterministic part of* $\mathbf{x}$ *with*

$$E[\boldsymbol{y}_2] = \boldsymbol{B}^\intercal \boldsymbol{\mu} \quad \textit{and} \quad Var(\boldsymbol{y}_2) = \mathbf{0}.$$

The density of the stochastic part $\boldsymbol{y}_1 = \boldsymbol{G}^\intercal \mathbf{x}$ *has the form*

$$f(\boldsymbol{y}_1) = \frac{1}{(2\pi)^{\frac{r}{2}} (\prod_{i=1}^r \lambda_i)^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{y}_1 - \boldsymbol{G}^\intercal \boldsymbol{\mu})^\intercal (\boldsymbol{G}^\intercal \boldsymbol{\Sigma} \boldsymbol{G})^{-1} (\boldsymbol{y}_1 - \boldsymbol{G}^\intercal \boldsymbol{\mu})\},$$

where λ_i *denote the corresponding* r *different non-zero eigen values and* $\boldsymbol{\Sigma}^-$ *is a generalized inverse of* $\boldsymbol{\Sigma}$*.*

D.1.2 Distributions of Quadratic Forms

Distributions of quadratic forms of normally distributed random vectors play an important role for testing linear hypotheses, compare Section 2.2.5.

Theorem D.1.6 (Distributions of quadratic forms).

(i) Let $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\Sigma} > 0$*. Then:*

$$y = (\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_p.$$

(ii) Let $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{I}_p)$*,* $\mathbf{B}$ *an* $n \times p$ $(n \leq p)$ *matrix and* $\boldsymbol{R}$ *a symmetric, idempotent* $p \times p$ *matrix with* $rk(\boldsymbol{R}) = r$*. Then:*

- $\mathbf{x}^\intercal \boldsymbol{R} \mathbf{x} \sim \chi^2_r$.
- *From* $\mathbf{B}\boldsymbol{R} = \mathbf{0}$ *follows that the quadratic form* $\mathbf{x}^\intercal \boldsymbol{R} \mathbf{x}$ *and the linear form* $\mathbf{B}\mathbf{x}$ *are independent.*

(iii) Let $X_1, \ldots, X_n$ *be independent random variables with* $X_i \sim N(\mu, \sigma^2)$ *and*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Then:

- $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$.
- S^2 *and* $\bar{X}$ *are independent.*

(iv) Let $\mathbf{x} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, $\boldsymbol{R}$ *and* $\boldsymbol{S}$ *be symmetric and idempotent* $n \times n$ *matrices with* $rk(\boldsymbol{R}) = r$ *and* $rk(\boldsymbol{S}) = s$ *and* $\boldsymbol{RS} = \mathbf{0}$. *Then:*

- $\mathbf{x}^\intercal \boldsymbol{R} \mathbf{x}$ *and* $\mathbf{x}^\intercal \boldsymbol{S} \mathbf{x}$ *are independent.*
- $\dfrac{s}{r} \dfrac{\mathbf{x}^\intercal \boldsymbol{R} \mathbf{x}}{\mathbf{x}^\intercal \boldsymbol{S} \mathbf{x}} \sim F_{r,s}$.

References

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall.

Durbin, J. and G. S. Watson (1950). Testing for serial correlation in least squares regression I. *37*, 409–428.

Durbin, J. and G. S. Watson (1951). Testing for serial correlation in least squares regression II. *38*, 159–178.

Durbin, J. and G. S. Watson (1971). Testing for serial correlation in least squares regression III. *58*, 1–19.

Fahrmeir, L. (1987). Asymptotic testing theory for generalized linear models. *Math. Operationsforsch. Statist. Ser. Statist. 18*, 65–76.

Fahrmeir, L., T. Kneib, and S. Lang (2007). *Regression*. Berlin: Springer.

Fahrmeir, L., R. Kuenstler, I. Pigeot, and G. Tutz (2007). *Statistik - Der Weg zur Datenanalyse*. Berlin: Springer.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer-Verlag.

Greene, W. H. (2000). *Econometric Analysis* (4. ed.). Upper Saddle River, NJ: Prentice Hall.

Hoaglin, D. and R. Welsch (1978). The hat matrix in regression and ANOVA. *American Statistician 32*, 17–22.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society, Series B, Methodological 49*, 127–145.

Judge, G. G., W. E. Griffith, R. C. Hill, H. Lütkepol, and T.-C. Lee (1980). *The Theory and Practice of Econometrics*. New York: Wiley.

Leeflang, P. S. H., D. R. Wittnik, M. Wedel, and P. A. Naert (2000). *Building Models for Marketing Decisions*. Boston: Kluwer.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall.

Mosler, K. and F. Schmid (2005). *Wahrscheinlichkeitsrechnung und schließende Statistik*. Berlin: Springer–Verlag.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society A 135*, 370–384.

Piegorsch, W. (1992). Complementary log regression for generalized linear models. *The American Statistician 46*, 94–99.

Pruscha, H. (2000). *Vorlesungen über Mathematische Statistik*. Stuttgard: B. G. Teubner.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.). New York: Springer.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Tutz, G. (2000). *Die Analyse kategorialer Daten – eine anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression*. München: Oldenbourg Verlag.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Wooldridge, J. M. (2006). *Introductory Econometrics* (3.rd ed.). Ohio: Thomson, Mason.